Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation

Paul-Gauthier Noé, Mohammad Mohammadamini, Driss Matrouf, Titouan Parcollet, Andreas Nautsch, Jean-François Bonastre

Adversarial Disentanglement of Speaker Representation for Attribute-Driven Privacy Preservation

Paul-Gauthier Noé[1], *Mohammad Mohammadamini*[1], *Driss Matrouf*[1], *Titouan Parcollet*[1,2], *Andreas Nautsch*[3], *Jean-François Bonastre*[1]

[1]Laboratoire Informatique d'Avignon (LIA), Avignon Université, France
[2]University of Cambridge, United-Kingdom
[3]Digital Security Department, EURECOM, France

paul-gauthier.noe@univ-avignon.fr

Abstract

In speech technologies, speaker's voice representation is used in many applications such as speech recognition, voice conversion, speech synthesis and, obviously, user authentication. Modern vocal representations of the speaker are based on neural embeddings. In addition to the targeted information, these representations usually contain sensitive information about the speaker, like the age, sex, physical state, education level or ethnicity. In order to allow the user to choose which information to protect, we introduce in this paper the concept of *attribute-driven privacy preservation* in speaker voice representation. It allows a person to hide one or more personal aspects to a potential malicious interceptor and to the application provider. As a first solution to this concept, we propose to use an adversarial autoencoding method that disentangles in the voice representation a given speaker attribute thus allowing its concealment. We focus here on the sex attribute for an Automatic Speaker Verification (ASV) task. Experiments carried out using the VoxCeleb datasets have shown that the proposed method enables the concealment of this attribute while preserving ASV ability.

1. Introduction

Data protection regulation such as the European *General Data Protection Regulation* (GDPR) and the privacy concern in speech technologies [1] call for more control on the personal information an user discloses while using such technologies. This paper presents and extends the *attribute-driven privacy preservation* introduced in [2]. The idea is to let the user decide what personal information they agree to disclose in their voice data while using a given voice service. In this work, we focus our efforts on the widely used x-vector [3] speaker's voice representation. Despite its effectiveness for many applications, this representation is known to contain sensitive information [4]. To answer this privacy issue, most of the speech privacy preservation systems impact the full speaker's representation [5, 6, 7, 8, 9] while in some use cases it is necessary and/or sufficient to hide only one or a few personal attributes in order to maintain the performance of the vocal system or to preserve the vocal richness as much as possible. To overcome this drawback, we propose to consider the disentanglement learning [10] to facilitate the control over information in speaker's voice representations.

Various works have been proposed on linguistic/non-linguistic [11, 12, 13, 14] and speaker/noise information disentanglement [15]. In [16] an adversarial approach is used to hide the speaker's identity in Automatic Speech Recognition (ASR) representations. Most of the investigations on disentanglement methods are conducted directly on acoustic feature sequences, and few works have been proposed to operate on speaker representations. In [17], an autoencoder is applied on x-vectors to disentangle the speaking style and the identity of the speaker within two subspaces. More recently, in [18], an unsupervised disentanglement method have been used over x-vectors to separate the speaker-discriminative information and the remaining noises for robust speaker recognition. However, none of these works considered to disentangle soft-biometric attributes directly from speaker representations. In [19], an adversarial approach is proposed for the extraction of gender-discriminative features with low speaker-discriminative information. Whereas this approach aims to enable the detection of a speaker attribute (the sex[1]) while avoiding speaker identification, our work allows the contrary: enabling speaker verification while avoiding the detection of a specific speaker attribute.

Indeed, as a first example and solution to the attribute-driven privacy preservation idea, this paper presents a method to disentangle and hide the sex attribute in x-vector embeddings using an adversarial autoencoding approach. Our goal is to allow the user to hide this specific personal attribute in the x-vector while preserving the remaining information in order to perform authentication-by-voice. While we focus here on ASV, this approch can be experimented on any application that is based on such speaker representation from disease detection [21] to voice conversion and anonymisation [7] for instance. The effectiveness of the method is assessed on sex concealment evaluation and speaker verification tasks using the VoxCeleb corpora [22, 23].

In this paper, Section 2 explains the attribute-driven privacy preservation concept; Section 3 presents the adversarial disentangling autoencoder[2] for hiding a binary attribute in x-vectors; Section 4 presents the experimental setup and the evaluation metrics; The results on the VoxCeleb corpora are then presented and discussed in Section 5; Finally, Section 6 concludes and presents potential future works.

2. Attribute-driven privacy preservation

The idea behind the attribute-driven privacy preservation is to not disclose one or a few personal information while enabling a desired task such as automatic speech recognition or automatic speaker verification for instance. This idea has been independently presented in [24] and is referred as *user-configurable*

[1]Throughout this paper, the term *sex* refers to the biological differences between female and male [20].

[2]Code is available at https://github.com/LIAvignon/adversarial-disentangling-autoencoder-for-spk-representation

privacy. As a first toy example, this paper focuses on the sex attribute and automatic speaker verification.

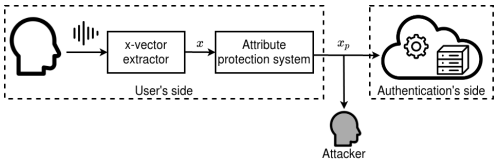


Figure 1: *Example of an attribute-driven privacy preservation use case. The released data x_p are processed by the authentication side but may also be intercepted by an unexpected attacker. Therefore, the user employs an attribute protection system in order to not disclose some aspects s/he decided to keep secret.*

Figure 1 illustrates the situation where an user conceals some information in its voice representation in order to avoid the authentication side and an attacker to infer these personal information. Indeed, they could try to detect for example the sex of the user using scores obtained from a sex classifier that operates on x-vectors. The role of the protection system is thus to make these scores unexploitable.

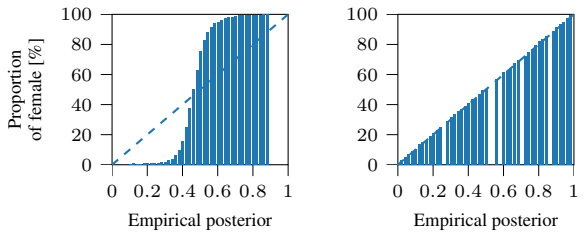
3. Adversarial disentangling autoencoder and attribute protection

This section presents a first solution to conceal the sex of the speaker in its x-vector representations.

3.1. Adversarial disentangling autoencoder

The proposed model, similarly to the one used in [25] for image processing, disentangles the sex information in x-vector speaker representations. It has four components: a pre-trained sex classifier, an encoder, a decoder and an adversarial sex classifier.
Notation: let $D = \{(x_1, y_1, \tilde{y}_1), ...(x_m, y_m, \tilde{y}_m)\}$ be a set of x-vectors with their binary sex label (0 for male, 1 for female) and as a soft label, the posterior probability $\tilde{y} = P(F|x)$ where F is the proposition: *"the speech segment has been uttered by a female speaker"*.
The pre-trained sex classifier is used as a feature extractor that predicts the posteriors $\tilde{y}$ as soft labels. Soft labelling is used instead of hard labelling to make the *sex* variable not strongly binary and to allow values between 0 and 1. Moreover, having probabilistic interpretation would be consistent with the Bayesian decision framework. As shown in Figure 2, the scores might not correspond to proper posterior probabilities. Therefore, a calibration step [26] is added in order to produce oracle calibrated scores.
The encoder encodes an input x-vector x into an embedding z. In addition to compressing the input information, it tries to cheat the adversarial sex classifier.
The decoder reconstructs the original x-vector x from the variable z and a condition w. During the training $w = \tilde{y}$.
The adversarial sex classifier tries to predict, from the encoded vector z, the sex class of the corresponding speech segment.

The training and protection flows in the model are illustrated in Figure 3. The sex information is disentangled from the rest using an adversarial training that opposes the encoder-decoder to the adversarial sex classifier. Thus, a first optimiser updates the neural parameters θ_d of the adversarial sex classifier with respect to the correct expected class prediction of x.



(a) *Without calibration* (b) *With oracle calibration*

Figure 2: *Empirical calibration plot [27] on the pre-trained classifier's scores on V2D. Applying PAVA [26] results in oracle calibrated scores (b).*

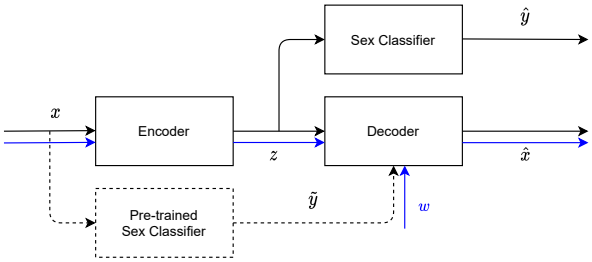


Figure 3: *Illustration of the Adversarial Disentangling Autoencoder. The black arrows illustrate the forward flow during the training phase. The dashed lines represent the part that is not updated during training: the pre-trained sex classifier is used to extract posteriors $\tilde{y}$ for feeding the decoder. The blue arrows illustrate the forward flow during the protection phase. Sex attribute control is done by setting w.*

Then, a second optimiser updates the parameters θ_E of the encoder and the parameters θ_D of the decoder in order to jointly cheat the adversarial sex classifier and reconstruct the x-vector. Therefore, the encoder-decoder and the sex classifier are trained in an adversarial manner in order to make the encoded vector z sex-independent. The two objective functions are respectively:

$$L_d(\theta_d|\theta_E) = -\frac{1}{m}\sum_{i=1}^{m}\log(\hat{y}_i), \tag{1}$$

$$L(\theta_E, \theta_D|\theta_d) = \frac{1}{m}\sum_{i=1}^{m}\left(r(\hat{x}_i, x_i) - \log(1-\hat{y}_i)\right), \tag{2}$$

where $z = E_{\theta_E}(x)$ and $\hat{x} = D_{\theta_D}(z, \tilde{y})$ are respectively the output of the encoder and the decoder, $\hat{y}$ is the adversarial sex classifier predicted score and $r(\hat{x}, x) = 1 - \frac{<\hat{x}, x>}{||\hat{x}||||x||}$ is the reconstruction error. During our experiments, both optimisers follow the stochastic gradient descent with the model parameters $\theta_E^{(t)}$, $\theta_D^{(t)}$ and $\theta_d^{(t)}$ being updated at each time step t as follow:

$$\theta_d^{(t+1)} = \theta_d^{(t)} \quad -\eta\nabla_{\theta_d} L_d(\theta_d^{(t)}|\theta_E^{(t)}, x^{(t)}, y^{(t)}), \tag{3}$$

$$\begin{bmatrix}\theta_E^{(t+1)} \\ \theta_D^{(t+1)}\end{bmatrix} = \begin{bmatrix}\theta_E^{(t)} \\ \theta_D^{(t)}\end{bmatrix} -\eta\nabla_{\theta_E,\theta_D} L(\theta_E^{(t)}, \theta_D^{(t)}|\theta_d^{(t)}, x^{(t)}, y^{(t)}), \tag{4}$$

with $x^{(t)}$ the training sample and $y^{(t)}$ its corresponding class at time t.

3.2. Attribute protection

Setting the value of the decoder input w permits to control in the x-vector the sex information which is expressed by the posterior $P(F|x)$. To achieve zero-evidence [28], zero log-likelihood ratios (LLRs) need to result: to protect the sex attribute, when x-vectors are taken as evidence, the likelihood for an x-vector given the female proposition needs to equal the likelihood for an x-vector given the male proposition. Such zero LLRs imply that updating any prior belief with them results in posterior belief of the same value. In this way, using a sex classifier becomes useless. This idea is motivated from perfect secrecy [29] and method validation in forensic sciences [30].

Different perspectives arise for setting w: when setting $w = \tilde{y}$, privacy is not preserved; for zero LLRs, the attacker posterior must remain the attacker prior π thus, w must be set to π. However, this prior is unknown in real application and an uninformed prior 0.5 is hence assumed in this work. Therefore, for every x-vector, w is set to 0.5 for protection. The uninformed prior suggests good performance on average, whereas particular cases of specific attacker priors will lead to different results. This work is not investigating to ensure zero LLRs through generative training methods.

4. Experimental and Evaluation Setup

This section first details the architecture, the training procedure of the proposed model and the corpora used (Section 4.1). Then, evaluation protocols are described (Section 4.2).

4.1. Model architecture, training and corpora

First, a standardisation and a length normalisation [31] are applied on the x-vectors before being fed into the model. The encoder consists of a single dense layer with ReLU activation functions [32]. Batch-normalisation [33] is applied accross the resulting 128 dimensional representation z. Then, the decoder takes as inputs the encoded vector z concatenated with the sex posterior $\tilde{y}$, and consists of a single dense layer with a hyperbolic tangent for the activation function followed by a last length normalization. The adversarial classifier is composed with two dense layers. The first one has 64 units with ReLU activation functions, and the second one has 1 unit with a sigmoid activation function. The model is trained using two standard stochastic gradient descents. Learning rates are set to 10^{-4} and the momentums are 0.9. A subset V1D (61616 segments per class) of VoxCeleb [22] is used to train the external sex classifier[3] that extracts the posteriors $\tilde{y}$. A subset V2D (397032 segments per class) of VoxCeleb2 development part [23] is used to train the adversarial disentangling autoencoder model. A subset V2T (9120 female segments and 22559 male segments) of VoxCeleb2 testing part is used to test how well the model protects the attribute.

4.2. Evaluation protocol

As explained in Section 3.2, the decoder input w is set to 0.5 in order to hide the sex information in the reconstructed x-vector. The following explains how to assess the level of protection and its impact on the utility i.e. the ASV performance.

[3] A one layer perceptron with a sigmoid as activation function. Scores are oracle calibrated using the pool-adjacent-violator algorithm [26].

Attribute privacy preservation assessment:
In order to assess the protection ability, several metrics are computed from the sex classifier scores. We compute the area under the receiver operating characteristic curve (AUC). An AUC of 50% would confirm the random prediction of the classifier. The $C_{\text{llr}}^{\text{min}}$ is also computed as a discrimination cost [34]. Recently, the Zero Evidence Biometric Recognition Assessment framework was introduced [28]. Originally presented in the context of speaker identity preservation, it can be applied to any binary decision task. It provides a measure D_{ECE} of the expected amount of private information disclosed as well as the worst-case score l_w i.e. the strongest strength-of-evidence in a set of segments. The latter comes with a categorical tag for better interpretation. For more details on these measures, refer to [28].
The Mutual Information (MI) can also be used as a privacy measure [35]. In our experiments, MI is used to measure the amount of information that the x-vector is sharing with the sex variable y. The MI between the x-vectors' components and y is estimated [36, 37] and its average over the x-vector's dimensions is given.
Automatic speaker verification evaluation:
As a first example of an attribute-driven privacy preservation application, we focus on ASV. An important aim of this work is indeed to hide the sex information in the x-vector representations while maintaining automatic speaker verification ability. Thus, ASV performance with the protected x-vectors are measured. The commonly used Probabilistic Linear Discriminant Analysis (PLDA) [38] is used to compute the likelihood-ratios from the comparisons of enrolment and probe x-vectors.

Table 1: *Description of the ASV datasets.*

	Number of segments		Number of speakers	
	Enrolment	Test	Enrolment	Test
Male	11282	11277	81	81
Female	4558	4562	39	39

Both enrolment and probe x-vectors are converted. The Equal Error Rate (EER) and the $C_{\text{llr}}^{\text{min}}$ are measured to assess the ASV. The PLDA have been trained on $200,000$ x-vectors randomly chosen from the VoxCeleb 1 and 2 training subsets. Their dimension is beforehand reduced to 128 with a linear discriminant analysis. Details on the enrolment and test sets are shown in Table 1.

5. Results

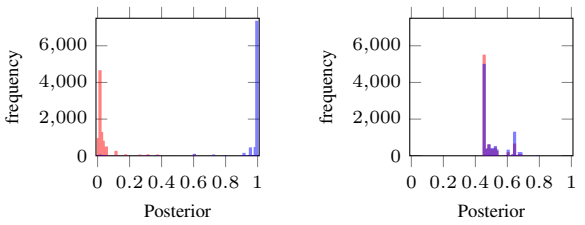
This section presents the results of the sex attribute concealment (Section 5.1) and the ASV performance on the protected x-vectors (Section 5.2).

5.1. Attribute privacy preservation results

Table 2 reports the sex classifier's AUC, the $C_{\text{llr}}^{\text{min}}$ and the ZEBRA measures obtained with original, unprotected reconstructed (i.e. the x-vectors have been passed through the system but with $w = \tilde{y}$) and protected x-vectors (i.e. with $w = 0.5$). At first, the AUC (lower than 0.5 for V2D with $w = 0.5$), suggests that the class labels might be swapped/the direction of LLR scores is negated. The association of class labels to particular posterior values is arbitrarily fixed but since the method can violates this *fixed* assumption, inference using these assumptions can suffers. We thus report results from labels' association that results in lower $C_{\text{llr}}^{\text{min}}$ and bigger D_{ECE}.

Table 2: *Effects of the reconstruction and the sex protection on the ability to predict, using a pre-trained classifier, the sex from x-vector on V2D and V2T. The first row reports results on original x-vectors. The second row refers to the x-vectors that have been passed through the system but without transformation ($w = \tilde{y}$). The last row refers to the protected x-vectors ($w = 0.5$).*

	AUC $_{10^{-2}}$		$C_{\text{llr}}^{\text{min}}{}_{10^{-2}}$		(D_{ECE}, $\log_{10}(l_w)$, tag)	
	V2D	V2T	V2D	V2T	V2D	V2T
Original vector x	98.84	99.09	15.97	13.18	(0.596, 2.910, C)	(0.619, 3.538, C)
$w = \tilde{y}$	98.00	97.29	19.19	17.60	(0.570, 2.859, C)	(0.584, 3.554, C)
$w = 0.5$	49.19	55.23	99.79	98.64	(0.001, 0.813, A)	(0.009, 0.393, A)



(a) *Original data* (b) *With protection ($w = 0.5$)*

Figure 4: *Histograms of the calibrated scores from the sex classifier on V2T original (a) and protected (b). Female scores are in blue and male scores are in red. When protection is applied (b), the score distributions overlap almost perfectly.*

With $w = \tilde{y}$, the sex from whom the reconstructed x-vector comes from can still be detected correctly. With $w = 0.5$, the AUC is getting close to 50% which corresponds to random prediction. There is also a significant increase of the $C_{\text{llr}}^{\text{min}}$ which confirms the difficulty to separate the score distributions also illustrated in Figure 4. Moreover, the drop of expected privacy disclosure (D_{ECE}) corroborates the difficulty for the adversary to benefit from the scores and the low strongest strength-of-evidence (A) suggests that no x-vectors are left with a poor protection. The mutual information measures are shown in Table 3.

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

	$I(\hat{x}, y)$ $_{10^{-2}}$ [bit per dimension]	
	V2D	V2T
Original vector x	18.7	19.0
$w = \tilde{y}$	20.3	19.81
$w = 0.5$	1.0	1.90

There is a significant decrease of mutual information when w is set to 0.5. Therefore, the system seems to reduce the dependency between the protected x-vector and the sex variable resulting in a more *sex-independent* speaker representation.

5.2. Automatic speaker serification results

Table 4 shows the results obtained for the ASV task on the transformed x-vectors. For both unprotected reconstructed ($w = \tilde{y}$) x-vectors and protected ($w = 0.5$) x-vectors the ASV performance slightly decreases. With protection, the EER increases from 1.72 to 2.36 and the $C_{\text{llr}}^{\text{min}}$ increases from 0.067 to 0.097. The ASV performance is slightly better for $w = 0.5$ in comparison to $w = \tilde{y}$. This suggests that the lose of performance is mostly due to the reconstruction error and that segments comparison for speaker verification could go without considering most of the sex information.

Table 4: *Effects of the reconstruction and the sex concealment on the ASV.*

	EER [%]	$C_{\text{llr}}^{\text{min}}$
Original vector x	1.72	0.067
$w = \tilde{y}$	2.89	0.118
$w = 0.5$	2.36	0.097

6. Conclusion

In this paper we have introduced the *attribute-driven privacy preservation* as the idea of enabling a speaker to hide only a few personal aspects in its voice representation while maintaining the remaining particularities and the performance of a desired task. As a first solution, we presented an adversarial autoencoding approach that disentangles and hide a given binary attribute in a x-vector neural embedding. This method is based on an encoder-decoder architecture combined with an additional classifier that tries to predict the attribute class from the encoded representation. Both are trained in an adversarial manner to make the encoded representation *attribute-independent*. This approach have been experimented on the sex attribute and with the aim of enabling speaker verification. Through the experiments conduced on the Voxceleb dataset, it has been shown that setting the sex variable representing the posterior to 0.5 for all segments enables to diminish the sex information in the speaker representation with only a slight alteration of the automatic speaker verification performance.
Even though the proposed approach reduces the amount of sex information contained in a x-vector, it is not guaranteed that an attacker in possession of the transformed data and their original sex class can not take advantage of the remaining sex information and train a new classifier that will detect the original sex. In future works, this approach will be tested on other attributes such as the age or the regional accent. The method will also be combined to a speech synthesizer, therefore, the attribute-driven privacy preservation will output speech signal just like speech anonymisation [9].

7. Acknowledgements

This work was supported by the VoicePersonae ANR-18-JSTS-0001 and the Robovox ANR-18-CE33-0014 projects.

8. References

[1] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps towards a Common Understanding," in *Proc. Interspeech*. ISCA, 2019, pp. 3695–3699.

[2] P.-G. Noé, M. Mohammadamini, D. Matrouf, T. Parcollet, and J.-F. Bonastre, "Adversarial disentanglement of speaker representation for attribute-driven privacy preservation," 2020, arXiv preprint :2012.04454, version 1.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. 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.

[4] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 726–733.

[5] M. Pathak, J. Portelo, B. Raj, and I. Trancoso, "Privacy-preserving speaker authentication," in *Information Security*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–22.

[6] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delacrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech & Language*, vol. 58, pp. 441 – 480, 2019.

[7] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker Anonymization Using X-vector and Neural Waveform Models," in *Proc. 10th Speech Synthesis Workshop*. ISCA, 2019, pp. 155–160.

[8] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[9] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy Initiative," in *Proc. Interspeech*. ISCA, 2020, pp. 1693–1697.

[10] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[11] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1432–1443, 2019.

[12] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. Interspeech*. ISCA, 2017, pp. 3364–3368.

[13] L. Yingzhen and S. Mandt, "Disentangled sequential autoencoder," in *Proc. 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80, 2018, pp. 5670–5679.

[14] S. Khurana, S. R. Joty, A. Ali, and J. Glass, "A factorial deep markov model for unsupervised disentangled representation learning from speech," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6540–6544.

[15] W.-N. Hsu, Y. Zhang, R. J. Weiss, Y.-A. Chung, Y. Wang, Y. Wu, and J. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5901–5905.

[16] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?" in *Proc. Interspeech*, 2019, pp. 3700–3704.

[17] J. Williams and S. King, "Disentangling style factors from speaker representations," in *Proc. Interspeech*. ISCA, 2019, pp. 3945–3949.

[18] R. Peri, M. Pal, A. Jati, K. Somandepalli, and S. Narayanan, "Robust speaker recognition using unsupervised adversarial invariance," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6614–6618.

[19] A. Nelus and R. Martin, "Gender discrimination versus speaker identification through privacy-aware adversarial feature extraction," in *Speech Communication; 13th ITG-Symposium*, 2018, pp. 1–5.

[20] V. Prince, "Sex vs. gender," *International Journal of Transgenderism*, vol. 8, no. 4, pp. 29–32, 2005.

[21] M. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Pathological speech detection using x-vector embeddings," *ArXiv*, vol. abs/2003.00864, 2020.

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

[23] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*. ISCA, 2018, pp. 1086–1090.

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

[25] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader networks: Manipulating images by sliding attributes," in *Advances in Neural Information Processing Systems*, 2017, pp. 5967–5976.

[26] N. Brummer and J. Preez, "The pav algorithm optimizes binary proper scoring rules," 2013.

[27] D. Ramos and J. Gonzalez-Rodriguez, "Reliable support: Measuring calibration of likelihood ratios," *Forensic Science International*, vol. 230, no. 1-3, pp. 156–169, 7 2013.

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

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

[30] S. E. Willis, L. Mc Kenna, S. Mc Dermott, A. Barrett, B. Rasmusson *et al.*, *ENFSI Guideline for Evaluative Reporting in Forensic Science*, European Network of Forensic Science Institutes, 2015.

[31] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*. ISCA, 2011, pp. 249–252.

[32] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd International Conference on Machine Learning*, ser. Machine Learning Research, vol. 37. PMLR, 2015, pp. 448–456.

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

[35] K. Kalantari, L. Sankar, and O. Kosut, "On information-theoretic privacy with general distortion cost functions," in *Proc. International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2865–2869.

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

[37] 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.

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