A comparative study of speech anonymization metrics

Mohamed Maouche, Brij Mohan Lal Srivastava, Nathalie Vauquier, Aurélien Bellet, Marc Tommasi, Emmanuel Vincent

A Comparative Study of Speech Anonymization Metrics

Mohamed Maouche[1], *Brij Mohan Lal Srivastava*[1], *Nathalie Vauquier*[1], *Aurélien Bellet*[1], *Marc Tommasi*[2], *Emmanuel Vincent*[3]

[1]Inria, France [2]Université de Lille, France
[3]Université de Lorraine, CNRS, Inria, LORIA, France

<firstname>.<lastname>@inria.fr

Abstract

Speech anonymization techniques have recently been proposed for preserving speakers' privacy. They aim at concealing speakers' identities while preserving the spoken content. In this study, we compare three metrics proposed in the literature to assess the level of privacy achieved. We exhibit through simulation the differences and blindspots of some metrics. In addition, we conduct experiments on real data and state-of-the-art anonymization techniques to study how they behave in a practical scenario. We show that the application-independent log-likelihood-ratio cost function $C_{\text{llr}}^{\text{min}}$ provides a more robust evaluation of privacy than the equal error rate (EER), and that detection-based metrics provide different information from linkability metrics. Interestingly, the results on real data indicate that current anonymization design choices do not induce a regime where the differences between those metrics become apparent.

Index Terms: anonymization, voice conversion, speaker recognition, privacy metrics.

1. Introduction

With the increasing popularity of smart devices, more users have access to voice-based interfaces. They offer simple access to modern technologies and enable the development of new services. The building blocks behind these speech-based technologies are no more handcrafted but learned from large sets of data. This is the case for instance of automatic speech recognition (ASR), where vast volumes of speech in different languages are needed and continuously collected to improve performance and adapt to new domains. The collection and exploitation of speech data raises privacy threats. Indeed, speech contains private or sensitive information about the speaker (e.g., gender, emotion, speech content) [1] and it is a biometric characteristic that can be used to recognize the speaker through, e.g., i-vector [2] or x-vector [3] based speaker verification.

To address this privacy issue, various anonymization techniques have been studied in the literature[1]. Their purpose is to transform speech signals in order to preserve all content except features related with the speaker identity. These techniques include noise addition [4], speech transformation [5], voice conversion [6, 7, 8], speech synthesis [9], or adversarial learning [10]. As a privacy preservation mechanism, they must achieve a suitable privacy/utility trade-off. The utility is typically assessed in terms of the accuracy of downstream processing steps (e.g., the word error rate achieved by an ASR system). The measurement of privacy is the topic we tackle in this paper.

Historically, the usual metrics employed in the speaker verification community have been used to assess the (in)ability of an attacker to recognize the speaker, which is considered as a proxy for privacy. The most widely used metric is the *equal error rate* (EER): it considers an attacker that makes a decision by comparing speaker similarity scores with a threshold and it assigns the same cost to false alarms and misses [11]. The *application-independent log-likelihood-ratio cost function* $C_{\text{llr}}^{\text{min}}$ generalizes the EER by considering optimal thresholds over all possible priors and all possible error costs [12]. In the following, we consider a third metric called *linkability* which has recently emerged from the biometric template protection community but has received little attention in the speech community so far [13]. This metric, denoted as $D_{\leftrightarrow}^{\text{sys}}$, estimates the distributions of scores for *mated* (same-speaker) vs. *non-mated* (different-speaker) trials and computes their overlap.

The goal of this paper is to assess the suitability of these three metrics for the evaluation of speaker anonymization. In addition to comparing the metrics in their form and substance, we generate simulated data to exhibit their blindspots. We also conduct experiments on real speech data processed by state-of-the-art anonymization techniques against different attackers (ignorant, semi-informed, or informed [14]). Overall, we aim to understand the complementary factors underlying different metrics and ensure that the anonymization techniques being evaluated were not designed to fool attackers that follow one specific speaker verification method but would fail with others.

We describe the attack model in Section 2 and introduce the metrics in Section 3. We present the simulations used to exhibit their blindspots in Section 4. Section 5 reports the results of the evaluation on real data with various anonymization techniques and attack types. We conclude in Section 6.

2. Attack Model

The attack scenario is depicted in Fig. 1. *Speakers* process their voice through an *anonymization* technique. This anonymization step takes as input one or more *private speech* utterances along

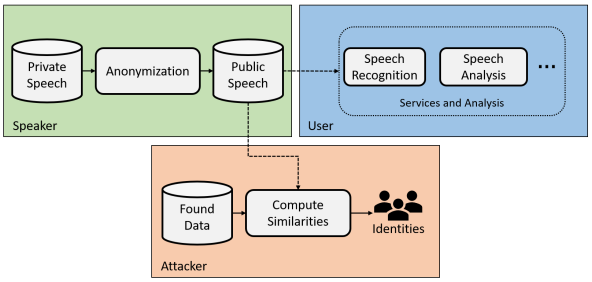

Figure 1: *Anonymization procedure and attack model.*

[1]In the legal community, the term "anonymization" means that this goal has been achieved. Following the VoicePrivacy Challenge [25], we use it to refer to the task, even when the technique has failed.

with some configuration parameters, and outputs a new speech signal or some kind of derived representation. The transformed utterances from one or more speakers form a *public speech* dataset that is processed by a third-party *user* for, e.g., ASR training/decoding or any other downstream task.

Given unprocessed or anonymized utterances from a known speaker, an *attacker* attempts to find which anonymized utterances in the public dataset are spoken by this speaker [15, 14]. Formally, an attacker has access to two sets of utterances: A (*enrollment/found data*) and B (*trial/public speech*), but knows the corresponding speakers in A only. The attacker designs a linkage function $LF(a, b)$ that outputs a score for any $a \in A$ and $b \in B$. Typically, this score is a similarity score obtained through a speaker verification system. The attacker then makes a decision (same vs. different) based on this score.

Anonymization techniques must achieve a suitable privacy/ utility trade-off. Utility is measured by the performance of the desired downstream task(s), e.g., the word error rate of an ASR system or the intelligibility for a human listener. Different privacy metrics exist in the literature.

3. Privacy Metrics

We describe three candidate privacy metrics, which model the attacker's decision making process or the score distribution.

3.1. Equal Error Rate (EER)

The EER is the classical metric used in speaker recognition. It assumes a threshold-based decision on the score. If $LF(a, b)$ is greater than a certain threshold t, the two utterances a and b are considered to be mated. Two types of errors can be made: false alarms with rate $P_{\text{fa}}(t)$, and misses with rate $P_{\text{miss}}(t)$. The EER is the error rate corresponding to the threshold t^* for which the two types of errors are equally likely:

$$\text{EER} = P_{\text{miss}}(t^*) = P_{\text{fa}}(t^*). \tag{1}$$

3.2. Log-Likelihood-Ratio Cost Function C_{llr} and $C_{\text{llr}}^{\text{min}}$

C_{llr} is also a common speaker recognition metric [12]. It is *application-independent* in the sense that it pools across all possible costs for false alarm vs. miss errors, and all possible priors for mated vs. non-mated trials. Let M (resp., $\overline{M}$) be the set of mated (resp., non-mated) trials and $|M|$ (resp., $|\overline{M}|$) its cardinality. Denoting by $\text{llr}(p)$ be the log-likelihood ratio for trial $p = (a, b)$, C_{llr} is defined as

$$C_{\text{llr}} = \frac{1}{\log 2}\left[\frac{1}{|M|}\sum_{p\in M}\log\left(1+e^{-\text{llr}(p)}\right) + \frac{1}{|\overline{M}|}\sum_{p\in\overline{M}}\log\left(1+e^{\text{llr}(p)}\right)\right]. \tag{2}$$

C_{llr} assesses the overall detection which includes both discrimination and calibration. In practice, discrimination alone is more relevant as a privacy metric. To measure it, a derived metric called $C_{\text{llr}}^{\text{min}}$ can be computed by optimal calibration of the scores $LF(p)$ into log-likelihood ratios using a monotonic rising transformation. This transformation is found via the Pool Adjacent Violators algorithm (PAV), see [17] for details.

3.3. Linkability

A linkability metric was proposed in [13] for biometric template protection systems. This metric can be generalized for any two sets of items. Denoting by H (resp., $\overline{H}$) the binary variable expressing whether two random utterances a and b are mated (resp., non-mated), the local linkability metric for a score $s = LF(a, b)$ is defined as $p(H \mid s) - p(\overline{H} \mid s)$. When the local metric is negative, an attacker can deduce with some confidence that the two utterances are from different speakers. The authors of [13] argued that the local metric should estimate the strength of the link described by a score rather than measure how much a score describes non-mated relationships. Therefore they propose a clipped version of the difference:

$$D_{\leftrightarrow}(s) = \max(0, p(H \mid s) - p(\overline{H} \mid s)). \tag{3}$$

The global linkability metric $D_{\leftrightarrow}^{\text{sys}}$ is the mean value of $D_{\leftrightarrow}(s)$ over all mated scores:

$$D_{\leftrightarrow}^{\text{sys}} = \int p(s \mid H) \cdot D_{\leftrightarrow}(s)\, ds.$$

In practice, $D_{\leftrightarrow}(s)$ is rewritten as $(2\cdot\omega\cdot\text{lr}(s))/(1+\omega\cdot\text{lr}(s))-1$ where the likelihood ratio $\text{lr}(s)$ is $p(s \mid H)/p(s \mid \overline{H})$ and the prior probability ratio ω is $p(H)/p(\overline{H})$, and $p(s \mid H)$ and $p(s \mid \overline{H})$ are computed via one-dimensional histograms.

3.4. Comparison of the Metrics

Based on the above definitions, we already note that the three metrics do not provide the same information. Both the EER and $C_{\text{llr}}^{\text{min}}$ measure the probability of error of an attacker that makes decisions based on a threshold on the linkage function (one particular threshold for EER and all possible ones for $C_{\text{llr}}^{\text{min}}$). Linkability measures something different: it evaluates how different the distributions of mated vs. non-mated scores are. There is no attacker making a decision and there is no threshold or, from another perspective, the best possible *oracle* attacker (not necessarily threshold-based) is assumed. In addition, if we consider how general are the metrics, on the one hand $C_{\text{llr}}^{\text{min}}$ is a direct extension of the EER as it does not focus on one single threshold. On the other hand, $D_{\leftrightarrow}^{\text{sys}}$ is evaluated over all the encountered mated scores. In the next section, we provide experimental examples that highlight the differences of information provided and generality of the metrics.

4. Exhibiting Differences and Blindspots through Simulation

We design two experiments over simulated scores in order to exhibit the differences between the metrics. The first experiment relies on discrete scores to highlight the lack of generality of the EER. The second experiment relies on Gaussian distributed scores to exhibit the differences between $C_{\text{llr}}^{\text{min}}$ and linkability. All of the metrics are integrated in the Voice Privacy Challenge 2020[2] and we developed an easy to use toolkit[3]

4.1. Discrete Scores

Let us assume that there are 8 trials $p_1, \ldots, p_8$ and that the score for the i-th trial is given by the integer $LF(p_i) = i$. The values of EER and $C_{\text{llr}}^{\text{min}}$ vary with the label (mated vs. non-mated) of each trial. In Table 1, we show 3 particular cases where only the labels of the last three trials (associated with scores 6, 7, and 8) change. We notice that this has an effect on $C_{\text{llr}}^{\text{min}}$ but not on the EER. This is because the EER searches for a single

[2] https://www.voiceprivacychallenge.org/#Soft

[3] https://gitlab.inria.fr/magnet/anonymization˙metrics

Table 1: $C_{\text{llr}}^{\text{min}}$ *and* EER *with discrete scores in* $\{1, \ldots, 8\}$. H *(resp.* $\overline{H}$*) denote mated (resp. non-mated) scores.*

Score	1	2	3	4	5	6	7	8	$C_{\text{llr}}^{\text{min}}$	EER
Case 1	$\overline{H}$	$\overline{H}$	H	$\overline{H}$	H	$\overline{H}$	H	H	0.50	0.25
Case 2	$\overline{H}$	$\overline{H}$	H	$\overline{H}$	H	H	$\overline{H}$	H	0.59	0.25
Case 3	$\overline{H}$	$\overline{H}$	H	$\overline{H}$	H	H	H	$\overline{H}$	0.65	0.25

threshold of the linkage function while $C_{\text{llr}}^{\text{min}}$ averages over all possible thresholds that the attacker might choose. We also notice that the EER indicates a privacy of 0.25 that is half of the best achievable privacy (0.5), while $C_{\text{llr}}^{\text{min}}$ increases from half of the best achievable privacy (0.5 over 1) to higher values (0.69).

4.2. Gaussian Scores

Since $D_{\leftrightarrow}^{\text{sys}}$ relies on density estimation, we now generate Gaussian distributed scores to compare $D_{\leftrightarrow}^{\text{sys}}$ and $C_{\text{llr}}^{\text{min}}$. We consider three Gaussians: $G_1 \sim \mathcal{N}(1, \sigma_1)$, $G_2 \sim \mathcal{N}(2, \sigma_2)$ and $G_3 \sim \mathcal{N}(3, \sigma_3)$. Each Gaussian G_i is used to sample either mated or non-mated scores according to a key $k_i \in \{H, \overline{H}\}$. In total, we have four different cases depending on the values of (k_1, k_2, k_3): *Mated higher* for $(\overline{H}, \overline{H}, H)$ or $(\overline{H}, H, H)$; *Non-mated higher* for $(H, \overline{H}, \overline{H})$ or $(H, H, \overline{H})$; *Mated in-between* for $(\overline{H}, H, \overline{H})$; *Non-mated in-between* for $(H, \overline{H}, H)$. We sample from those three distributions in order to obtain $5,000$ mated and $5,000$ non-mated scores. Multiple standard deviations are chosen to obtain different degrees of overlap between the distributions: $(\sigma_1, \sigma_2, \sigma_3) \in \{0.1, 0.5, 1, 1.5\}^3$.

The results are presented in Fig. 2. We consider that $C_{\text{llr}}^{\text{min}}$ and $D_{\leftrightarrow}^{\text{sys}}$ are equivalent when $C_{\text{llr}}^{\text{min}}$ is equal to $1 - D_{\leftrightarrow}^{\text{sys}}$ (diagonal line). The two metrics agree to a large extent only when the mated scores are higher. When the non-mated scores are higher, $C_{\text{llr}}^{\text{min}}$ is always close to 1 while $D_{\leftrightarrow}^{\text{sys}}$ varies depending on the overlap between the distributions. In the two remaining cases when the mated scores are surrounded by the non-mated scores or vice-versa, $C_{\text{llr}}^{\text{min}}$ is lower-bounded by 0.6 and the two metrics do not agree on the strength of anonymization. This is explained by the fact that threshold-based decision is meaningful in the *mated higher* case and its performance is then strongly related to the overlap between distributions, while it fails partially or totally in the three other cases.

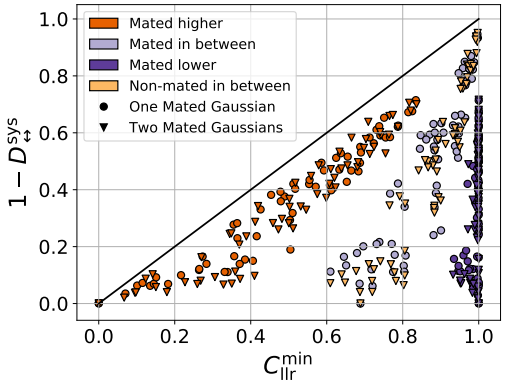

Figure 2: $C_{\text{llr}}^{\text{min}}$ *vs.* $1 - D_{\leftrightarrow}^{\text{sys}}$ *on simulated Gaussian scores.*

To illustrate why this an issue and how this may happen in practice, in Figure 3, we draw (simulated) x-vectors for multiple utterances of two speakers, which have all been anonymized by mapping them to another (pseudo) speaker's voice. Each utterance of speaker A has been randomly mapped to the left or the right cluster, while the utterances of speaker B have been mapped to the center cluster. The resulting score distributions match the *non-mated in-between* case above. As expected, the two metrics strongly disagree: $D_{\leftrightarrow}^{\text{sys}} = 0.99$ (low privacy) and $C_{\text{llr}}^{\text{min}} = 0.81$ (high privacy). While this situation is unlikely to occur with unprocessed data (scores are then expected to match the *mated higher* case), it becomes likely once the utterances have been anonymized and the anonymization design choices (see [18] for example choices) result in multimodal score distributions.

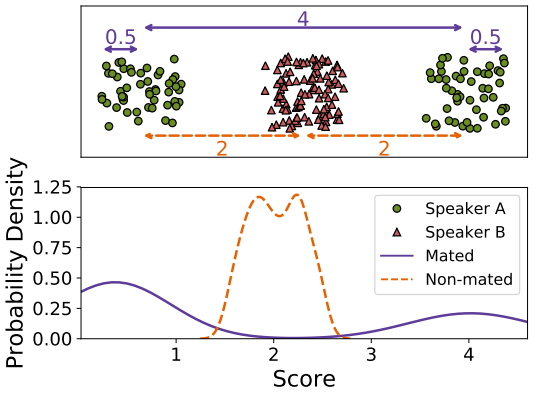

Figure 3: *Simulated 'non-mated in-between' data. Top: x-vectors visualized in 2D. Bottom: resulting score distributions.*

5. Evaluation on Real Anonymized Speech

In order to further compare $D_{\leftrightarrow}^{\text{sys}}$, $C_{\text{llr}}^{\text{min}}$ and the EER, we conduct a second experiment on real speech data. In the following, we present the dataset, the anonymization techniques and the attackers considered. Then we discuss the results.

The experiment is conducted on LibriSpeech [19]. The *train-clean-460* set (∼1k speakers, ∼130k utterances and 460 hours of speech) is used to train the x-vector model and the probabilistic linear discriminant analysis (PLDA) model with Kaldi [20]. Part of the *test-clean* set (40 speakers, 1,496 utterances) is anonymized to form the *trial/public* data. The remaining part (29 speakers, 438 utterances) is considered as unprocessed *enrollment/found* data.

5.1. Anonymization Techniques and Target Selection

We use the following four anonymization techniques. Except for the first one, these are voice conversion techniques which map the input (source) signal to another (target) speaker's voice.

VoiceMask (VM) [21] is a frequency warping method. It has two parameters, α and β, they are chosen uniformly at random from a predefined range which is found to produce intelligible speech while perceptually concealing the speaker identity.

VTLN-based VC [22] clusters each speaker's data into unsupervised pseudo-phonetic classes. For each source speaker class, the closest target speaker class is found and the corresponding warping parameters are applied to the input signal.

The third approach is based on **disentangled representation (DAR)** [23, 24]. It uses a speaker encoder and a content

encoder to separate speaker and content information and replace the source speaker information by that of the target speaker.

Finally, the primary baseline of the **VoicePrivacy Challenge 2020 (VPC)** [25] uses a neural synthesizer [9, 26] to synthesize speech given the target x-vector and fundamental frequency and bottleneck features extracted from the source.

VTLN and DAR require speakers to be anonymized using target speakers from a given pool. Following [14], we evaluate three different target selection strategies: (1) CONST: all utterances of all source speakers are mapped to one single target speaker; (2) PERM: each source speaker has all her utterances mapped to one specific target speaker; (3) RAND: each utterance of each speaker is mapped to a random target speaker. Rather than an actual target speaker, VPC constructs a target x-vector by averaging several x-vectors from the pool.

5.2. Attacker Knowledge and Linkage Function

Following [14], we also consider different attackers based on their knowledge about the anonymization. (1) Ignorant: the attacker has no knowledge of the anonymization and uses unprocessed enrollment data; (2) Informed: the attacker has complete knowledge of the anonymization technique including the target speakers, and he/she processes the enrollment data accordingly; (3) Semi-informed: the attacker knows the anonymization technique and the target selection strategy but not the particular target speaker selected for a given source speaker, and she processes the enrollment data accordingly. The attacker performs linkage attacks by computing the x-vectors of a trial utterance and an enrollment utterance and comparing them using one of three linkage functions: PLDA affinity, cosine distance, or Euclidean distance. This results in a total of 72 combinations of anonymization techniques, target selection strategies, attacker knowledge levels, and linkage functions.

5.3. Results

Figures 4 and 5 compare the resulting metrics, where each dot corresponds to one of the 72 combinations above. The comparison between the EER and $C_{\text{llr}}^{\text{min}}$ (Fig. 4) shows a clear relation between the two metrics. In some cases the EER is stable and $C_{\text{llr}}^{\text{min}}$ varies a little bit but not significantly so. Regarding the comparison between $D_{\leftrightarrow}^{\text{sys}}$ and $C_{\text{llr}}^{\text{min}}$, we see a clear difference between Fig. 5 on real data and Fig. 2 on simulated Gaussian scores: on real data, the two metrics follow a clear relation.

These results can be explained by the fact that, with few exceptions, the score distributions for the specific target selection and attack strategies considered here fall into the *mated higher* case, as can be seen from the colors associated with the dots. It is however likely that advanced target selection strategies aiming for score distributions akin to Fig. 2 will be developed in the near future, as these would provide an advantage against attackers making threshold-based decisions. For that reason, we believe $D_{\leftrightarrow}^{\text{sys}}$ should be privileged as a privacy metric, since it provides very similar results to established metrics with current target selection and attack strategies, while being more robust to advanced strategies that will likely be developed soon.

6. Conclusion

In this study, we compare three metrics to assess the effectiveness of anonymization: the EER, the application-independent log-likelihood-ratio min cost function $C_{\text{llr}}^{\text{min}}$, and the linkability $D_{\leftrightarrow}^{\text{sys}}$. The EER and $C_{\text{llr}}^{\text{min}}$ assume that the attacker makes threshold-based decisions on the linkage score, while $D_{\leftrightarrow}^{\text{sys}}$ implicitly models a more powerful, non-threshold-based *oracle* attacker. The comparison on real speech data processed via 4 anonymization techniques with different target selection strategies and with 9 attackers suggests that these metrics behave similarly. Yet, experiments on simulated data highlight fundamental differences. Specifically, the EER may yield a fixed value for situations involving different levels of privacy correctly captured by $C_{\text{llr}}^{\text{min}}$, and $C_{\text{llr}}^{\text{min}}$ becomes less informative than $D_{\leftrightarrow}^{\text{sys}}$ when the mated scores are lower or interleaved with non-mated scores. While such situations were unlikely to occur in the field of speaker verification, which involves unprocessed speech data, we expect them to become frequent in the field of anonymization when more advanced target selection and attack strategies are built. For this reason, we advocate for the use of $D_{\leftrightarrow}^{\text{sys}}$ as a robust privacy metric capable of handling both current approaches and future developments in this field.

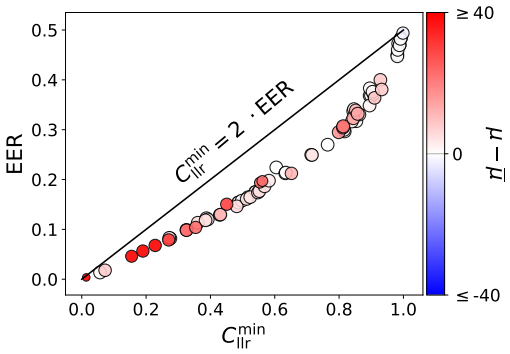

Figure 4: $C_{\text{llr}}^{\text{min}}$ *vs.* EER *on real data. The color scale* $\mu - \bar{\mu}$ *is the difference of the means of mated and non-mated scores.*

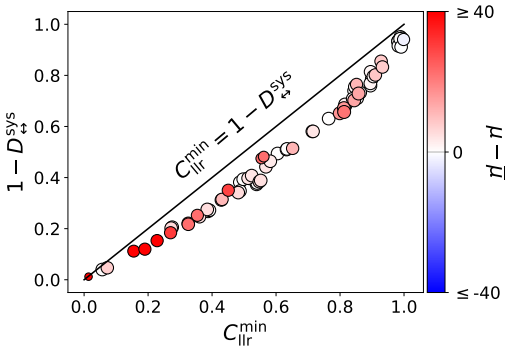

Figure 5: $C_{\text{llr}}^{\text{min}}$ *vs.* $1 - D_{\leftrightarrow}^{\text{sys}}$ *on real data. The color scale* $\mu - \bar{\mu}$ *is the difference of the means of mated and non-mated scores.*

7. Acknowledgments

This work was supported by the French National Research Agency under project DEEP-PRIVACY (ANR-18-CE23-0018) and by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 825081 COMPRISE (`https://www.compriseh2020.eu/`). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

8. References

[1] A. Nautsch, A. Jimenez, 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 and Language*, vol. 58, pp. 441–480, 2019.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

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

[4] K. Hashimoto, J. Yamagishi, and I. Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5500–5504.

[5] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X.-Y. Li, Y. Wang, and Y. Deng, "Voicemask: Anonymize and sanitize voice input on mobile devices," *arXiv preprint arXiv:1711.11460*, 2017.

[6] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification via voice transformation," in *2009 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009, pp. 529–533.

[7] M. Pobar and I. Ipšić, "Online speaker de-identification using voice transformation," in *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2014, pp. 1264–1267.

[8] F. Bahmaninezhad, C. Zhang, and J. H. L. Hansen, "Convolutional neural network based speaker de-identification." in *Odyssey*, 2018, pp. 255–260.

[9] 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 *10th ISCA Speech Synthesis Workshop (SSW)*, 2019, pp. 155–160.

[10] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-preserving adversarial representation learning in ASR: Reality or illusion?" in *Interspeech*, 2019, pp. 3700–3704.

[11] ISO/IEC 19795-1:2006, "Information Technology — Biometric performance testing and reporting — Part 1: Principles and framework," 2006.

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

[13] M. Gomez-Barrero, J. Galbally, C. Rathgeb, and C. Busch, "General framework to evaluate unlinkability in biometric template protection systems," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1406–1420, 2017.

[14] 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 *2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 2802–2806.

[15] J. Qian, F. Han, J. Hou, C. Zhang, Y. Wang, and X.-Y. Li, "Towards privacy-preserving speech data publishing," in *2018 IEEE Conference on Computer Communications (INFOCOM)*, 2018, pp. 1079–1087.

[16] 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 *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2802–2806.

[17] D. A. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification I: Fundamentals, Features, and Methods*, 2007, pp. 330–353.

[18] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design choices for x-vector based speaker anonymization," in *Interspeech*, submitted.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[21] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *16th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2018, pp. 82–94.

[22] D. Sundermann and H. Ney, "VTLN-based voice conversion," in *3rd IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2003, pp. 556–559.

[23] J.-C. Chou and H.-Y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," in *Interspeech*, 2019, pp. 664–668.

[24] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6924–6932.

[25] 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 *Interspeech*, submitted. [Online]. Available: https://hal.inria.fr/hal-02562199

[26] X. Wang and J. Yamagishi, "Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis," in *10th ISCA Speech Synthesis Workshop (SSW)*, 2019, pp. 1–6.