Privacy and utility of x-vector based speaker anonymization

Brij Mohan Lal Srivastava, Mohamed Maouche, Md Sahidullah, Emmanuel Vincent, Aurélien Bellet, Marc Tommasi, Natalia Tomashenko, Xin Wang, Junichi Yamagishi

Privacy and utility of x-vector based speaker anonymization

Brij Mohan Lal Srivastava, Mohamed Maouche, Md Sahidullah, Emmanuel Vincent, Aurélien Bellet, Marc Tommasi, Natalia Tomashenko, Xin Wang, and Junichi Yamagishi

***Abstract*—We study the scenario where individuals (*speakers*) contribute to the publication of an anonymized speech corpus. Data *users* leverage this public corpus for downstream tasks, e.g., training an automatic speech recognition (ASR) system, while *attackers* may attempt to de-anonymize it using auxiliary knowledge. Motivated by this scenario, speaker anonymization aims to conceal speaker identity while preserving the quality and usefulness of speech data. In this article, we study x-vector based speaker anonymization, the leading approach in the VoicePrivacy Challenge, which converts the speaker's voice into that of a random pseudo-speaker. We show that the strength of anonymization varies significantly depending on how the pseudo-speaker is chosen. We explore four design choices for this step: the distance metric between speakers, the region of speaker space where the pseudo-speaker is picked, its gender, and whether to assign it to one or all utterances of the original speaker. We assess the quality of anonymization from the perspective of the three actors involved in our threat model, namely the speaker, the user and the attacker. To measure privacy and utility, we use respectively the linkability score achieved by the attackers and the decoding word error rate achieved by an ASR model trained on the anonymized data. Experiments on LibriSpeech show that the best combination of design choices yields state-of-the-art performance in terms of both privacy and utility. Experiments on Mozilla Common Voice further show that it guarantees the same anonymization level against re-identification attacks among 50 speakers as original speech among 20,000 speakers.**

***Index Terms*—speaker anonymization, privacy, linkability, voice conversion**

I. Introduction

SPEECH is a rich source of personal information including sensitive attributes such as identity [1], accent [2], or pathological condition [3]. When the speaker's goal is not biometric authentication but some other voice-based interaction, for example, exchanging with voice assistants or customer helplines, speaker anonymization is desirable. Indeed, the availability of efficient techniques to infer these attributes from speech as well as recent advances in voice cloning [4] pose severe privacy and security risks [5].

Throughout this article, we consider the following threat model. Given a public dataset of anonymized speech contributed by several speakers, an attacker records/finds a sample of speech of a speaker and attempts to find which utterances in the anonymized dataset are spoken by this speaker, possibly leveraging some knowledge about the anonymization method. A good speaker anonymization method must defeat such *linkage attacks* by concealing speaker identity, while preserving the utility of speech for data *users* as measured for instance by the perceived speech naturalness and/or the performance of downstream tasks such as training an automatic speech recognition (ASR) system.[1] Figure 1 shows the three actors involved in this model, namely the *speaker*, the *attacker* and the *user*, along with their actions. The goals of the speaker and the user are intimately linked, while the attacker operates independently.

Speaker anonymization methods have been studied for just over a decade. They include noise addition [7], speech transformation [8], voice conversion [9]–[11], speech synthesis [12], [13], or adversarial learning [14]. In this study, we focus on x-vector based anonymization [13], [15], which converts the original speaker's voice into that of a *target (pseudo-)speaker*, due to the naturalness of its output and its promising results so far. In order to implement and assess such an anonymization method, the following questions arise from the speaker's and user's perspectives: Q1: *How to optimally choose and assign the target pseudo-speaker?* Q2: *How well is utility preserved?* Q3: *How much residual speaker information remains?* Furthermore, the attacker must address the following questions: Q4: *Can privacy protection be defeated using some knowledge of the anonymization method?* Q5: *How does the number of possible speakers affect the re-identification performance?*

In this article, we extend the two target pseudo-speaker generation strategies in [13] (fully random, or at a fixed distance from the original as measured by cosine distance between x-vectors) into a whole family of strategies based on four design choices: the distance metric between x-vectors, the region of x-vector space where the pseudo-speaker is picked, its gender, and whether to assign it to one or all utterances of the original speaker. Our experiments suggest an optimal combination of design choices to balance privacy and utility (answering Q1). We train and/or evaluate ASR models on anonymized speech to assess utility (answering Q2). We show that some speaker information remains in the pitch sequence and apply two different pitch transformation techniques to remove it (answering Q3). We conduct these experiments for three types of attackers [16], where stronger attackers

B. M. L. Srivastava, M. Maouche, A. Bellet and M. Tommasi are with Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France (email: brij.srivastava@inria.fr, mohamed.maouche@inria.fr, aurelien.bellet@inria.fr, marc.tommasi@univ-lille.fr). M. Sahidullah and E. Vincent are with Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France (email: md.sahidullah@inria.fr, emmanuel.vincent@inria.fr). N. Tomashenko is with Laboratoire Informatique d'Avignon (LIA), Avignon Université, France (email: natalia.tomashenko@univ-avignon.fr). X. Wang and J. Yamagishi are with the National Institute of Informatics, Tokyo, Japan (email: wangxin@nii.ac.jp, jyamagis@nii.ac.jp).

[1] Legally speaking, the term "anonymization" refers to a method that fully achieves this goal. Following [6], we use it in a broader sense to refer to a method that aims to achieve this goal, even when it has failed to do so.

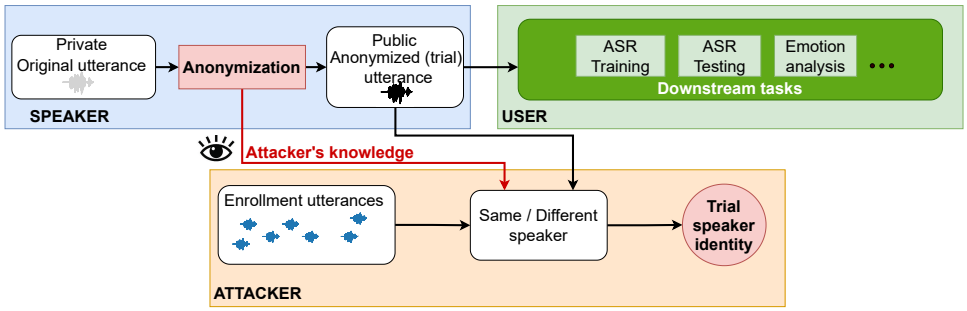


Fig. 1. Considered threat model. *Speakers* anonymize speech to conceal their identity before publication; *attackers* use biometric technology and knowledge of the anonymization method to re-identify it; *users* (e.g., speech technology companies) use the published data for downstream tasks such as ASR training.

have more knowledge about the anonymization method (answering Q4). Finally, we conduct additional experiments with more than 20,000 possible speakers (answering Q5). These contributions significantly extend our preliminary study [17], which provided less detail, did not include utterance- vs. speaker-level target assignment and pitch transformation, did not evaluate privacy against the strongest (*Semi-Informed*) attacker or with a large number of possible speakers, and did not evaluate utility for ASR training.

The structure of the article is as follows. In Section II, we introduce x-vector based anonymization and position it among other related anonymization methods. Section III presents the four considered design choices. Sections IV and V describe the main experimental setup and the corresponding results. Experiments with more speakers are conducted in Section VI. We conclude in Section VII.

II. X-VECTOR BASED ANONYMIZATION

Speaker anonymization aims to conceal the speaker's identity from a speech signal such that it cannot be used to clone the speaker's voice or to re-identify the speaker through automatic speaker verification (ASV) or automatic speaker identification (ASI). Early methods based on voice conversion required both the original and the target speaker to be part of the training set for the voice conversion system. Jin et al. [9] convert all speakers into a single target. Bahmaninezhad et al. [11] convert a given speaker into the average of all speakers of the same gender. Pobar and Ipšić [10] pre-train a set of speaker transformations and identify the speaker at test time to select one of the corresponding transformations. These methods are hardly applicable in practice since, in the context of anonymization, the amount of speech from the original speaker is often limited to one utterance. To relax this constraint, Magariños et al. [18] find the closest source speaker in the training set and apply one of the corresponding transformations, while Justin et al. [12] transcribe speech into a diphone sequence and re-synthesize it using a single target. Although they do not require the original speaker to belong to the training set anymore, these two methods suffer from three limitations. First, they still result in a limited set of target speakers or speaker transformations, which prevents the original speaker from choosing an arbitrary unseen speaker as the target. Second, using a real speaker's voice as the target raises ethical concerns. Third, the phonetic transcription step in [12] is error-prone. This motivates the objective of converting the original speaker's voice into an arbitrary, imaginary *pseudo-speaker*'s voice without relying on a transcription step. Speaker embeddings such as x-vectors [19] (a low-dimensional representation extracted from an intermediate layer of an ASI model) provide the continuous representation needed to define and generate such pseudo-speakers.

Fang et al. [13] address this objective using a speaker-independent speech synthesis system. They select x-vectors within an external pool of speakers and average them to obtain a target *pseudo-speaker* x-vector. This x-vector, along with a representation of the original linguistic and intonation contents, is provided as input to a neural source-filter (NSF) based speech synthesizer [20] to produce anonymized speech.

In the following, we use the anonymization system shown in Fig. 2, that is a variant of the one in [13]. This system represents speaker identity, linguistic content and intonation using x-vectors $\mathbf{v}$,[2] bottleneck (BN) features $\mathbf{B}$ [21] (a low-dimensional representation extracted from an intermediate layer of an ASR model) and pitch sequences $\mathbf{p}$, respectively. It comprises four steps: Step 1 *(Feature extraction)* extracts pitch and BN features and the x-vector from the input signal. Step 2 *(X-vector anonymization)* generates a target x-vector $\mathbf{v}^*$ by averaging N^* candidate x-vectors from an external pool

[2]Following [13], we use raw x-vectors to represent speaker identity instead of x-vectors compressed and rotated by linear discriminant analysis (LDA), as classically done in the context of ASV. Unless the projected dimension is carefully chosen after several experiments, the impact of the LDA transformation on speaker-specific information cannot be ascertained. Hence we defer experiments with LDA-transformed x-vectors to a future study.

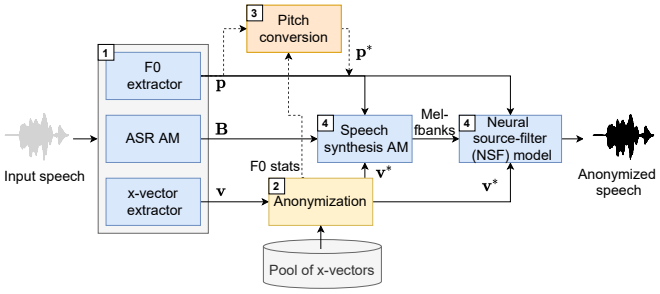


Fig. 2. General architecture of the anonymization system.

of speakers.[3] Step 3 *(Pitch conversion)* is an optional step which receives the target pitch statistics from the anonymization module and transforms the original pitch sequence into $\mathbf{p}^*$. Step 4 *(Speech synthesis)* synthesizes a speech waveform from the anonymized x-vector $\mathbf{v}^*$ and the original $\mathbf{B}$ and $\mathbf{p}$ (or optionally $\mathbf{p}^*$) features using an acoustic model (AM) and the NSF model. With the exception of Step 3 which is new (see Section V-D), this system is identical to the first anonymization baseline for the VoicePrivacy Challenge [6]. We refer to [22] for details on the feature dimensions and the architectures of the models in Steps 1 and 4.

We note that there have been other interesting attempts to generate a target *pseudo-speaker* x-vector for speaker anonymization in the systems submitted to the first VoicePrivacy Challenge. Mawalim et al. [23] modified the significant elements of the source speaker x-vector that were determined using singular value decomposition and variant analysis to anonymize the identity. Perero-Codosero et al. [24] transformed the original x-vector using an autoencoder with adversarial training to suppress speaker, gender and accent information. Turner et al. [25] fitted a Gaussian mixture model based generative model over the external pool of speakers, and then proposed to sample target x-vectors from this model to preserve the distributional properties of x-vectors. Readers are referred to Tomashenko et al. [26] for an in-depth analysis of the objective and subjective evaluation results achieved by the two challenge baselines and the 16 submissions.

III. Anonymization Design Choices

Now given the ability to generate arbitrary external targets in Step 2 (yellow box in Fig. 2), the question arises of which strategy the speaker shall employ to select the candidate x-vectors and achieve a suitable privacy-utility tradeoff. Fang et al. [13] select candidate x-vectors at random within the whole pool or within a fixed interval of distances from the original x-vector. Han et al. [15] select a single target x-vector at random within a maximum distance from the original x-vector. In the following, we expand these initial strategies into a broader range of strategies governed by the choice of the distance metric between x-vectors, the region of x-vector space where the candidates are selected, their gender, and the assignment of the resulting target x-vector to one or all utterances of the original speaker. These four design choices, which are illustrated in Fig. 3, are detailed below. For the sake of focus, we do not explore other design choices such as the size or the diversity of the anonymization pool.

A. Distance metric: cosine vs. PLDA

To design advanced candidate selection strategies, the speaker must first choose a distance metric which dictates the properties of the x-vector space. We compare two such metrics.

The first one is the cosine distance, which was used by [13]. For a pair of x-vectors ω_i and ω_j, it is defined as

$$d_{\cos}(\omega_i, \omega_j) = 1 - \frac{\omega_i \cdot \omega_j}{||\omega_i||_2 ||\omega_j||_2}. \quad (1)$$

[3]There is no guarantee that averaging produces a valid x-vector, but all our experiments show that the synthesized anonymized speech is of good quality.

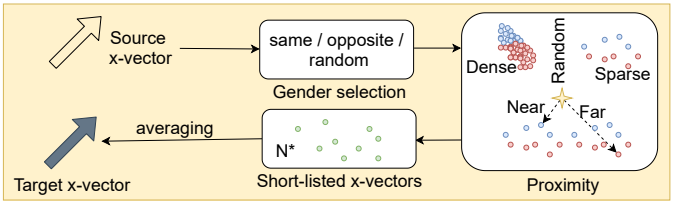


Fig. 3. Zoomed-in view of the x-vector anonymization step in Fig. 2 showing the design choices for the generation of the target x-vector.

The second metric is based on probabilistic linear discriminant analysis (PLDA) [27], that is the log-likelihood ratio of the hypotheses that ω_i and ω_j belong to the same speaker ($\mathcal{H}_s$) vs. different speakers ($\mathcal{H}_d$). Previous studies [28] have shown that PLDA yields state-of-the-art performance as the x-vector similarity metric in the context of ASV. This is attributed to its formulation which estimates the factorized within-speaker and between-speaker variability in speaker space, making it a superior metric even for short utterances [29]. More specifically, PLDA models x-vectors ω as $\omega = m + Vy + Dz$, where m is the center of the x-vector space, the columns of V capture speaker variability (eigenvoices) with y depending only on the speaker, and the columns of D encode channel variability (eigenchannels) with z varying from one recording to another. The parameters m, V and D are trained on x-vectors extracted using the x-vector extractor in Step 1 from the dataset used to train that extractor itself (see [22] for details on this dataset). The log-likelihood ratio score

$$d_{\text{PLDA}}(\omega_i, \omega_j) = \log \frac{p(\omega_i, \omega_j|\mathcal{H}_s)}{p(\omega_i, \omega_j|\mathcal{H}_d)} \quad (2)$$

can be computed in closed form [30]. We propose to use $-d_{\text{PLDA}}$ as the "distance" between a pair of x-vectors.

B. Proximity: random, far, near, dense, or sparse

We propose three alternative criteria resulting in five different "proximity" choices to restrict the region of x-vector space from which candidate x-vectors are selected.

1) Random: The simplest candidate x-vector selection strategy is to select N^* x-vectors with a given gender uniformly at random from the pool. Note that this strategy does not allow us to choose particular regions of interest in the x-vector space.

2) Far/near: Alternatively, the chosen distance metric can be used to find candidate x-vectors which resemble most (*near*) or least (*far*) the original speaker $\mathbf{v}$. In essence, we rank all the x-vectors in the pool in increasing order of their distance from $\mathbf{v}$ and select either the top N (*near*) or the bottom N (*far*). To introduce some randomness, $N^* < N$ x-vectors are selected out of these N uniformly at random.

3) Dense/sparse: Another alternative is to identify clusters of x-vectors in the pool and rank them based on their cardinality. We construct these clusters using the Affinity Propagation [31] algorithm (see detailed procedure in Section IV-B). We filter out the cluster which is closest to the original speaker, then randomly select one cluster among those with most

(*dense*) or least (*sparse*) members.[4] We then randomly select half of the members of that cluster.

In all five cases, the selected candidate x-vectors are averaged to obtain the target (pseudo-speaker) x-vector $\mathbf{v}^*$.

C. Gender selection: same, opposite, or random

In practice, instead of applying one of these five proximity choices to the entire speaker pool, we apply it to a gender-dependent pool which consists of either all males or all females of the original pool. We propose three possible gender selection choices: *same* where all speakers in the pool have the same gender as the original speaker; *opposite* where they all have the opposite gender; and *random* where either of the two gender-dependent pools is selected at random. This allows us to avoid averaging candidate x-vectors from both genders with each other, and to assess the impact of gender selection on privacy and utility.

D. Assignment: speaker- or utterance-level

The generation of the anonymized waveform is conditioned upon the x-vector sequence, whose length is equal to the number of frames in the original utterance. All the x-vectors in this sequence are identical to each other to indicate a single pseudo-speaker throughout the utterance. In theory, these x-vectors should also be identical across all utterances spoken by this pseudo-speaker but, according to [32], x-vectors also contain channel, duration, and phonetic information, in addition to speaker and gender. Hence, the x-vectors computed for different utterances may exhibit some variations due to utterance-specific properties. To assess the effect of these variations on privacy and utility, we propose two assignment strategies for the target x-vector: speaker-level (*perm*) or utterance-level (*rand*). In the former case, we average the utterance-level x-vectors of all utterances of the original speaker into a single speaker-level x-vector $\mathbf{v}$, we generate a corresponding target x-vector $\mathbf{v}^*$, and we use it to anonymize all utterances of that speaker. In the latter case, we consider the utterance-level x-vector $\mathbf{v}_u$ for each utterance u of the original speaker, we generate a corresponding target x-vector $\mathbf{v}_u^*$ (using the same distance metric, proximity, and gender across all utterances), and we use it to anonymize that utterance only.

IV. Experimental setup

The VoicePrivacy Challenge [6] assumed that the attacker does not have access to the anonymization system and that the user is unaware that speech has been anonymized. Privacy and utility were consequently assessed using an ASV system and an ASR system trained on *original* (non-anonymized) speech. This resulted in overestimated privacy and underestimated utility with respect to an attacker or a user who have access to the anonymization scheme [16]. Also, the utility for ASR training was not evaluated. Following [16], we advocate for a complete study of the utility/privacy trade-off, which is key to the success of downstream tasks.

[4]Note that the terms *sparse* and *dense* do not directly reflect the density of x-vectors, since they do not take the diameter of the clusters into account. However, we find that this relation holds in practice.

A. Data

The experiments in Section V rely on the same datasets as the VoicePrivacy Challenge. Among the components of the anonymization system, the ASR AM is trained on the *train-clean-100* and *train-other-500* subsets of LibriSpeech [33], the x-vector extractor is trained on VoxCeleb1 [34] and VoxCeleb2 [35], and the speech synthesis AM and NSF model are trained on the *train-clean-100* subset of LibriTTS [36]. The *train-other-500* subset of LibriTTS is used as the external pool of speakers for x-vector anonymization. The development and test sets are built from the *dev-clean* and *test-clean* subsets of LibriSpeech, respectively.[5] Each of these two sets consists of *trial* utterances from 40 speakers and *enrollment* utterances from a subset of 29 speakers (see Section IV-C). Details about the number of male and female speakers and the number of utterances in each dataset can be found in [6].

In Section VI, we employ the same trained models and the same external pool of speakers but we build mutiple test sets from the Mozilla Common Voice [37] English corpus in order to study the attacker's performance against a larger number of enrolled speakers. Following the approach in [38, Appendix A.1], we select 24,610 male speakers with a total speech duration greater than 10 s after removing silent frames using voice activity detection (VAD). All utterances with a signal-to-noise ratio (SNR) above 75 dB are used for enrollement, in the limit of a total duration of 2 min per speaker.[6] The remaining utterances from 20 speakers whose total duration is greater than 5 min are selected for trial. The resulting numbers of utterances and trials are given in Table I.

TABLE I
STATISTICS FOR THE MOZILLA COMMON VOICE ENROLLMENT AND TRIAL SETS AND NUMBER OF TRIALS.

Common Voice-enroll	Number of speakers	24,610
	Number of utterances	320,085
Common Voice-trial	Number of speakers	20
	Number of utterances	4,696
Number of trials	Same-speaker trials	4,696
	Different-speaker trials	115,563,864

B. Algorithm settings

The *dense* and *sparse* anonymization choices are implemented as follows. We use Affinity Propagation [31] to cluster the speakers in the external pool. This non-parametric clustering method determines the number of clusters automatically via a message passing algorithm. Two parameters govern the number of clusters: the *preference* parameter assigns a higher weight to samples which are likely candidates for centroids, and the *damping factor* weights the so-called responsibility and availability messages. In our experiments, equal *preference* is assigned to all samples and the *damping factor* is set to

[5]The VoicePrivacy Challenge involves development and evaluation sets built from LibriSpeech and VCTK. Due to space limitations, we focus on LibriSpeech.

[6]The SNR was computed using the WADA-SNR [39] algorithm available at https://gist.github.com/johnmeade/d8d2c67b87cda95cd253f55c21387e75.

0.5. Out of 1,160 speakers in the pool, 80 clusters are found, including 46 male and 34 female. The number of speakers per cluster ranges from 6 to 36. Candidate x-vector selection is achieved by picking either the 10 clusters with least members (*sparse*) or the 10 clusters with most members (*dense*). The remaining clusters are ignored. During anonymization, one of the 10 clusters is selected at random and 50% of its members are averaged to produce the target x-vector.

C. Privacy evaluation

As explained in Section I, privacy protection can be seen as a contest between two entities: a *speaker* who publishes anonymized utterances, and an *attacker* who attempts to uncover the speaker's identity by comparing these utterances with utterances whose speaker is known. Following the classical ASV terminology adopted in the VoicePrivacy Challenge, these are called *trial* and *enrollment* utterances, respectively, and each such comparison is called a *trial*. The attacker has full control over the enrollment set and the speaker identities within it. Hence he/she may use some knowledge about the anonymization scheme to transform the enrollment data and reduce the mismatch with the trial data. To assess the strength of anonymization against attackers with increasing knowledge, we perform the evaluation in four scenarios:

- *Baseline*: The speaker does not perform any anonymization. The attacker uses original speech for enrollment and an ASV system trained on original speech. This offers the lowest possible privacy protection.
- *Ignorant*: The speaker anonymizes his/her speech, unbeknownst to the attacker who still uses original speech for enrollment and an ASV system trained on original speech.
- *Lazy-Informed*: The speaker anonymizes his/her speech. The attacker anonymizes the enrollment data using the same anonymization system and the same design choices. However, he/she is not aware of the random numbers drawn by the speaker to obtain the *random* target gender (Section III-C) or the candidate x-vectors (Section III-B). Hence, different pseudo-speakers are assigned to the trial and enrollment utterances of a given speaker.[7]
- *Semi-Informed* The speaker anonymizes his/her speech. The attacker anonymizes the enrollment data using the same system and design choices. In addition, he/she anonymizes the training dataset for the ASV system and re-trains it. This scenario is the one in which the speaker is most "vulnerable" despite anonymization, hence we consider it as the most trustworthy assessment of privacy.[8]

In Section V, privacy is assessed in terms of the *linkability* [40], [41], denoted as $D_{\leftrightarrow}^{\text{sys}}$, achieved by an x-vector-PLDA ASV system trained on the *train-clean-360* subset of LibriSpeech (anonymized in the *Semi-Informed* scenario, original otherwise). This metric computes the overlap between the distributions of PLDA scores of same-speaker and different-speaker trials. It behaves similarly to the equal error rate (EER) and log-likelihood ratio cost function [42] used in the VoicePrivacy Challenge, but it does not rely on any restrictive assumption (e.g., threshold-based decision) which makes it a more trustworthy metric [41]. For the sake of reproducibility, we use the same set of trials as in [6].[9] Linkability varies from 0 to 1, where lower values indicate higher privacy. The 95% confidence interval on the linkability computed via the jackknife method [43] varies from ± 0.0001 to ± 0.0002.

Formally, the local linkability metric $D_{\leftrightarrow}(\theta)$ for two random utterances i and j with a score $\theta = d_{\cos}(\omega_i, \omega_j)$ or $\theta = -d_{\text{PLDA}}(\omega_i, \omega_j)$ is defined as $p(\mathcal{H}_s \mid \theta) - p(\mathcal{H}_d \mid \theta)$. When the local metric is negative, an attacker can deduce with some confidence that the two utterances are from different speakers. The authors in [40] argued that the local metric should estimate the strength of the link described by a score rather than measure how much a score describes different-speaker relationships. Therefore they proposed a clipped version of the difference:

$$D_{\leftrightarrow}(\theta) = \max(0, p(\mathcal{H}_s \mid \theta) - p(\mathcal{H}_d \mid \theta)). \tag{3}$$

The global linkability metric $D_{\leftrightarrow}^{\text{sys}}$ is then defined as the mean value of $D_{\leftrightarrow}(\theta)$ over all same-speaker scores:

$$D_{\leftrightarrow}^{\text{sys}} = \int p(\theta \mid \mathcal{H}_s) \cdot D_{\leftrightarrow}(\theta)\ d\theta.$$

In practice, $D_{\leftrightarrow}(\theta)$ is rewritten as $(2 \cdot \alpha \cdot \text{lr}(\theta))/(1 + \alpha \cdot \text{lr}(\theta)) - 1$ where the likelihood ratio $\text{lr}(\theta)$ is $p(\theta \mid \mathcal{H}_s)/p(\theta \mid \mathcal{H}_d)$ and the prior probability ratio α is $p(\mathcal{H}_s)/p(\mathcal{H}_d)$, and $p(\theta \mid \mathcal{H}_s)$ and $p(\theta \mid \mathcal{H}_d)$ are computed via one-dimensional histograms.

In Section VI, we also evaluate the average *rank* of the true speaker and the *top-k precision* achieved for closed-set ASI. Instead of training speaker classification systems on subsets of Common Voice, which would overfit the speakers therein, we compute the PLDA scores between each trial utterance and all enrollment utterances (one per speaker, including the true speaker) using the same x-vector and PLDA models as in Section V and sort them in decreasing order. The higher the rank and the lower the top-k precision, the higher the privacy.

D. Utility evaluation

In Section V-A, we evaluate the utility for ASR decoding in terms of the word error rate (WER) achieved by an ASR system trained on the *train-clean-360* subset of LibriSpeech and applied to the anonymized utterances. In Sections V-B and V-C, we evaluate the utility both for ASR decoding and training in terms of the WER achieved by an ASR system trained either on the original or the anonymized *train-clean-360* dataset and used to decode either original or anonymized speech. For more details on the ASR system architecture, see [6]. A lower WER indicates higher utility. The 95% confidence

[7]The *Ignorant* and *Lazy-Informed* scenarios were called OA and AA in the VoicePrivacy Challenge. The *Lazy-Informed* scenario was also called *Semi-Ignorant* in [17].

[8]The *Informed* scenario in [16] where the attacker is aware of the random numbers drawn by the speaker is not part of our study, since it falls into a security problem rather than just a privacy problem.

[9]As classically assumed in the speaker verification literature, the two speakers in each trial have the same original gender. In practice though, the gender of the original speaker may be unknown to the attacker. Hence, the resulting linkability values can be seen as worst-case values from the speaker's point of view and best-case values from the attacker's point of view.

interval on the WER varies from $\pm 0.2\%$ for the lowest WER values to $\pm 0.4\%$ for the highest ones.

V. Results and Discussion

The design choices introduced in Section III result in 54 possible combinations, among which 48 combinations correspond to 2 distances $\times$ 4 non-*random* proximities $\times$ 3 gender selections $\times$ 2 assignments, and 6 combinations correspond to *random* proximity with 3 gender selections $\times$ 2 assignments. To assess the impact of these choices, our experiments are organized according to the three actors in our threat model. First, the speaker finds the **two** most promising combinations of design choices on the development set in terms of privacy in the *Ignorant* and *Lazy-Informed* scenarios and utility for ASR decoding. This is motivated by the high computational cost of anonymizing the *train-clean-360* subset of LibriSpeech and retraining ASV and ASR systems on it, which prevents the evaluation of privacy in the *Semi-Informed* scenario and utility for ASR training for all 54 combinations. Second, the user assesses the utility of these two combinations for both ASR training and decoding. Third, the attacker quantifies the resulting privacy in the *Semi-Informed* scenario, which leads us to identify the best combination among these two. Finally, we show how the proposed pitch transformation further improves privacy.

A. Speaker's perspective

We first evaluate the design choices from the speaker's perspective in terms of privacy in the *Ignorant* and *Lazy-Informed* scenarios and utility for ASR decoding on the development set. The results are displayed in the form of swarm plots, i.e., scatter plots where each dot represents the privacy or utility value associated with one combination of design choices. In order to avoid overlapping dots with similar values, the dots are spread horizontally.

1) Distance: Figure 4 evaluates the effect of the chosen distance metric on privacy. We observe that both cosine distance and PLDA result in similarly low linkability in the *Ignorant* case but PLDA marginally outperforms cosine distance (i.e., it results in a lower linkability) in the *Lazy-Informed* case. Since both distance measures perform similarly in terms of utility (see Fig. 9(a)), PLDA has an advantage. Therefore we consider only PLDA as the distance metric in the following experiments.

2) Proximity: Next, we assess the five choices of target proximity described in Section III-B, namely *random*, *near*, *far*, *sparse* and *dense*. The distance metric is fixed to PLDA and the values of N and N^* are fixed to 200 and 100, respectively.[10]

We observe in Fig. 5 that, although selecting candidate x-vectors *far* from the original speaker achieves the lowest linkability in the *Ignorant* case together with the *random* strategy, it is largely outperformed in the *Lazy-Informed* case by selection from *sparse* or *dense* clusters and by the *random* strategy. This shows that clustering based pseudo-speaker mapping results in more robust anonymization as compared to simple distance-based mapping.

[10]We noticed a sharp decline in utility for smaller values of N^*.

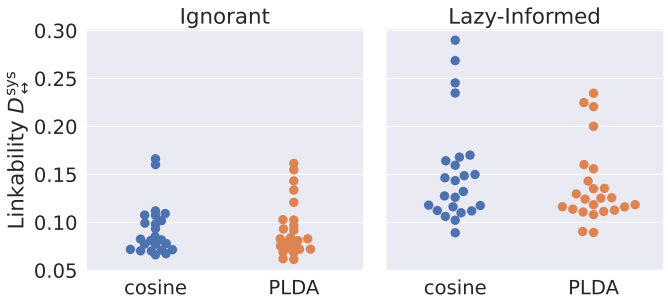


Fig. 4. Privacy against *Ignorant* and *Lazy-Informed* attackers depending on the distance choice. Each swarm plot shows the 24 linkability values on the development set resulting from all combinations of 4 proximity (excluding *random*), 3 gender selection, and 2 assignment choices.

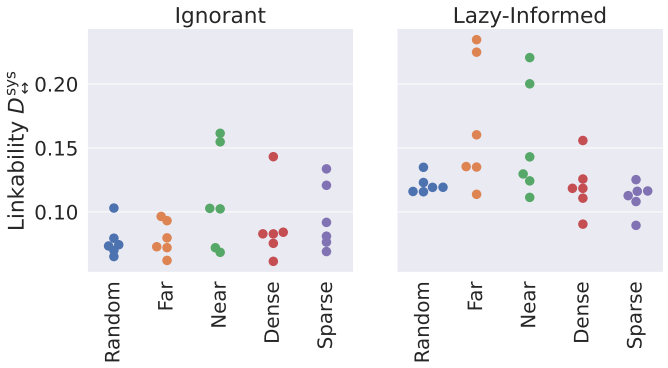


Fig. 5. Privacy against *Ignorant* and *Lazy-Informed* attackers depending on the proximity choice. Distance is fixed to PLDA. Each swarm plot shows the 6 linkability values on the development set resulting from all combinations of 3 gender selection and 2 assignment choices.

Compared to the *sparse* selection strategy, the *dense* strategy provides comparable privacy protection in the *Lazy-Informed* case, but much higher utility (see Fig. 9(b)). This can be attributed to the fact that speakers in *sparse* clusters stand out more from the crowd than those in *dense* clusters, therefore they are more likely to suffer from poor ASR performance.

Finally, *random* target selection yields similar privacy protection in the *Lazy-Informed* case and slightly better utility as compared to *dense*. Hence we consider the *random* and *dense* strategies to be the best choices for proximity.

3) Gender selection: We now investigate the gender selection strategy described in Section III-C. The distance is fixed to PLDA and proximity to *dense* or *random*. As per the results shown in Fig. 6 it is hard to find the best choice for gender selection in terms of privacy since the linkability is not consistently lower for any specific choice.

In order to make a suitable choice, we introduce the additional requirement that the chosen anonymization scheme obfuscates the original speaker's gender. The different anonymization schemes can be visually compared in Fig. 7. *Same* gender selection (Fig. 7 (b)) causes male and female clusters to move apart. A similar result is observed with *opposite* gender selection (not shown in the figure). On the

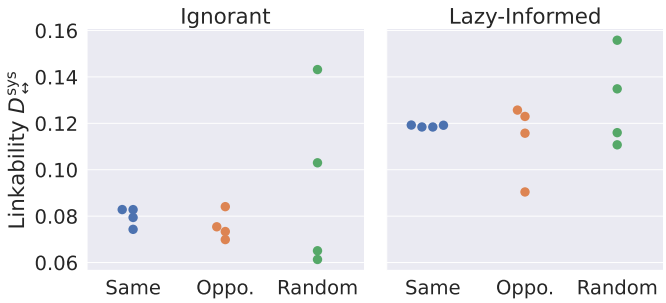


Fig. 6. Privacy against *Ignorant* and *Lazy-Informed* attackers depending on the gender selection choice. Distance is fixed to PLDA and proximity to *dense* or *random*. Each swarm plot shows the 4 linkability values on the development set resulting from all combinations of the 2 proximity and 2 assignment choices.

contrary, *random* gender-selection (Fig. 7(c) and 7(d)) results in a non-separable boundary between genders.

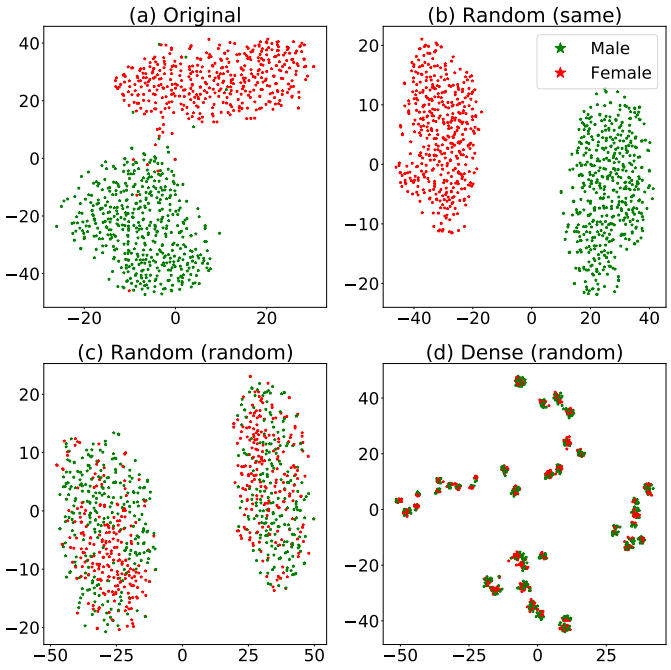


Fig. 7. t-SNE visualization of speaker-level x-vectors from the LibriSpeech *train-clean-360* dataset anonymized using different proximity (*random* or *dense*) and gender selection (*same* or *random*, in parentheses) choices. Gaussian pitch normalization (see Section V-D) was used in all three cases.

Furthermore, we conduct gender identification experiments over the original and anonymized x-vectors shown in Fig. 7 to measure the degree of gender obfuscation caused by *same* vs. *random* gender selection. We employ the k-nearest neighbour algorithm with 5-fold cross-validation to predict the gender of speakers in the LibriSpeech *train-clean-360* dataset which contains 921 speakers. The mean cross-validation accuracy for each dataset reported in Table II corroborates the visual observations in Fig. 7.

4) Assignment: Finally the choice of pseudo-speaker assignment is examined from the speaker's perspective as described in Section III-D. The distance is fixed to PLDA, proximity to *dense* and gender selection to *random*. The results

TABLE II
GENDER IDENTIFICATION ACCURACY OVER THE ORIGINAL AND ANONYMIZED X-VECTORS IN FIG. 7.

Anonymization scheme	**Mean cross-validation accuracy (%)**
Original	98.58
Random (*same*)	100.00
Random (*random*)	70.46
Dense (*random*)	53.31

reported in Fig. 8 show that *utterance-level* assignment results in lower linkability than *speaker-level* assignment. However, the WER resulting from *utterance-level* assignment is higher than from *speaker-level* assignment (see Fig. 9(d)).

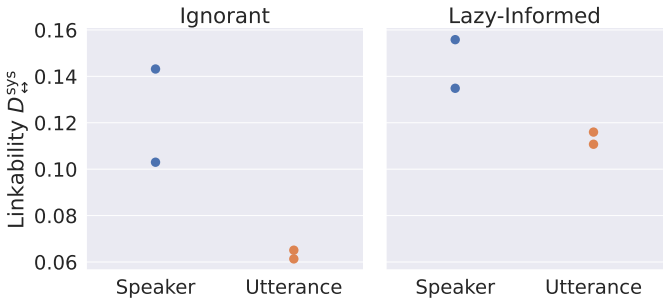


Fig. 8. Privacy against *Ignorant* and *Lazy-Informed* attackers depending on the assignment choice. Distance is fixed to PLDA, proximity to *dense* or *random*, and gender selection to *random*. Each swarm plot shows the 2 linkability values on the development set resulting from the 2 proximity choices.

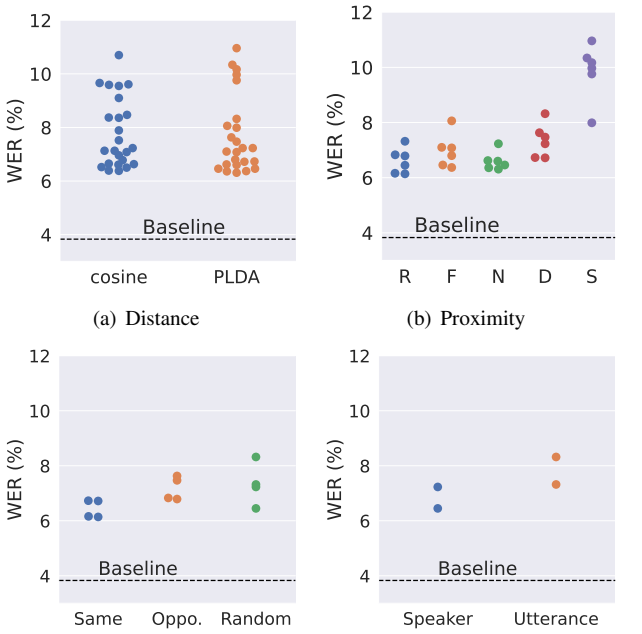
In the following, in order to conform with the requirements of the VoicePrivacy Challenge [22, Section 3.2], we choose *speaker-level* assignment. This ensures that "all trial utterances from a given speaker appear to be uttered by the same pseudo-speaker".

Based on these indications, the speaker may choose specific parameters according to their application needs. For the sake of further experimentation, we choose distance as **PLDA**, proximity as **random** or **dense**, gender selection as **random** and assignment as **speaker-level** to be the two best combinations of design choices based on our observations.

B. User's perspective

We now present complementary results from the user's perspective. Recall that in our threat model the user exploits the anonymized speech data for some downstream task. His/her primary concern is hence the utility of the data for that task. So far, we have only evaluated the utility for ASR decoding using an ASR system trained on the original *train-clean-360* dataset (see Fig. 9). We now evaluate the utility for ASR decoding using an ASR system trained on anonymized data, as well as the utility for ASR training. To do so, we anonymize the *train-clean-360* dataset using either of the two best combinations of design choices, and we retrain the ASR system on it.

Figure 10 shows the resulting utility values. The four bars in each plot represent the four decoding scenarios: OO indicates original (non-anonymized) speech decoded by the ASR model trained on original speech, AO indicates anonymized speech



(a) Distance (b) Proximity

(c) Gender-Selection (d) Assignment

Fig. 9. Utility of anonymized speech for ASR decoding compared to original (Baseline) speech depending on the design choices made by the speaker. Each swarm plot shows the WER values obtained on the development set using an ASR system trained on the original *train-clean-360* dataset. The design choices in subfigures a, b, c, and d are fixed or vary in the same way as in Figs. 4, 5, 6 and 8, respectively.

decoded by the ASR model trained on original speech, OA indicates original speech decoded by the ASR model retrained on anonymized speech, and AA indicates anonymized speech decoded by the ASR model retrained on anonymized speech.

Fig. 10. Utility of original or anonymized speech for ASR training and ASR decoding depending on the proximity choice (*random* or *dense*) made by the speaker. Distance is fixed to PLDA, gender selection to *random*, and assignment to *speaker-level*. See Section V-B for the definition of OO, AO, OA, and AA.

We observe a WER degradation in the AO and OA scenarios, which indicates a mismatch between training and test data. The degradation is higher when original speech is decoded using the retrained model (OA) than when anonymized speech is decoded using the original model (AO). This asymmetry suggests a "loss of generalization" of the ASR model trained on anonymized speech, due to the unintentional exclusion of certain factors of variability of original speech.

Fortunately, the WERs on the test set in the AA scenario are almost as low as those in the OO scenario. This indicates that anonymization yields viable speech data for ASR training and ASR decoding with a WER similar to original speech, provided that training and decoding are both conducted on anonymized speech. No significant difference is observed depending on the proximity choice made by the speaker.

C. Attacker's perspective

Finally, we present complementary results from the attacker's perspective. The primary objective of the attacker is to find the original speaker's identity of anonymized speech utterances, i.e., to achieve high linkability. So far, we have only reported the linkability achieved by *Ignorant* and *Lazy-Informed* attackers on the development set. We now present the results achieved by these attackers and by a *Semi-Informed* attacker on both the development and test sets.

The results for the two best combinations of design choices are shown in Fig. 11. We observe that the linkability increases as the strength of the attacker increases. It goes up to 0.44 with *random* proximity, but stays below 0.22 with *dense* proximity, even for the strongest (*Semi-Informed*) attacker. This indicates the robustness of *dense* over *random* proximity. Therefore, we ultimately recommend the following combination of choices to the speaker: **PLDA** distance, **dense** proximity, **random** gender selection, and **speaker-level** assignment. We recall that the latter choice is a requirement set by the VoicePrivacy challenge. Whenever *speaker-level* assignment is not required, we recommend *utterance-level* assignment for higher privacy.

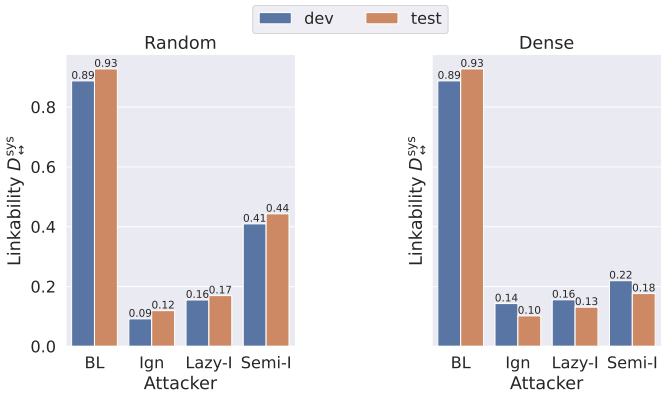


Fig. 11. Privacy against *Ignorant* (Ign), *Lazy-Informed* (Lazy-I) and *Semi-Informed* (Semi-I) attackers depending on the proximity choice (*random* or *dense*) made by the speaker, compared to original speech (BL). Distance is fixed to PLDA, gender selection to *random*, and assignment to *speaker-level*.

D. Pitch conversion

Sticking with the best combination of design choices, we bring one last improvement: we explore three pitch conversion methods to further enhance privacy and increase the naturalness of anonymized speech. Indeed, the pitch sequence $\mathbf{p}$ might reveal some information about the speaker [44] which is carried over to the synthesized speech. Also, keeping the pitch

sequence $\mathbf{p}$ unchanged while possibly changing the gender of the x-vector results in inconsistent features which affect the naturalness of the synthesized speech.

The three conversion methods operate on nonzero pitch values only. Indeed, zero pitch values correspond to unvoiced or silent frames, and must remain equal to zero to preserve the phonetic content of the utterance. Conversely, nonzero pitch values must remain nonzero. In the rest of this section, the term "pitch sequence" and the notation $\mathbf{p}$ refer to a sequence stripped off its zero values.

The first method called logarithm Gaussian pitch conversion [45] was recently employed in [46] for voice anonymization. The target pitch sequence $\mathbf{p}^*$ is obtained by linearly scaling the original pitch sequence $\mathbf{p}$ in the logarithmic domain as

$$\log(\mathbf{p}^*) = \frac{\log(\mathbf{p}) - \mu_{\text{src}}}{\sigma_{\text{src}}}\sigma_{\text{tgt}} + \mu_{\text{tgt}}, \tag{4}$$

where μ_{src}, σ_{src} are the mean and standard deviation of $\mathbf{p}$, and μ_{tgt}, σ_{tgt} are the mean and standard deviation of the target pseudo-speaker's pitch "sequence" $\mathbf{p}_{\text{ps}}$. The latter is obtained by concatenating the pitch sequences of all utterances of the N^* candidate speakers composing the pseudo-speaker; it is stored by the x-vector anonymization module (Step **2** in Fig. 2) and passed to the pitch conversion module (Step **3**).

In addition, we propose two other methods, which we call *percentile* and *minmax* based pitch conversion. Percentile based pitch conversion maps each percentile of the original pitch distribution to the corresponding percentile of the target pitch distribution. To do so, the sequences $\mathbf{p}$ and $\mathbf{p}_{\text{ps}}$ are sorted in ascending order, yielding $\mathbf{p}^{\text{sorted}}$ and $\mathbf{p}_{\text{ps}}^{\text{sorted}}$. Each value $\mathbf{p}[i]$ in the original sequence is converted into a percentile $\varrho[i]$:

$$\varrho[i] = \frac{\text{rank of } \mathbf{p}[i] \text{ in } \mathbf{p}^{\text{sorted}}}{\text{length}(\mathbf{p}^{\text{sorted}})} \times 100. \tag{5}$$

Then, the converted pitch value $\mathbf{p}^*[i]$ corresponding to $\varrho[i]$ is picked in $\mathbf{p}_{\text{ps}}^{\text{sorted}}$ as

$$\mathbf{p}^*[i] = \mathbf{p}_{\text{ps}}^{\text{sorted}}\left[\left\lfloor \frac{\text{length}(\mathbf{p}_{\text{ps}}^{\text{sorted}}) \times \varrho[i]}{100} \right\rfloor\right] \tag{6}$$

where $\lfloor\cdot\rfloor$ denotes rounding down to the nearest integer. This mapping is an instance of one-dimensional optimal transport between the two distributions [47]. To the best of our knowledge, this pitch conversion method is new.

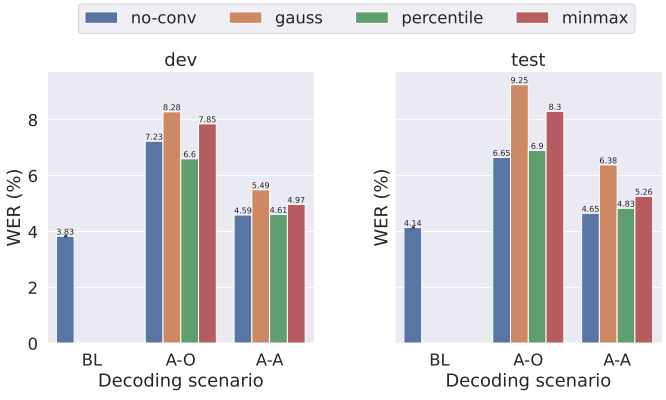
Minmax based pitch conversion linearly scales the range of pitch values, such that the minimum and maximum values in the original sequences are mapped to the minimum and maximum values of the target pseudo-speaker:

$$\mathbf{p}^*[i] = \left[(\mathbf{p}[i] - \min(\mathbf{p})) \times \frac{\max(\mathbf{p}_{\text{ps}}) - \min(\mathbf{p}_{\text{ps}})}{\max(\mathbf{p}) - \min(\mathbf{p})}\right] + \min(\mathbf{p}_{\text{ps}}). \tag{7}$$

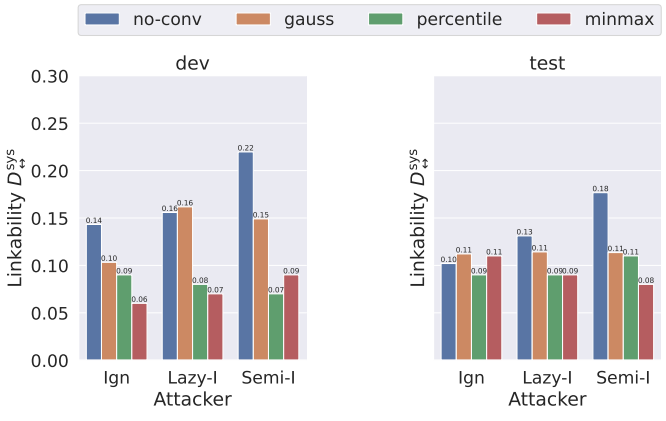
One benefit of percentile or minmax based conversion is that the converted pitch values belong to the range of pitch values for the N^* candidate speakers composing the pseudo-speaker, while in case of Gaussian normalization some converted pitch values may be beyond that range or even beyond the range of valid pitch values for male or female speakers.

It is observed in Fig. 12(a) that logarithm Gaussian pitch conversion and to a lesser extent minmax based conversion significantly increase the WER, while percentile based conversion maintains a WER close to the original. Figure 12(b) shows that percentile and minmax based conversion substantially improve privacy, especially against the *Semi-Informed* attacker, while logarithm Gaussian conversion results in a more modest improvement or no improvement. We conclude that percentile based conversion is a suitable pitch conversion method which increases privacy with no significant loss of utility. According to informal listening results (not show in the figure), it also improves the naturalness of cross-gender voice conversion.

Overall, the experiments in Section V exhibited the benefits of the proposed improvements to x-vector based voice conversion in terms of privacy against the strongest (*Semi-Informed*) attacker. Specifically, x-vector based voice conversion with the best combination of design choices and with percentile based pitch conversion reduces the attacker's linkability by one order of magnitude with respect to original speech. This is to be contrasted with signal processing based methods such as [48] which offer almost no protection against such a strong attacker.



(a) Utility



(b) Privacy

Fig. 12. Utility and privacy resulting from logarithm Gaussian (gauss), percentile or minmax based pitch conversion on top of x-vector anonymization, compared to original pitch (no-conv). Top: utility for ASR decoding with an ASR model trained on original (A-O) or anonymized (A-A) speech, compared to original speech (BL). Bottom: privacy against *Ignorant* (Ign), *Lazy-Informed* (Lazy-I) and *Semi-Informed* (Semi-I) attackers.

VI. Large-scale speaker study

Similarly to other studies following the VoicePrivacy Challenge setup, all experiments above have relied on a small set of 29 enrolled speakers. In this section we analyze the attacker's performance against a larger number of enrolled speakers. This number reflects the attacker's knowledge: a smaller number means that the attacker was able to narrow down the list of speakers who may have uttered the trial utterances using contextual information.[11] Our goal is to study whether the speaker's identity gets hidden in the crowd or is still revealed to some extent within a large set of enrolled speakers.

Previous research has studied the impact of the number of speakers from a voice spoofing perspective where an attacker aims to be accepted through an ASV authentication system by finding the closest (trial) impostor to a given (enrolled) target speaker [49], [50]. The attacker has access to a speech sample of the target speaker and to the ASV scoring mechanism. The authors showed that the chances of acceptance reach up to 50% as the number of impostors approaches 10^5. A similar problem was posed by the Multi-Target Speaker Detection Challenge [51] which aims to identify and assess the membership of a speaker to a set of blacklisted speakers. The authors showed that the performance for both tasks degrades as the number of speakers in the blacklist increases. In the following, we do not assess the worst-case performance like [50]. Instead, we measure the overall speaker recognition performance as the number of different-speaker trials increases manifold.

A. Experimental setup

We use the Mozilla Common Voice English corpus because of its large number of speakers. To the best of our knowledge this is the first time this corpus is used for ASV and privacy related experiments.

As mentioned in Section IV-A, we consider a total population of 24,610 male speakers. Out of these, utterances from 20 speakers are selected as the public trial data subjected to re-identification attack. After computing PLDA scores between the trial and enrollment utterances, we get 4,696 same-speaker scores and 115,563,864 different-speaker scores (see Table I).

We measure the attacker's performance against an increasing number of enrolled speakers. In the first step, we select for enrollement the same 20 speakers as for trial, and we use the corresponding same- and different-speaker scores. In the second step, we add 20 other speakers for enrollement and we include the corresponding different-speaker scores. In the following steps, we double the number of other speakers at each step, i.e., the total number of enrolled speakers increases from 20 to 20,500. The added speakers are randomly sampled 5 times from the entire speaker population to avoid any bias.

B. Open-set results

Figure 13 reports the performance of different attackers in terms of linkability. The linkability of original speech is equal to 0.80 with 20 speakers and decreases to 0.70 with more than a few hundred speakers. The linkability of anonymized speech is much lower. For *Ignorant* and *Lazy-Informed* attackers, it starts from 0.18 and decreases to 0.06. The linkability curve for the *Semi-Informed* attacker is surprisingly below those of the two other attackers, but it also follows a decreasing trend.

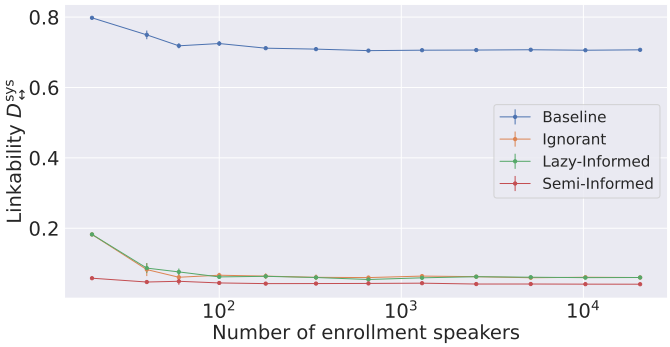


Fig. 13. Open-set ASV performance of *Ignorant*, *Lazy-Informed* and *Semi-Informed* attackers as a function of the number of enrolled speakers, compared to original speech (Baseline).

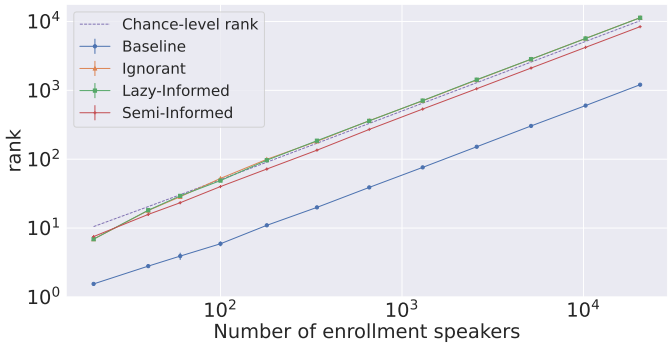
C. Closed-set results

The above evaluation in terms of linkability, which assumes that the attackers rely on open-set ASV, hides the fact that the chance of finding the true speaker of a trial utterance decreases very quickly as the number of enrolled speakers increases. To highlight it, we perform closed-set ASI as explained in Section IV-C and report the rank of the true speaker. The higher the rank, the lower the ASI performance. Since increasing the number of speakers is expected to increase the rank, we also report the normalized rank, that is the rank divided by the number of speakers, and the *chance-level* rank, that is the expected rank when the attacker selects the true speaker at random (see Appendix A).

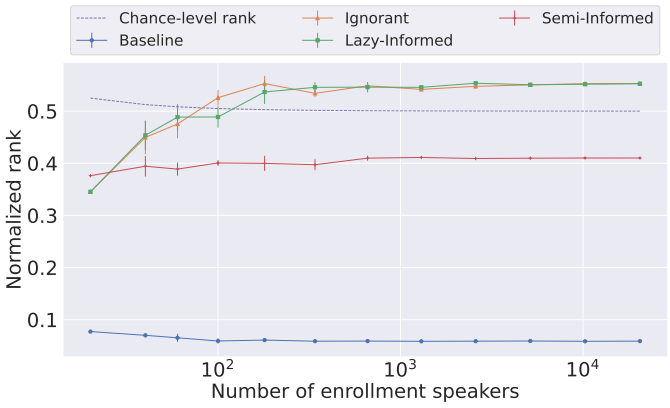
Figure 14(a) shows that the rank of the true speaker increases almost linearly as a function of the number of speakers. On original speech, it remains much lower than chance-level even with thousands of speakers, which can be attributed to the distinct characteristics of speakers in the population. On anonymized speech, it converges close to the chance-level rank for all attackers. The normalized rank plot in Fig. 14(b) shows that, beyond a few hundred speakers, the *Ignorant* and *Lazy-Informed* attackers perform more poorly than chance-level, while the *Semi-Informed* attacker maintains a consistent performance that is slightly better than chance-level.

In addition, we study the top-k precision obtained by the attackers compared to the baseline performance for $k \in \{1, 20\}$. We observe in Fig. 15 that the precision drops much faster on anonymized data than original data, i.e., finding the true speaker of an *anonymized* utterance among a set of S speakers is equivalent to finding the true speaker of an *original* utterance among S' speakers, where S' increases at a much faster rate than S. The plot for $k = 1$ shows that without anonymization the true speaker can be uniquely identified with 40% accuracy among 20,500 speakers, while after anonymization the risk

[11]The attacker may obtain this contextual information by inspecting the metadata and/or the statistics of the public, anonymized dataset, or by simply listening to individual utterances.

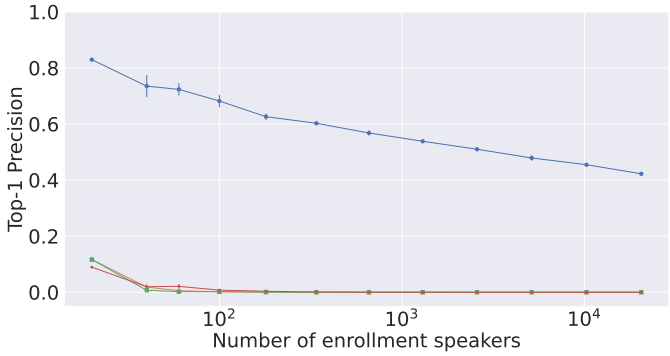


(a) Rank

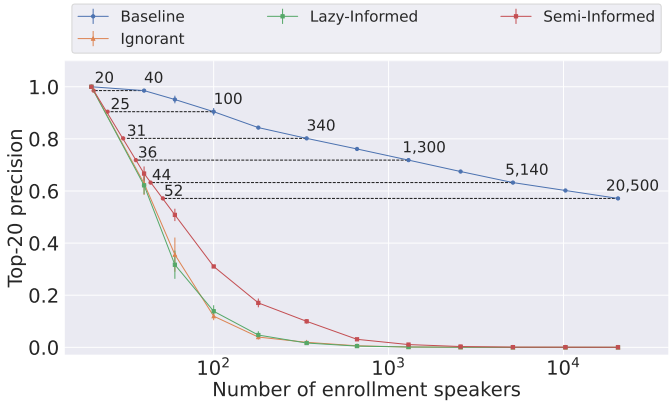


(b) Normalized rank

Fig. 14. Closed-set ASI performance of *Ignorant*, *Lazy-Informed* and *Semi-Informed* attackers as a function of the number of speakers, compared to original speech (Baseline).



(a) $k = 1$



(b) $k = 20$

Fig. 15. Top-k ASI precision of *Ignorant*, *Lazy-Informed* and *Semi-Informed* attackers as a function of the number of speakers, compared to original speech (Baseline). The numbers of speakers needed to achieve an equivalent drop in precision before vs. after anonymization are highlighted.

of being uniquely identified becomes negligible beyond a few dozen speakers. For $k = 20$, our anonymization scheme provides the same level of protection against a *Semi-Informed* attacker among 52 speakers than raw speech among 20,500 speakers. Additional results for $k \in \{10, 50\}$ (not shown here) follow a similar trend.

VII. Conclusion

Our work aimed to answer three questions from the speaker's and user's perspectives, and two questions from the attacker's perspective to realistically assess the privacy and utility of x-vector based speaker anonymization (see Section I).

To answer Q1, we introduced four design choices and studied their effect on the privacy of anonymized speech and its utility for ASR training and decoding. Based on our findings, we recommended the following optimal combination of choices: PLDA distance, *dense* proximity, *random* gender selection, and *utterance-level* assignment (unless otherwise required). To answer Q3, we then investigated three pitch conversion methods for removing the residual speaker information carried by the pitch sequence and to enhance the naturalness of the synthesized speech. While classical logarithm Gaussian conversion resulted in little or no improvement of privacy, the proposed percentile based conversion method significantly improved privacy with little loss of utility. Overall, x-vector based voice conversion with the best combination of design choices and with percentile based pitch conversion reduced the linkability against the strongest (*Semi-Informed*) attacker by one order of magnitude with respect to original speech (answering Q4), and it increased the WER on LibriSpeech *test-clean* from 4.1% to 4.8% only in the situation when ASR training and decoding are both conducted on anonymized speech (answering Q2).

To answer Q5, we further evaluated the proposed anonymization scheme as a function of the number of enrolled speakers, which reflects the attacker's ability to narrow down the list of speakers who may have uttered the trial utterances using contextual information. We conducted closed-set ASI by incrementally adding thousands of speakers in the population and observed that the rank of the true speaker quickly increases and converges close to chance-level after anonymization. Another interesting observation can be made by looking at top-k precision curves: the loss of precision before anonymization that is seen after adding thousands of speakers in the enrollment set is equivalent to adding only a couple of speakers after anonymization. Specifically, the best combination of design choices offers the same level of protection against re-identification attacks among 52 speakers as original speech among 20,500 speakers.

In the future, we plan to study the worst-case performance of the proposed speaker anonymization scheme and characterize which speakers are easier to re-identify. We also plan to provide analytical lower bounds on privacy by using techniques

inspired from differential privacy [52].

APPENDIX A
DERIVATION OF THE CHANCE-LEVEL RANK

Let $R \in \{1, \ldots, N_{\text{spk}}\}$ be the set of all possible ranks for a given speaker that can be obtained with probability $P(R)$. The expected rank is equal to:

$$\mathbb{E}(R) = \sum_{R=1}^{N_{\text{spk}}} R \cdot P(R). \tag{8}$$

To obtain the chance-level rank, we set $P(R) = \frac{1}{N_{\text{spk}}}$. Hence

$$\mathbb{E}(R) = \frac{1}{N_{\text{spk}}} \sum_{R=1}^{N_{\text{spk}}} R = \frac{1}{N_{\text{spk}}} \frac{N_{\text{spk}}(N_{\text{spk}}+1)}{2} = \frac{N_{\text{spk}}+1}{2}. \tag{9}$$

When the rank is normalized, we divide the chance-level rank by N_{spk} to obtain the normalized chance-level rank

$$\frac{N_{\text{spk}}+1}{2N_{\text{spk}}} \approx 0.5. \tag{10}$$

ACKNOWLEDGMENT

This work was supported in part by the French National Research Agency under projects HARPOCRATES (ANR-19-DATA-0008) and DEEP-PRIVACY (ANR-18-CE23-0018), by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 825081 COMPRISE (https://www.compriseh2020.eu/), and jointly by the French National Research Agency and the Japan Science and Technology Agency under project VoicePersonae. Experiments presented in this paper were partially 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).

REFERENCES

[1] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[2] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, "The accented English speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6918–6922.

[3] R. Gupta, T. Chaspari, J. Kim, N. Kumar, D. Bone, and S. Narayanan, "Pathological speech processing: State-of-the-art, current challenges, and future directions," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6470–6474.

[4] V. Vestman, T. Kinnunen, R. G. Hautamäki, and M. Sahidullah, "Voice mimicry attacks assisted by automatic speaker verification," *Computer Speech and Language*, vol. 59, pp. 36–54, 2020.

[5] 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 *Interspeech*, 2019, pp. 3695–3699.

[6] 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*, 2020, pp. 1693–1697.

[7] 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.

[8] 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.

[9] 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 & Understanding (ASRU)*, 2009, pp. 529–533.

[10] 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.

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

[12] T. Justin, V. Štruc, S. Dobrišek, B. Vesnicer, I. Ipšić, and F. Mihelič, "Speaker de-identification using diphone recognition and speech synthesis," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 4, 2015, pp. 1–7.

[13] 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 ISCA Speech Synthesis Workshop*, 2019, pp. 155–160.

[14] 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.

[15] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa, "Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[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] 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*, 2020, pp. 1713–1717.

[18] C. Magariños, P. Lopez-Otero, L. Docio-Fernandez, E. Rodriguez-Banga, D. Erro, and C. Garcia-Mateo, "Reversible speaker de-identification using pre-trained transformation functions," *Computer Speech and Language*, vol. 46, pp. 36–52, 2017.

[19] 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.

[20] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5916–5920.

[21] D. Yu and M. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Interspeech*, 2011, pp. 237–240.

[22] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J.-F. Bonastre, P.-G. Noé, M. Todisco, and J. Patino, "The VoicePrivacy 2020 Challenge evaluation plan," 2020. [Online]. Available: https://www.voiceprivacychallenge.org/vp2020/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf

[23] C. O. Mawalim, K. Galajit, J. Karnjana, and M. Unoki, "X-Vector Singular Value Modification and Statistical-Based Decomposition with Ensemble Regression Modeling for Speaker Anonymization System," in *Interspeech*, 2020, pp. 1703–1707.

[24] J. M. Perero-Codosero, F. M. Espinoza-Cuadros, and L. A. Hernández-Gómez, "X-vector anonymization using autoencoders and adversarial training for preserving speech privacy," *Computer Speech & Language*, vol. 74, p. 101351, 2022.

[25] H. Turner, G. Lovisotto, and I. Martinovic, "Speaker anonymization with distribution-preserving x-vector generation for the VoicePrivacy challenge 2020," *arXiv preprint arXiv:2010.13457*, 2020.

[26] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022.

[27] S. Ioffe, "Probabilistic linear discriminant analysis," in *9th European Conference on Computer Vision (ECCV)*, 2006, pp. 531–542.

[28] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.

[29] I. Salmun, I. Opher, and I. Lapidot, "On the use of PLDA i-vector scoring for clustering short segments," in *Odyssey*, 2016, pp. 407–414.
[30] J. Rohdin, S. Biswas, and K. Shinoda, "Constrained discriminative PLDA training for speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1670–1674.
[31] D. Dueck, "Affinity propagation: Clustering data by passing messages," Ph.D. dissertation, University of Toronto, 2009.
[32] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 726–733.
[33] 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.
[34] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Interspeech*, 2017, pp. 2616–2620.
[35] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018, pp. 1086–1090.
[36] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," *Interspeech*, pp. 1526–1530, 2019.
[37] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
[38] B. M. L. Srivastava, "Speaker anonymization — representation, evaluation and formal guarantees," Ph.D. dissertation, University of Lille, 2021.
[39] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Interspeech*, 2008, pp. 2598–2601.
[40] 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.
[41] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi, and E. Vincent, "A comparative study of speech anonymization metrics," in *Interspeech*, 2020, pp. 1708–1712.
[42] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.
[43] H. Abdi and L. Williams, "Jackknife," in *Encyclopedia of Research Design*, 2010, pp. 1–10.
[44] M. Farrús, M. Wagner, J. Anguita, and J. Hernando, "How vulnerable are prosodic features to professional imitators?" in *Odyssey*, 2008.
[45] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with STRAIGHT for Mandarin," in *4th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, vol. 4, 2007, pp. 410–414.
[46] P. Champion, D. Jouvet, and A. Larcher, "A study of F0 modification for x-vector based speech pseudonymization across gender," in *2nd AAAI Workshop on Privacy-Preserving Artificial Intelligence*, 2021.
[47] C. Villani, *Optimal Transport: Old and New*. Springer, 2009.
[48] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," in *Interspeech*, 2021, pp. 1099–1103.
[49] A. Sholokhov, T. Kinnunen, V. Vestman, and K. A. Lee, "Voice biometrics security: Extrapolating false alarm rate via hierarchical bayesian modeling of speaker verification scores," *Computer Speech and Language*, vol. 60, p. 101024, 2020.
[50] ——, "Extrapolating false alarm rates in automatic speaker verification," in *Interspeech*, 2020, pp. 4218–4222.
[51] S. Shon, N. Dehak, D. Reynolds, and J. Glass, "MCE 2018: The 1st multi-target speaker detection and identification challenge evaluation," in *Interspeech*, 2019, pp. 356–360.
[52] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.