Design Choices for X-vector Based Speaker Anonymization

Brij Mohan Lal Srivastava1, Natalia Tomashenko2, Xin Wang3, Emmanuel Vincent4, Junichi Yamagishi5, Mohamed Maouche3, Aurélien Bellet1, Marc Tommasi2

1Inria, France
2Laboratoire Informatique d’Avignon (LIA), Avignon Université, France
3National Institute of Informatics, Tokyo, Japan
4Université de Lorraine, CNRS, Inria, LORIA, France
5Université de Lille, France

organisers@lists.voiceprivacychallenge.org

Abstract

The recently proposed x-vector based anonymization scheme converts any input voice into that of a random pseudo-speaker. In this paper, we present a flexible pseudo-speaker selection technique as a baseline for the first VoicePrivacy Challenge. We explore several design choices for the distance metric between speakers, the region of x-vector space where the pseudo-speaker is picked, and gender selection. To assess the strength of anonymization achieved, we consider attackers using an x-vector based speaker verification system who may use original or anonymized speech for enrollment, depending on their knowledge of the anonymization scheme. The Equal Error Rate (EER) achieved by the attackers and the decoding Word Error Rate (WER) over anonymized data are reported as the measures of privacy and utility. Experiments are performed using datasets derived from LibriSpeech to find the optimal combination of design choices in terms of privacy and utility.

Index Terms: speaker anonymization, VoicePrivacy challenge, voice conversion, PLDA, x-vectors

1. Introduction

Privacy protection methods for speech fall into four broad categories [1]: deletion, encryption, distributed learning, and anonymization. The VoicePrivacy initiative [1] specifically promotes the development of anonymization methods which aim to suppress personally identifiable information in speech while leaving other attributes such as linguistic content intact. Recent studies have proposed anonymization methods based on noise addition [2], speech transformation [3], voice conversion [4–6], speech synthesis [7,8], and adversarial learning [9]. We focus on voice conversion / speech synthesis based methods due to the naturalness of their output and their promising results so far.

In order to implement a speaker anonymization scheme based on voice conversion or speech synthesis, we must address the following questions: 1. What is the best representation to characterize speaker information in a speech signal? 2. Which distance metric is most appropriate to explore various regions of the speaker space? 3. How to optimally select target speakers from a small pool of speakers? 4. How to combine the distance metric and target selection in order to strike balance between privacy protection and loss of utility?

Classically, speaker anonymization methods that rely on a voice conversion or speech synthesis system select a random target speaker from a pool of speakers which must be included in the training set for that system. This constraint severely restricts the user’s freedom to choose an arbitrary unseen speaker as the target for anonymization. Moreover, several targets cannot be mixed together to create an imaginary sample in speaker space, i.e., a pseudo-speaker. In a previous experimental study [10], we specified three criteria to be satisfied by voice conversion algorithms for speaker anonymization: 1) non-parallel, 2) many-to-many, and 3) source- and language-independent. Although the algorithms compared in [10] satisfied these criteria, they did not allow conversion conditioned on a continuous speaker representation, such as x-vectors [11].

Recently, Fang et al. [8] proposed to identify x-vectors at a fixed distance from the “user” x-vector and to combine them to produce a pseudo-speaker representation. This representation, along with the “user” linguistic representation, is provided as input to a Neural Source-Filter (NSF) [12] based speech synthesizer to produce anonymized speech. Han et al. [13] extended [8] by proposing a metric privacy framework where an x-vector based pseudo-speaker is selected so as to satisfy a given privacy budget. Based on these studies, we answer Question 1 by choosing x-vectors as the appropriate speaker representation. In addition, the freedom to generate previously unseen pseudo-speakers by combining existing speakers from a small dataset exponentially increases the choices for the user.

The user may select pseudo-speakers at random in the entire x-vector space or based on specific properties, such as density of speakers, gender majority, etc. They must also choose a similarity metric between x-vectors since this dictates the properties of the vector space. Previous studies [14] have shown that Probabilistic Linear Discriminant Analysis (PLDA) yields state-of-the-art speaker verification performance, superior to the cosine distance. This is attributed to the formulation of PLDA which estimates the factorized within-speaker and between-speaker variability in speaker space. Hence, the PLDA score provides a good estimate of the log-likelihood ratio between same-speaker and different-speaker hypotheses, making it a superior measure of speaker affinity even for short speech segments [15].

In this paper, we establish that a greater level of anonymization is achieved when the distance between x-vectors is measured by PLDA instead of the cosine distance as used by Fang et al. [8] (answering Question 2). Then, we introduce a design choice called proximity which allows us to pick the pseudo-speaker in dense, sparse, far, or near regions of speaker space. We further explore the flexibility of this anonymization scheme by exploring the influence of gender selection. These design choices are evaluated using attackers which may or may not know the anonymization scheme applied (answering Question 3). Finally, we suggest the optimal combination of distance metric and design choices based on qualitative and quantitative measures to balance privacy and utility (answering Question 4).
We describe the general anonymization framework and the proposed design choices in Section 2. The datasets and evaluations metrics are briefly explained in Section 3. We present the experiments and discuss their results in Section 4. Section 5 concludes the paper.

2. Anonymization design choices

The general anonymization scheme follows the method proposed in [16] and shown in Fig. 1. It comprises three steps: Step 1 (Feature extraction) extracts fundamental frequency (F0) and bottleneck (BN) features and the source speaker’s x-vector from the input signal. Step 2 (X-vector anonymization) anonymizes this x-vector using an external pool of speakers. Step 3 (Speaker synthesis) synthesizes a speech waveform from the anonymized x-vector and the original BN and F0 features using an acoustic model (AM) and the NSF model.

![General anonymization scheme](image)

**Figure 1:** General anonymization scheme.

Step 2 (yellow box in Fig. 1) is the focus of this paper. It aims to generate a pseudo-speaker and comprises two sub-steps: 1) select \( N^* \) candidate target x-vectors from the anonymization pool; 2) average them to obtain the pseudo-speaker’s x-vector. In the following, we introduce various design choices for pseudospeaker selection. In all cases, a single target pseudo-speaker x-vector is selected for a given source speaker \( S \), and all the utterances of \( S \) are mapped to it, following the perm strategy described in [10]. This strategy has been shown to perform robust anonymization compared to other strategies described in [10].

2.1. Distance Metric: Cosine vs. PLDA

We compare two metrics to identify candidates for target x-vectors. The first one is the cosine distance, which was used by [8]. It is defined as

\[
1 - \frac{u \cdot v}{||u||_2 ||v||_2}
\]

for a pair of x-vectors \( u \) and \( v \). The second one is PLDA [17], which represents the log-likelihood ratio of same-speaker (\( H_s \)) and different-speaker (\( H_d \)) hypotheses. PLDA models x-vectors \( \omega \) as \( \omega = m + V y + D z \), where \( m \) is the center of the acoustic space, the columns of \( V \) represent speaker variability (eigenvoices) with \( y \) depending only on the speaker, and the columns of \( D \) capture channel variability (eigenchannels) with \( z \) varying from one recording to another. The parameters \( m, V \) and \( D \) are trained using x-vectors from the training set for the x-vector model, which is used to generate the anonymization pool. The log-likelihood ratio score

\[
\text{PLDA} = \log \frac{p(\omega, \omega | H_s)}{p(\omega, \omega | H_d)}
\]

can be computed in closed form [18]. We propose to use minus-PLDA as the “distance” between a pair of x-vectors.

2.2. Proximity: Random

The simplest candidate x-vector selection strategy called random consists of simply selecting \( N^* \) (set to 100) x-vectors uniformly at random from the same gender as the source in the anonymization pool. Note that this strategy does not allow us to choose particular regions of interest in x-vector space.

2.3. Proximity: Near vs. Far

The notion of distance can be used to define regions in x-vector space which closely resemble (near) or least resemble (far) the source speaker \( S \). In essence, we rank all the x-vectors in the anonymization pool in increasing order of their distance from \( S \) 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. The variability of results is controlled by a fixed random seed. The values of \( N \) and \( N^* \) are fixed to 200 and 100 respectively in our experiments. We noticed a sharp decline in utility for a smaller value of \( N^* \).

2.4. Proximity: Sparse vs. Dense

A simple mapping to far or near regions might produce biased pseudo-speaker estimates and the actual region where the output x-vector lies may not be optimal with respect to the distance from the source speaker. In order to pick the target pseudo-speaker in a specific region, we identify clusters of x-vectors in the anonymization pool which are then ranked based on their density. The density of each cluster is determined by the number of members belonging to that cluster.

We use Affinity Propagation [19] to determine the number of clusters and their members in the anonymization pool. Affinity Propagation is a non-parametric clustering method where the number of clusters is determined automatically through a message passing protocol. Two parameters determine the final number of clusters: preference assigns prior weights to samples which may be likely candidates for centroids, and damping factor is a floating-point multiplier to responsibility and availability messages. In our experiments, equal preference is assigned to each sample and the damping factor is set to 0.5. Out of 1160 speakers in the anonymization pool, 80 clusters were found, including 46 male and 34 female. The number of speakers per cluster ranges from 6 (sparse) to 36 (dense).

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 (\( N^* \)) are averaged to produce the pseudo-speaker. The 50% candidate x-vectors for a given cluster remain fixed for a given random seed.

2.5. Gender-selection: Same, Opposite, or Random

We observe clear clustering of the two genders in x-vector space using both cosine and PLDA distances. Hence, we pose gender selection as a design choice to study its impact on anonymization and intelligibility. We have the gender information for the source speaker as well as the speakers in the anonymization pool. Hence this design choice can be combined with all proximity choices. We study three different types of gender selection: same where the candidate target x-vectors are constrained to be of the same gender as the source; opposite where they are constrained to be of the opposite gender; and random where the target gender is selected at random before picking candidate x-vectors of that gender.
3. Experimental setup

3.1. Data
Following the rules of the VoicePrivacy Challenge, we use three publicly available datasets for our experiments. VoxCeleb-1.2 [20, 21] and the train-clean-100 and train-other-500 subsets of LibriSpeech [22] and LibriTTS [23] are used to train the models described in Section 2. The development and test sets are built from LibriSpeech dev-clean and test-clean, respectively. Details about the number of speakers, utterances, and trials in the enrollment and trial sets can be found in [1].

3.2. Evaluation methodology
We evaluate the above design choices in terms of privacy and utility. We define utility as the objective intelligibility of anonymized speech measured by the Word Error Rate (WER). The primary metric for privacy is the Equal Error Rate (EER).

3.2.1. Attack model
Privacy protection can be seen as a game between two entities: a “user” who publishes anonymized speech to hide his/her identity, and an “attacker” who attempts to uncover the user’s identity by conducting speaker verification trials over enrolled speakers. The attacker may possibly use some knowledge about the anonymization scheme to transform the enrollment data.

To assess the strength of anonymization against attackers with increasing amounts of knowledge, we perform the evaluation in three stages. The first scenario (Baseline) refers to the case when the user does not perform any anonymization before publication and the attacker also uses non-anonymized speech for enrollment. This attacker typically achieves low error rate (i.e., the user identity is accurately predicted) since there is no anonymization. In the second scenario (Ignorant), the user publishes anonymized speech, unbeknownst to the attacker who still uses non-anonymized speech for enrollment. Finally, in the Semi-Ignorant scenario, both the user and the attacker use anonymized speech for publication and enrollment respectively. However, the parameters of anonymization used by the attacker might differ from the user’s parameters.

The final scenario is the one in which the user is most vulnerable, hence it is considered as the lower bound for privacy in the context of this study. Note that there can be even stronger attacks [10] when the attacker has the exact knowledge of the anonymization parameters and uses it to generate large amounts of training data. This scenario is referred to in [10] as the Informed scenario. However it is not very realistic, so we do not consider it here.

3.2.2. Metrics
In all scenarios, the attacker implements the attack using a pretrained x-vector-PLDA based Automatic Speaker Verification (ASV$_{eval}$) system. Privacy protection is assessed in terms of the rate of failure of the attacker, as measured by the EER. The EER is computed from the distribution of PLDA scores generated by ASV$_{eval}$. In addition, a pretrained Automatic Speech Recognition (ASR$_{eval}$) system is used to decode anonymized speech and compute the WER for utility evaluation. Both evaluation systems are trained on disjoint data from that used to train the anonymization system. For more details, see [1].

Although we use Kaldi [24] to implement ASV$_{eval}$, we do not use it to compute the EER. Instead we use the PLDA scores output by ASV$_{eval}$ as inputs to the cllr toolkit2 to compute the ROCCH-EER [25]. The ROCCH-EER has interesting properties from the privacy perspective [26]. Its value does not exceed 50% which is considered as the upper-bound for anonymization since it implies complete overlap between genuine and impostor PLDA score distributions [27]. The higher the ROCCH-EER and the lower the WER, the better.

4. Experimental results
All the experiments are performed using the publicly available recipe of the VoicePrivacy Challenge.3 Figure 2 shows the EER values achieved by the considered anonymization scheme for different design choices. The corresponding WERs are reported in Table 1. To qualitatively analyze the effect of anonymization over the source speakers’ x-vectors, we also compute the average PLDA distance between original and anonymized x-vectors over all trial utterances in the test set. Figure 3 shows the average PLDA distance obtained for different design choices.

---

2https://gitlab.eurecom.fr/nautsch/cllr
3https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

---

![Figure 2: ROCCH-EER (%) obtained by ASV$_{eval}$ on the test set by an Ignorant or a Semi-Ignorant attacker for different design choices.](image)

(a) Distance
(b) Proximity
(c) Gender selection

---

Figure 2: ROCCH-EER (%) obtained by ASV$_{eval}$ on the test set by an Ignorant or a Semi-Ignorant attacker for different design choices. a) Distance: cosine vs. PLDA. Proximity is fixed to far and gender to same. b) Proximity: random, near, far, sparse, or dense. Distance is fixed to PLDA and gender to same. c) Gender: same, opposite, or random. Distance is fixed to PLDA and proximity to same.
Figure 3: Average PLDA distance between original and anonymized x-vectors for different design choices. Comparison of: a) Distance with proximity as far and gender as same; b) Proximity with distance as PLDA and gender as same; c) Gender with distance as PLDA and proximity as dense. (Darker left bars: male speakers, Lighter right bars: female speakers)

Table 1: WER (%) obtained by ASR_{test} on the dev and test sets.

<table>
<thead>
<tr>
<th>Distance</th>
<th>Proximity</th>
<th>Gender-selection</th>
<th>Dev WER (%)</th>
<th>Test WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (no anonymization)</td>
<td></td>
<td></td>
<td>3.83</td>
<td>4.15</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td>Same</td>
<td>6.28</td>
<td>6.58</td>
</tr>
<tr>
<td>Cosine</td>
<td>Far</td>
<td>Same</td>
<td>6.30</td>
<td>6.81</td>
</tr>
<tr>
<td></td>
<td>Near</td>
<td>Same</td>
<td>6.38</td>
<td>6.71</td>
</tr>
<tr>
<td></td>
<td>Sparse</td>
<td>Same</td>
<td>6.42</td>
<td>6.79</td>
</tr>
<tr>
<td></td>
<td>Dense</td>
<td>Same</td>
<td>10.14</td>
<td>10.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random</td>
<td>6.45</td>
<td>6.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Opposite</td>
<td>6.86</td>
<td>6.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Same</td>
<td>7.22</td>
<td>7.19</td>
</tr>
</tbody>
</table>

4.1. Distance

Our first experiment aims to identify the distance metric which is most suitable for the selection of candidate target x-vectors. To do so, we fix the proximity as far and the gender selection strategy as same, and we consider cosine distance vs. PLDA. We observe in Fig. 2(a) that cosine distance and PLDA result in a comparable high ROCCH-EER in the Ignorant case but PLDA consistently outperforms cosine distance (i.e., it results in a higher ROCCH-EER) in the Semi-Ignorant case. We also notice in Fig. 3 that the average PLDA distance between original and anonymized x-vectors is lower with cosine distance as compared to PLDA. For these reasons, we use PLDA to measure distances in x-vector space in the following experiments.

4.2. Proximity

Our second experiment assesses the five choices of target proximity described in Sections 2.2, 2.3 and 2.4. The distance metric is fixed to PLDA and the gender selection strategy to same. We observe in Fig. 2(b) that although x-vector selection from a far region achieves the greatest level of anonymization in the Ignorant case, it is outperformed by selection from sparse or dense regions in the Semi-Ignorant case. We notice in Fig. 3 that the target x-vectors are not too far from the source in the case of sparse or dense when compared to far. This may be due to the fact that same gender selection allows only same-gender clusters which lie nearby the source x-vectors. Random target selection provides similar privacy protection and average PLDA distance as sparse or dense.

Although random target selection produces comparable privacy protection and utility to dense, it limits the flexibility to select different regions in x-vector space. Compared to the sparse selection strategy, the dense strategy provides slightly better privacy protection in the Semi-Ignorant case, as well as higher utility (see Table 1). This might be due to fewer members in sparse clusters, hence a smaller value of N∗ as pointed out in Section 2.3. Consequently we select the dense strategy in our third experiment.

4.3. Gender selection

Our third experiment concerns the gender selection strategy in Section 2.5. The distance is fixed to PLDA and proximity to dense. When we look at male trials in Fig. 2(c), it is not clear which gender selection strategy is the best among same and opposite, but female trials show that random strategy outperforms the rest. We also observe in Fig. 3 that the mean distance is much higher in the case of random and opposite gender selection, which is intuitive since it allows selection of dense clusters from other genders as well. However, we notice that utility suffers in the case of opposite gender selection (see Table 1) due to limitations of cross-gender voice conversion. Hence we can conclude that random gender selection is the best choice.

5. Conclusions

We presented a flexible speaker anonymization scheme as the primary baseline for the first VoicePrivacy Challenge. In particular we proposed three design choices for target selection in x-vector space, namely distance metric, proximity, and gender selection which can be combined to obtain various anonymization systems. We objectively evaluated these choices in terms of ROCHC-EER to measure privacy protection and decoding WER to measure utility. We also reported the average PLDA distance between the source and the target. We showed that the previously used cosine distance is not the best choice of distance in x-vector space and it should be replaced by PLDA. Then we explored interesting regions in the x-vector space for picking the target pseudo-speaker during anonymization. We observed that when the target is picked in a dense region and the target gender is selected at random, robust privacy protection can be achieved against both Ignorant and Semi-Ignorant attackers with a reasonable loss of utility. In the future, we will evaluate the best design choices with additional utility metrics, e.g., the WER obtained after retraining ASR_{test} on anonymized data.

6. Acknowledgments

This work was supported in part by ANR and JST under projects DEEP-PRIVACY, HARPOCRATES, and VoicePersonae, 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 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).

1716
7. References


