Introducing the VoicePrivacy Initiative


1LIA, University of Avignon, France 2Inria, France 3NII, Tokyo, Japan 4Université de Lorraine, CNRS, Inria, LORIA, France 5EURECOM, France 6University of Edinburgh, UK
organisers@lists.voiceprivacychallenge.org

Abstract

The VoicePrivacy initiative aims to promote the development of privacy preservation tools for speech technology by gathering a new community to define the tasks of interest and the evaluation methodology, and benchmarking solutions through a series of challenges. In this paper, we formulate the voice anonymization task selected for the VoicePrivacy 2020 Challenge and describe the datasets used for system development and evaluation. We also present the attack models and the associated objective and subjective evaluation metrics. We introduce two anonymization baselines and report objective evaluation results.

Index Terms: privacy, anonymization, speech synthesis, voice conversion, speaker verification, automatic speech recognition

1. Introduction

Recent years have seen mounting calls for the preservation of privacy when treating or storing personal data. This is not least the result of the European general data protection regulation (GDPR). While there is no legal definition of privacy [1], speech data encapsulates a wealth of personal information that can be revealed by listening or by automated systems [2]. This includes, e.g., age, gender, ethnic origin, geographical background, health or emotional state, political orientations, and religious beliefs, among others [3, p. 62]. In addition, speaker recognition systems can reveal the speaker’s identity. It is thus of no surprise that efforts to develop privacy preservation solutions for speech technology are starting to emerge. The VoicePrivacy initiative aims to gather a new community to define the tasks of interest and the evaluation methodology, and to benchmark these solutions through a series of challenges.

Current methods fall into four categories: deletion, encryption, distributed learning, and anonymization. Deletion methods [4, 5] are meant for ambient sound analysis. They delete or obfuscate any overlapping speech to the point where no information about it can be recovered. Encryption methods [6, 7] such as fully homomorphic encryption [8] and secure multi-party computation [9], support computation upon data in the encrypted domain. They incur significant increases in computational complexity, which require special hardware. Decentralized or federated learning methods aim to learn models from distributed data without accessing it directly [10]. The derived data used for learning (e.g., model gradients) may still leak information about the original data, however [11].

Anonymization refers to the goal of suppressing personally identifiable attributes of the speech signal, leaving all other attributes intact. Past and recent attempts have focused on noise addition [12], speech transformation [13], voice conversion [14–16], speech synthesis [17, 18], or adversarial learning [19]. In contrast to the above categories of methods, anonymization appears to be more flexible since it can selectively suppress or retain certain attributes and it can easily be integrated within existing systems. Despite the appeal of anonymization and the urgency to address privacy concerns, a formal definition of anonymization and attacks against it is missing. Furthermore, the level of anonymization offered by existing solutions is unclear and not meaningful because there are no common datasets, protocols and metrics.

For these reasons, the VoicePrivacy 2020 Challenge focuses on the task of speech anonymization. This paper is intended as a general reference about the Challenge for researchers, engineers and privacy professionals. Details for participants are provided in the evaluation plan [20] and on the challenge website.

The paper is structured as follows. The anonymization task and the attack models, the datasets, and the metrics are described in Sections 2, 3, and 4, respectively. The two baseline systems and the corresponding objective evaluation results are presented in Section 5. We conclude in Section 6.

2. Anonymization task and attack models

Privacy preservation is formulated as a game between users who publish some data and attackers who access this data or data derived from it and wish to infer information about the users [21, 22]. To protect their privacy, the users publish data that contain as little personal information as possible while allowing one or more downstream goals to be achieved. To infer personal information, the attackers may use additional prior knowledge.

Focusing on speech data, a given privacy preservation scenario is specified by: (i) the nature of the data: waveform, features, etc., (ii) the information seen as personal: speaker identity, traits, spoken contents, etc., (iii) the downstream goal(s): human communication, automated processing, model training, etc., (iv) the data accessed by the attackers: one or more utterances, derived data or model, etc., (v) the attackers’ prior knowledge: previously published data, privacy preservation method applied, etc. Different specifications lead to different privacy preservation methods from the users’ point of view and different attacks from the attackers’ point of view.

2.1. Privacy preservation scenario

VoicePrivacy 2020 considers the following scenario, where the terms “user” and “speaker” are used interchangeably. Speakers want to hide their identity while still allowing all other downstream goals to be achieved. Attackers have access to one or more utterances and want to identify the speakers.

1In the legal community, the term “anonymization” means that this goal has been achieved. Here, it refers to the task to be addressed, even when the method being evaluated has failed. We expect the VoicePrivacy initiative to lead to the definition of new, unambiguous terms.

2https://www.voiceprivacychallenge.org/
2.2. Anonymization task

To hide his/her identity, each speaker passes his/her utterances through an anonymization system. The resulting anonymized utterances are referred to as trial data. They sound as if they had been uttered by another speaker called pseudo-speaker, which may be an artificial voice not corresponding to any real speaker.

The task of challenge participants is to design this anonymization system. In order to allow all downstream goals to be achieved, this system should:  (a) output a speech waveform, (b) hide speaker identity as much as possible, (c) distort other speech characteristics as little as possible, (d) ensure that all trial utterances from a given speaker appear to be uttered by the same pseudo-speaker, while trial utterances from different speakers appear to be uttered by different pseudo-speakers.

Requirement (c) is assessed via utility metrics: automatic speech recognition (ASR) decoding error rate using a model trained on original, i.e., unprocessed data and subjective speech intelligibility and naturalness (see Section 4). Requirement (d) and additional downstream goals including ASR training will be assessed in a post-evaluation phase (see Section 6).

2.3. Attack models

The attackers have access to: (a) one or more anonymized trial utterances, (b) possibly, original or anonymized enrollment utterances for each speaker. They do not have access to the anonymization system applied by the user. The protection of personal information is assessed via privacy metrics, including objective speaker verifiability and subjective speaker verifiability and linkability. These metrics assume different attack models.

The objective speaker verifiability metrics assume that the attackers have access to a single anonymized trial utterance and several enrollment utterances. Two sets of metrics are used for original vs. anonymized enrollment data (see Section 4.1). In the latter case, we assume that the trial and enrollment utterances of a given speaker have been anonymized using the same system, but the corresponding pseudo-speakers are different.

The subjective speaker verifiability metric (Section 4.2) assumes that the attackers have access to a single anonymized trial utterance and a single original enrollment utterance. Finally, the subjective speaker linkability metric (Section 4.2) assumes that the attackers have access to several anonymized trial utterances.

3. Datasets

Several publicly available corpora are used for the training, development and evaluation of speaker anonymization systems.

3.1. Training set

The training set comprises the 2,800 h VoxCeleb-1.2 speaker verification corpus [23,24] and 600 h subsets of the LibriSpeech [25] and LibriTTS [26] corpora, which were initially designed for ASR and speech synthesis, respectively. The selected subsets are detailed in Table 1 (top).

3.2. Development set

The development set comprises LibriSpeech dev-clean and a subset of the VCTK corpus [27] denoted as VCTK-dev (see Table 1, middle). With the above attack models in mind, we split them into trial and enrollment subsets. For LibriSpeech dev-clean, the speakers in the enrollment set are a subset of those in the trial set. For VCTK-dev, we use the same speakers for enrollment and trial and we consider two trial subsets, denoted as common and different. The common trial subset is composed of utterances #1 – 24 in the VCTK corpus that are identical for all speakers. This is meant for subjective evaluation of speaker verifiability/linkability in a text-dependent manner. The enrollment and different trial subsets are composed of distinct utterances for all speakers.

3.3. Evaluation set

Similarly, the evaluation set comprises LibriSpeech test-clean and a subset of VCTK called VCTK-test (see Table 1, bottom).

4. Utility and privacy metrics

Following the attack models in Section 2.3, we consider objective and subjective privacy metrics to assess anonymization performance in terms of speaker verifiability and linkability. We also propose objective and subjective utility metrics to assess whether the requirements in Section 2.2 are fulfilled.

4.1. Objective metrics

For objective evaluation, we train two systems to assess speaker verifiability and ASR decoding error. The first system denoted $ASV_{eval}$ is an automatic speaker verification (ASV) system, which produces log-likelihood ratio (LLR) scores. The second system denoted $ASR_{eval}$ is an ASR system which outputs a word error rate (WER). Both are trained on LibriSpeech train-clean-360 using Kaldi [28].

4.1.1. Objective speaker verifiability

The $ASV_{eval}$ system for speaker verifiability evaluation relies on x-vector speaker embeddings and probabilistic linear discriminant analysis (PLDA) [29]. Three metrics are computed: the equal error rate (EER) and the LLR-based costs $C_{ln}$ and $C_{min}$. Denoting by $P_a(\theta)$ and $P_{fa}(\theta)$ the false alarm and miss rates at threshold $\theta$, the EER corresponds to the threshold $\theta_{EER}$ at which the two detection error rates are equal, i.e., $EER = P_a(\theta_{EER}) = P_{fa}(\theta_{EER})$. $C_{ln}$ is computed from PLDA
4.1.2. ASR decoding error

ASReval is based on the state-of-the-art Kaldi recipe for LibriSpeech involving a factorized time delay neural network (TDNN-F) acoustic model (AM) [32, 33] and a trigram language model. As shown in Fig. 2, the (1) original and (2) anonymized trial data is decoded using the provided pretrained ASR AM trained on LibriSpeech train-clean-100 and train-other-500 using Kaldi. A 512-dimensional x-vector encoding spoken content are extracted using a TDNN and bottleneck (BN) features; (2) x-vector anonymization; (3) speech synthesis (SS) from the anonymized x-vector and the original F0+BN features. In Step 1, 256-dimensional BN features encoding spoken content are extracted using a TDNN-F ASR AM trained on LibriSpeech train-clean-100 and train-other-500 using Kaldi. A 512-dimensional x-vector encoding the speaker is extracted using a TDNN trained on VoxCeleb1.2 with Kaldi. 

4.2. Subjective metrics

Subjective metrics include speaker verifiability, speaker linkability, speech intelligibility, and speech naturalness. They will be evaluated using listening tests carried out by the organizers.

4.2.1. Subjective speaker verifiability

To evaluate subjective speaker verifiability, listeners are given pairs of one anonymized trial utterance and one distinct original enrollment utterance of the same speaker. Following [34], they are instructed to imagine a scenario in which the anonymized sample is from an incoming telephone call, and to rate the similarity between the voice and the original voice using a scale of 1 to 10, where 1 denotes ‘different speakers’ and 10 denotes ‘the same speaker’ with highest confidence. The performance of each anonymization system will be visualized through detection error tradeoff (DET) curves.

4.2.2. Subjective speaker linkability

The second subjective metric assesses speaker linkability, i.e., the ability to cluster several utterances into speakers. Listeners are asked to place a set of anonymized trial utterances from different speakers in a 1- or 2-dimensional space according to speaker similarity. This relies on a graphical interface, where each utterance is represented as a point in space and the distance between two points expresses subjective speaker dissimilarity.

4.2.3. Subjective speech intelligibility

Listeners are also asked to rate the intelligibility of individual samples (anonymized trial utterances or original enrollment utterances) on a scale from 1 (totally unintelligible) to 10 (totally intelligible). The results can be visualized through DET curves.

4.2.4. Subjective speech naturalness

Finally, the naturalness of the anonymized speech will be evaluated on a scale from 1 (totally unnatural) to 10 (totally natural).

5. Baseline software and results

Two anonymization baselines are provided.4 We briefly introduce them and report the corresponding objective results below.

5.1. Anonymization baselines

The primary baseline shown in Fig. 3 is inspired from [17] and comprises three steps: (1) extraction of x-vector [29], pitch (F0) and bottleneck (BN) features; (2) x-vector anonymization; (3) speech synthesis (SS) from the anonymized x-vector and the original F0+BN features. In Step 1, 256-dimensional BN features encoding spoken content are extracted using a TDNN-F ASR AM trained on LibriSpeech train-clean-100 and train-other-500 using Kaldi. A 512-dimensional x-vector encoding the speaker is extracted using a TDNN trained on VoxCeleb1.2 with Kaldi. In Step 2, for every source x-vector, an
Table 3: Speaker verifiability achieved by the pretrained ASV$_{eval}$ model. The primary baseline is used for anonymization.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gender</th>
<th>Anonymization</th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Enroll</td>
<td>Trial</td>
<td>EER (%)</td>
</tr>
<tr>
<td>LibriSpeech</td>
<td>Female</td>
<td>original</td>
<td>original</td>
<td>8.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>anonymized</td>
<td>original</td>
<td>50.28</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>original</td>
<td>original</td>
<td>35.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>anonymized</td>
<td>original</td>
<td>7.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>anonymous</td>
<td>26.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>original</td>
<td>29.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>anonymized</td>
<td>2.86</td>
</tr>
<tr>
<td>VCTK</td>
<td>Female</td>
<td>original</td>
<td>original</td>
<td>50.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>anonymized</td>
<td>original</td>
<td>29.48</td>
</tr>
<tr>
<td>(different)</td>
<td>Male</td>
<td>original</td>
<td>original</td>
<td>1.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>anonymized</td>
<td>original</td>
<td>55.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>anonymized</td>
<td>26.10</td>
</tr>
</tbody>
</table>

Table 4: ASR decoding error achieved by the pretrained ASV$_{eval}$ model. The primary baseline is used.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Anonymization</th>
<th>Dev. WER (%)</th>
<th>Test WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibriSpeech</td>
<td>original</td>
<td>3.83</td>
<td>4.14</td>
</tr>
<tr>
<td></td>
<td>anonymized</td>
<td>6.30</td>
<td>6.77</td>
</tr>
<tr>
<td>VCTK</td>
<td>original</td>
<td>10.79</td>
<td>12.81</td>
</tr>
<tr>
<td>(comm.+diff.)</td>
<td>anonymized</td>
<td>15.50</td>
<td>15.53</td>
</tr>
</tbody>
</table>

anneanized x-vector is computed by finding the N farthest x-vectors in an external pool (LibriTTS train-other-500) according to the PLDA distance, and by averaging N’ randomly selected vectors among them. In Step 3, an SS AM generates Mel-filterbank features given the anonymized x-vector and the F0+BN features, and a neural source-filter (NSF) waveform model [35] outputs a speech signal given the anonymized x-vector, the F0, and the generated Mel-filterbank features. The SS AM and NSF models are both trained on LibriTTS train-clean-100. See [20, 36] for further details.

The secondary baseline is a simpler, formant-shifting approach provided as additional inspiration [37].

5.2. Objective evaluation results

Table 3 reports the values of objective speaker verifiability metrics obtained before/after anonymization with the primary baseline. The EER and C$_{min}$ metrics behave similarly, while interpretation of C$_{mp}$ is more challenging due to non-calibration. We hence focus on the EER below. On all datasets, anonymization of the trial data greatly increases the EER. This shows that the anonymization baseline effectively increases the users’ privacy. The EER estimated with original enrollment data (49 to 58%), which is comparable to or above the chance value (50%), suggests that full anonymization has been achieved. However, anonymized enrollment data result in a much lower EER (26 to 35%), which suggests that F0+BN features retain some information about the original speaker. If the attackers have access to such enrollment data, they will be able to re-identify users almost half of the time. Note also that the EER is larger for female than males on average. This further demonstrates that failing to define the attack model or assuming a naive attack model leads to a greatly overestimated sense of privacy [22].

Table 4 reports the WER achieved before/after anonymization with the primary baseline. While the absolute WER stays below 7% on LibriSpeech and 16% on VCTK, anonymization incurs a large WER increase of 21 to 70% relative.

The results achieved by the secondary baseline are inferior and detailed in [20]. Overall, there is substantial potential for challenge participants to improve over the two baselines.

6. Conclusions

The VoicePrivacy initiative aims to promote the development of private-by-design speech technology. Our initial event, the VoicePrivacy 2020 Challenge, provides a complete evaluation protocol for voice anonymization systems. We formulated the voice anonymization task as a game between users and attackers, and highlighted three possible attack models. We also designed suitable datasets and evaluation metrics, and we released two open-source baseline voice anonymization systems. Future work includes evaluating and comparing the participants’ systems using objective and subjective metrics, computing alternative objective metrics relating to, e.g., requirement (d) in Section 2.2, and drawing initial conclusions regarding the best anonymization strategies for a given attack model. A revised, stronger evaluation protocol is also expected as an outcome.

In this regard, it is essential to realize that the users’ downstream goals and the attack models listed above are not exhaustive. For instance, beyond ASR decoding, anonymization is extremely useful in the context of anonymized data collection for ASR training [19]. It is also known that the EER becomes lower when the attackers generate anonymized training data and re-trains ASV$_{eval}$ on this data [22]. In order to assess these aspects, we will ask volunteer participants to share additional data with us and run additional experiments in a post-evaluation phase.

7. Acknowledgment

VoicePrivacy was born at the crossroads of projects VoicePersonae, COMPRIZE (https://www.compriseh2020.eu/), and DEEP-PRIVACY. Project HARPOCRATES was designed specifically to support it. The authors acknowledge support by ANR, JST, and the European Union’s Horizon 2020 Research and Innovation Program, and they would like to thank Md Sahidullah and Fuming Fang. 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).
8. References


