The VoicePrivacy 2020 Challenge
Evaluation Plan

Version 1.1

Natalia Tomashenko1, Brij Mohan Lal Srivastava2, Xin Wang3, Emmanuel Vincent4, Andreas Nautsch5, Junichi Yamagishi3,6, Nicholas Evans5, Jean-François Bonastre1, Paul-Gauthier Noé1, Massimiliano Todisco5, and Jose Patino5

1Laboratoire Informatique d’Avignon (LIA), Avignon Université, France
2Inria, France
3National Institute of Informatics, Tokyo, Japan
4Université de Lorraine, CNRS, Inria, LORIA, France
5Audio Security and Privacy Group, EURECOM, France
6University of Edinburgh, UK

https://voiceprivacychallenge.org

1 Context

Recent years have seen mounting calls for the preservation of privacy when treating or storing personal data. This is not least the result of recent European privacy legislation, e.g., the general data protection regulation (GDPR). While there is no legal definition of privacy [1], speech data is likely to fall within the scope of privacy regulation. This is because speech data contains much more than just the spoken words. Speech encapsulates a wealth of personal, personal data, e.g., age and gender, health and emotional state, racial or ethnic origin, geographical background, social identity, and socio-economic status [2]. Speaker recognition systems can also reveal the speaker’s identity. Since machines can now decipher spoken language with ever-impressive accuracy, there is no reason why political orientations, religious and philosophical beliefs could not also be derived from speech data.

Given that speech data can qualify as personal data, it is thus of no surprise that efforts to develop privacy preservation solutions for speech technology are starting to emerge. The study of privacy is also gaining traction; the Security and Privacy in Speech Communication (SPSC)1 Special Interest Group (SIG) of the International Speech Communication Association (ISCA) has recently been formed to champion research in the domain.

There are arguably two general solutions to preserve privacy in speech data: cryptography, and anonymization, also referred to as de-identification. Cryptographic solutions can render speech data unreadable and can only be read in plain text with access to a private key. Some techniques, e.g.

1https://www.spsc-sig.org
homomorphic encryption, even support computation upon data in the encrypted domain. These methods are typically specific to the given application, are challenging to integrate within existing systems and almost always result in significant increases to computational complexity and/or communication overheads. As such, cryptographic solutions can be relatively inflexible and cumbersome. Even then, they can only be implemented with specialist cryptography expertise.

Anonymization is different. Instead of conserving and protecting sensitive data through encryption, anonymization techniques, e.g. [3], have the goal of suppressing personally identifiable information within a speech signal, leaving all other aspects of the speech signal intact. Anonymization techniques can be highly flexible and can be integrated easily within existing systems. They usually have low computational complexity and rarely require any additional expertise outside of the speech domain. Despite the appeal of anonymization techniques and the urgency to address privacy concerns, there are currently only few solutions. In addition, a formal definition of anonymization is missing, the level of anonymization offered by some solutions is somewhat unclear, and there are no common datasets, protocols and metrics to support further work.

The VoicePrivacy initiative is spearheading the effort to develop privacy preservation solutions for speech technology. It aims to gather a new community to define the task and metrics and to benchmark initial solutions using the first common datasets and protocols. In following best practice, VoicePrivacy takes the form of a competitive challenge. Challenge participants are required to process a dataset of speech signals in order to anonymize them, while protecting the linguistic content and speech naturalness. The challenge will run from early 2020 and conclude with a special session/event held in conjunction with Interspeech 2020, at which challenge results will be made publicly available. This document describes plans for the challenge, the dataset, protocol and the set of metrics that will be used for assessment, in addition to evaluation rules and guidelines for registration and submission.

2 Challenge objectives

The grand objective of the VoicePrivacy challenge is to foster progress in the development of anonymization techniques for speech data. The specific technical goals are summarised as follows. They are to:

- facilitate the development of novel techniques which suppress speaker-discriminative information within speech signals;
- promote techniques which provide effective anonymization while protecting intelligibility and naturalness;
- provide a level playing field to facilitate the comparison of different anonymization solutions using a common dataset and protocol;
- investigate metrics for the evaluation and meaningful comparison of different anonymization solutions.

VoicePrivacy participants will be provided with a common datasets, protocols and a suite of software packages that will enable them to gauge anonymization performance. While it is likely that this will evolve in future challenges, no formal ranking of systems will be made for this first edition. This is because recommendations for the assessment of anonymization solutions form one of the technical goals.
3 Task

A privacy preservation task is typically formulated as a game involving one or more users who publish some data and an attacker (also called an adversary) who gains legal or illegal access to this data or to derived data and attempts to infer personal information about the users. To protect their privacy, the users publish data that contain as little personal information as possible while allowing one or more desired goals to be fulfilled. To infer personal information about the users, the attacker may use additional knowledge.

Considering speech data as an example, the definition of the privacy preservation task depends on the following scenario specifications [4]:

- the nature of the data (e.g., speech waveform, speech features, text, etc.),
- the pieces of information considered as personal (e.g., speaker identity, speaker traits, spoken contents, etc.),
- the desired goal(s) (e.g., human dialogue, multi-party human conversation, automatic speech recognition (ASR) training, emotion recognition, etc.),
- the data accessed by the attacker (e.g., a single utterance, multiple utterances, features computed from these utterances, an ASR model trained on these utterances, etc.),
- the additional knowledge available to the attacker (e.g., utterances previously released by the user, the measures adopted by users to protect their privacy, etc.).

Different scenario specifications lead to different privacy preservation methods from the user’s point of view and different attack methods from the attacker’s point of view.

3.1 Scenario

In VoicePrivacy 2020, we consider the following specific scenario, where the terms “user” and “speaker” are used interchangeably:

- Speakers want to hide their identity while allowing any desired goal to be potentially achieved.
- The attacker has access to a single utterance and wants to identify the corresponding speaker.

3.2 Anonymization task

In order to hide his/her identity, each speaker passes his/her utterances through an anonymization system before publication. The resulting anonymized utterances are called trial utterances. They sound as if they were uttered by another speaker, which we call pseudo-speaker which may be an artificial voice not corresponding to any real speaker.

The task of challenge participants is to develop this anonymization system. In order to allow any desired goal to be potentially achieved, this system should satisfy the following requirements:

- output a speech waveform,
- hide speaker identity as much as possible,
- distort other speech characteristics as little as possible,
• 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.

The third requirement will be assessed via a range of utility metrics (also called usability metrics). Specifically, ASR performance using a model trained on clean (non-anonymized) data and subjective speech intelligibility and naturalness will be measured during the challenge (see Section 6.1 and 6.2), and additional desired goals including ASR training will be assessed in a post-evaluation stage (see Section 8).

The fourth requirement is motivated by the fact that, in a multi-party human conversation, each speaker cannot change his/her anonymized voice over time and the anonymized voices of all speakers must be distinguishable from each other. This will be assessed via a new, specifically designed metric in the post-evaluation stage.

3.3 Attack model

We assume that the attacker has access to various amounts of data:

• one or more anonymized trial utterances,

• possibly, several additional utterances for each speaker, which may or may not have been anonymized and are called enrollment utterances.

The attacker does not have access to the anonymization system applied by the user though.

The level of protection of personal information will be assessed via a range of privacy metrics, including objective speaker verifiability metrics and subjective speaker verifiability and linkability metrics. These metrics assume different attack models:

• The objective speaker verifiability metrics assume that the attacker has access to a single anonymized trial utterance and several enrollment utterances. During the challenge, two sets of metrics will be computed, corresponding to the two situations when the enrollment utterances are clean or they have been anonymized (see Section 6.1). In the latter case, we assume that the utterances have been anonymized in the same way as the trial data using the same anonymization system, i.e., all enrollment utterances from a given speaker are converted into the same pseudo-speaker, and enrollment utterances from different speakers are converted into different pseudo-speakers. We also assume that the pseudo-speaker corresponding to a given speaker in the enrollment set is different from the pseudo-speaker corresponding to that same speaker in the trial set. In the post-evaluation stage, we will consider alternative anonymization procedures for the enrollement data, corresponding to stronger attack models (see Section 8).

• The subjective speaker veriability metric assumes that the attacker has access to a single anonymized trial utterance and a single clean enrollment utterance (see Section 6.2).

• The subjective speaker linkability metric assumes that the attacker has access to a several anonymized trial utterances and several clean enrollment utterances (see Section 6.2).

4 Data

Several publicly available corpora are used for training, development and evaluation of speaker anonymization systems. They are comprised of subsets from the following corpora:
• **LibriSpeech** [2] is a corpus of read English speech derived from audiobooks and designed for ASR research. It contains about 1,000 hours of speech sampled at 16 kHz.

• **LibriTTS** [3] is a corpus of English speech derived from the original LibriSpeech corpus and designed for text-to-speech (TTS). It contains approximately 585 hours of read English speech sampled at 24 kHz.

• **VCTK** [4] is a corpus of read speech of 109 native speakers of English with various accents. It was originally aimed for TTS synthesis systems and contains about 44 hours of speech sampled at 48 kHz.

• **VoxCeleb-1,2** [5, 8, 9] is an audiovisual corpus extracted from videos uploaded to YouTube and designed for speaker verification research. It contains about 2,770 hours of speech (16 kHz) from about 7,360 speakers, covering a wide range of accents and languages.

The detailed description of datasets provided for training, development and evaluation is given below.

### 4.1 Training data

For training an anonymization system the following corpora can be used: **VoxCeleb-1,2**, **LibriSpeech-train-clean-100**, **LibriSpeech-train-other-500**, **LibriTTS-train-clean-100** and **LibriTTS-train-other-500**. Summary statistics for these data are given in Table 1.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Size,h</th>
<th>Number of Speakers</th>
<th>Number of Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>VoxCeleb-1,2</td>
<td>2,794</td>
<td>2,912</td>
<td>4,451</td>
</tr>
<tr>
<td>LibriSpeech: train-clean-100</td>
<td>100</td>
<td>125</td>
<td>126</td>
</tr>
<tr>
<td>LibriSpeech: train-other-500</td>
<td>497</td>
<td>564</td>
<td>602</td>
</tr>
<tr>
<td>LibriTTS: train-clean-100</td>
<td>54</td>
<td>123</td>
<td>124</td>
</tr>
<tr>
<td>LibriTTS: train-other-500</td>
<td>310</td>
<td>560</td>
<td>600</td>
</tr>
</tbody>
</table>

### 4.2 Development set

Subsets from two different corpora are provided as development sets for anonymization systems. The first one is a subset of the **LibriSpeech-dev-clean** dataset. The second one (denoted as **VCTK-dev** ) is obtained from the VCTK corpus. As explained in Section 3, we split these datasets into a trial subset, which is to be anonymized, and an enrollment subset, which is to be anonymized with different pseudo-speakers in order to compute objective speaker verification metrics. The anonymized enrollment data shall be stored in a different folder than the clean enrollment data, since both clean and anonymized enrollment data will be used to computed two different sets of metrics.

---


[4] VCTK, release version 0.92: [https://datashare.is.ed.ac.uk/handle/10283/3443](https://datashare.is.ed.ac.uk/handle/10283/3443)

The statistics of these subsets are summarized in Table 2. For the LibriSpeech-dev-clean dataset, the speakers in the enrollment set are a subset of those in the trial set. For the VCTK-dev dataset, we created two subsets of trial utterances, denoted as common part and different part. Both include trials from the same set of speakers, but from disjoint subsets of utterances. The common part of the trials is composed of utterances #1 – 24 in the VCTK corpus that are identical for all speakers: the elicitation paragraph⁶ (utterances #1 – 5) and rainbow passage⁷ (utterances #6 – 24). This part is meant for subjective evaluation of speaker verifiability/linkability in a text-dependent manner. The enrollment subset and the different part of the trials are composed of distinct utterances for all speakers (utterations with indexes ≥ 25).

<table>
<thead>
<tr>
<th>Subset</th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Librispeech: dev-clean</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speakers in enrollment</td>
<td>15</td>
<td>14</td>
<td>29</td>
</tr>
<tr>
<td>Speakers in trials</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>Enrollment utterances</td>
<td>167</td>
<td>176</td>
<td>343</td>
</tr>
<tr>
<td>Trial utterances</td>
<td>1,018</td>
<td>960</td>
<td>1,978</td>
</tr>
<tr>
<td>VCTK-dev</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speakers (same in enrollment and trials)</td>
<td>15</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>Enrollment utterances</td>
<td>300</td>
<td>300</td>
<td>600</td>
</tr>
<tr>
<td>Trial utterances (common part)</td>
<td>344</td>
<td>351</td>
<td>695</td>
</tr>
<tr>
<td>Trial utterances (different part)</td>
<td>5,422</td>
<td>5,255</td>
<td>10,677</td>
</tr>
</tbody>
</table>

### 4.3 Evaluation data

Similarly to the development data, subsets from two different corpora are used for evaluation. The first one is a subset of the LibriSpeech-test-clean dataset. The second one (denoted as VCTK-test) is obtained from the VCTK corpus. We split those datasets into enrollment and trial subsets as summarized in Table 3. For the VCTK dataset, we created two subsets of trial utterances, denoted as common part and different part, in a similar manner as for the development set.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Librispeech: test-clean</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speakers in enrollment</td>
<td>16</td>
<td>13</td>
<td>29</td>
</tr>
<tr>
<td>Speakers in trials</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>Enrollment utterances</td>
<td>254</td>
<td>184</td>
<td>438</td>
</tr>
<tr>
<td>Trial utterances</td>
<td>734</td>
<td>762</td>
<td>1496</td>
</tr>
<tr>
<td>VCTK-test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speakers (same in enrollment and trials)</td>
<td>15</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>Enrollment utterances</td>
<td>300</td>
<td>300</td>
<td>600</td>
</tr>
<tr>
<td>Trial utterances (common part)</td>
<td>346</td>
<td>354</td>
<td>700</td>
</tr>
<tr>
<td>Trial utterances (different part)</td>
<td>5,328</td>
<td>5,420</td>
<td>10,748</td>
</tr>
</tbody>
</table>

⁶Elicitation paragraph: [http://accent.gmu.edu/pdfs/elicitation.pdf](http://accent.gmu.edu/pdfs/elicitation.pdf)

⁷Rainbow passage: [https://www.dialectsarchive.com/the-rainbow-passage](https://www.dialectsarchive.com/the-rainbow-passage)
5 Baseline

Our baseline system is based on the speaker anonymization method proposed in [3] and shown in Figure 1. The anonymization is performed in three steps:

- **Step 1: Feature extraction**: extract the speaker x-vector [10], the fundamental frequency (F0) and bottleneck (BN) features from the original audio waveform.

- **Step 2: X-vector anonymization**: anonymize the x-vector of the source speaker using an external pool of speakers.

- **Step 3: Speech synthesis**: synthesize the speech waveform from the anonymized x-vector and the original BN and F0 features using an acoustic model and a neural waveform model.

![Figure 1: Baseline anonymization system.](image)

In order to implement these steps, four different models are required as shown in Figure 1. Details for training these components are presented in Table 4.

In Step 1, to extract BN features, an ASR acoustic model (AM) is trained (#1 in Table 1). We assume that these BN features represent the linguistic content of the speech signal. The ASR AM has a factorized time delay neural network (TDNN-F) model architecture [12,13] and is trained using the Kaldi toolkit [14]. To encode speaker information, an x-vector extractor with a TDNN model topology (#2 in Table 1) is also trained using Kaldi.

In Step 2, for a given source speaker, a new anonymized x-vector is computed by averaging a set of candidate x-vectors from the speaker pool for which the similarity to the x-vector of the source speaker is in the given range. The cosine distance \(\cos(x_1, x_2)\) or, optionally, probabilistic linear discriminant analysis (PLDA) distance is used as a similarity measure between two x-vectors \(x_1\) and \(x_2\). The candidate x-vectors for averaging are chosen in two steps. First, for a given x-vector, \(N\) the most farthest candidates from the speaker pool are selected. Second, a smaller subset of \(N^*\) x-vector candidates from this set are chosen randomly\(^9\). The x-vectors for the speaker pool are extracted from a disjoint dataset (LibriTTS-train-other-500).

In Step 3, two modules are used to generate the speech waveform: a speech synthesis AM that generates Mel-filterbank features given the F0, the anonymized x-vector, and the BN features, and a neural source-filter (NSF) waveform model [11] that produces a speech waveform given the F0, the anonymized x-vector, and the generated Mel-filterbanks. Both models (#3 and #4 in Table 1) are trained on the same corpus (LibriTTS-train-clean-100).

More details about the baseline recipe can be found in the provided scripts.

\(^8\)Baseline is available online: [https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020](https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020)

\(^9\)In the baseline, the following parameter values are used: \(N = 200\) and \(N^* = 100\).
Table 4: Baseline anonymization system: models and corpora. The model indexes are the same as in Figure 1. Superscript numbers represent feature dimensions.

<table>
<thead>
<tr>
<th>#</th>
<th>Model</th>
<th>Description</th>
<th>Output features</th>
<th>Training dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ASR AM</td>
<td>TDNN-F&lt;br&gt;Input: MFCC&lt;sup&gt;40&lt;/sup&gt; + i-vectors&lt;sup&gt;100&lt;/sup&gt;&lt;br&gt;17 TDNN-F hidden layers&lt;br&gt;Output: 6032 triphone ids&lt;br&gt;LF-MMI and CE criteria</td>
<td>BN&lt;sup&gt;256&lt;/sup&gt; features extracted from the final hidden layer</td>
<td>Librispeech: train-clean-100 train-other-500</td>
</tr>
<tr>
<td>2</td>
<td>X-vector extractor</td>
<td>TDNN&lt;br&gt;Input: MFCC&lt;sup&gt;30&lt;/sup&gt;&lt;br&gt;7 hidden layers + 1 stats pooling layer&lt;br&gt;Output: 7232 speaker ids&lt;br&gt;CE criterion</td>
<td>speaker x-vectors&lt;sup&gt;512&lt;/sup&gt;</td>
<td>VoxCeleb: 1, 2</td>
</tr>
<tr>
<td>3</td>
<td>Speech synthesis AM</td>
<td>Autoregressive (AR) network&lt;br&gt;Input: F0&lt;sup&gt;1&lt;/sup&gt; + BN&lt;sup&gt;256&lt;/sup&gt; + x-vectors&lt;sup&gt;512&lt;/sup&gt;&lt;br&gt;FF * 2 + BLSTM + AR + LSTM * 2 + highway-postnet&lt;br&gt;MSE criterion</td>
<td>Mel-filterbanks&lt;sup&gt;80&lt;/sup&gt;</td>
<td>LibriTTS: train-clean-100</td>
</tr>
<tr>
<td>4</td>
<td>NSF model</td>
<td>sinc-Ir-NSF in [11]&lt;br&gt;Input: F0&lt;sup&gt;1&lt;/sup&gt; + Mel-fbanks&lt;sup&gt;80&lt;/sup&gt; + x-vectors&lt;sup&gt;512&lt;/sup&gt;&lt;br&gt;STFT criterion</td>
<td>speech waveform</td>
<td>LibriTTS: train-clean-100</td>
</tr>
</tbody>
</table>

Pool of speaker x-vectors | LibriTTS: train-other-500 |

6 Utility and privacy metrics

Following the attack model assumed in Section 3.3, we consider objective and subjective privacy metrics to assess speaker re-identification and linkability. We also propose objective and subjective utility metrics in order to assess the fulfillment of the user goals specified in Section 3.2. Specifically, we consider ASR performance using a model trained on clean data and subjective speech intelligibility and naturalness.

Objective metrics will be computed by the participants themselves using the provided evaluation scripts and evaluation data, while subjective metrics will be computed the organizers using data provided by the participants.

6.1 Objective metrics

For objective evaluation of anonymization performance, two systems will be trained to assess the following characteristics: (1) speaker verifiability and (2) ability of the anonymization system to preserve linguistic information in the anonymized speech. The first system denoted as ASV<sub>eval</sub> is an automatic speaker verification (ASV) system, which produces log-likelihood ratio (LLR) scores. The second system denoted as ASR<sub>eval</sub> is an automatic speech recognition (ASR) system which outputs a word error rate (WER) metric.

Both ASR<sub>eval</sub> and ASV<sub>eval</sub> are trained on the LibriSpeech-train-clean-360 dataset using the Kaldi speech recognition toolkit [14]. The statistics of this dataset are given in Table 5. Training of ASR<sub>eval</sub> and ASV<sub>eval</sub> and evaluation will be done with the provided recipes<sup>10</sup>.

<sup>10</sup>Evaluation scripts: https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020
Table 5: Statistics of the training dataset for the ASV$_\text{eval}$ and ASR$_\text{eval}$ evaluation systems.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Size,h</th>
<th>Number of Speakers</th>
<th>Number of Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>LibriSpeech: train-clean-360</td>
<td>363.6</td>
<td>439</td>
<td>482</td>
</tr>
</tbody>
</table>

Figure 2: ASR evaluation for (1) clean speech data and (2) anonymized speech data. WER is estimated on the trial utterances of the development and evaluation datasets.

6.1.1 Word error rate (WER)

ASR performance will be assessed using ASR$_\text{eval}$ which is based on the state-of-the-art Kaldi recipe for LibriSpeech involving a TDNN-F acoustic model and a trigram language model. The recipe has been adapted to run on the LibriSpeech-train-clean-360 dataset rather than the full LibriSpeech training corpus in order to speed up the evaluation process.

As shown in Figure 2, the anonymized development and evaluation data will be decoded using the provided pretrained system ASR$_\text{eval}$ and the word error rate (WER) will be calculated:

$$\text{WER} = \frac{N_{\text{sub}} + N_{\text{del}} + N_{\text{ins}}}{N_{\text{ref}}}$$

with $N_{\text{sub}}$, $N_{\text{del}}$, and $N_{\text{ins}}$, the number of substitution, deletion, and insertion errors, respectively, and $N_{\text{ref}}$ the number of words in the reference. This will be compared to the WER achieved on clean data. The WER is calculated only on the trial part of the development and evaluation datasets.

6.1.2 Speaker verifiability metrics

The ASV$_\text{eval}$ system for speaker verification is based on state-of-the-art x-vector speaker embeddings with a probabilistic linear discriminant analysis (PLDA) backend [10] as implemented in Kaldi [14]. It will be applied as shown in Figure 3:

1. Compute PLDA (LLR) scores for (a) clean enrollment data and (b) anonymized trial data;
2. Compute PLDA (LLR) scores for (a) anonymized enrollment data and (b) anonymized trial data;
3. For steps 1 and 2, estimate the corresponding speaker verifiability metrics: equal error rate (EER) and log-likelihood-ratio cost function ($C_{\text{llr}}$).

The number of target and impostor trials in the development and evaluation datasets are given in Tables 6 and 7, respectively.
Figure 3: ASV evaluation for (1) clean trial and enrollment data; (2) anonymized trial data and clean enrollment data; and (3) anonymized trial and enrollment data.

Table 6: Number of trials in the development datasets.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Trials</th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Librispeech: dev-clean</td>
<td>Target</td>
<td>704</td>
<td>644</td>
<td>1,348</td>
</tr>
<tr>
<td></td>
<td>Impostor</td>
<td>14,566</td>
<td>12,796</td>
<td>27,362</td>
</tr>
<tr>
<td>VCTK-dev</td>
<td>Target (common part)</td>
<td>344</td>
<td>351</td>
<td>695</td>
</tr>
<tr>
<td></td>
<td>Target (different part)</td>
<td>1,781</td>
<td>2,015</td>
<td>3,796</td>
</tr>
<tr>
<td></td>
<td>Impostor (common part)</td>
<td>4,810</td>
<td>4,911</td>
<td>9,721</td>
</tr>
<tr>
<td></td>
<td>Impostor (different part)</td>
<td>13,219</td>
<td>12,985</td>
<td>26,204</td>
</tr>
</tbody>
</table>

Table 7: Number of trials in the evaluation datasets.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Trials</th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Librispeech: test-clean</td>
<td>Target</td>
<td>548</td>
<td>449</td>
<td>997</td>
</tr>
<tr>
<td></td>
<td>Impostor</td>
<td>11,196</td>
<td>9,457</td>
<td>20,653</td>
</tr>
<tr>
<td>VCTK-test</td>
<td>Target (common part)</td>
<td>346</td>
<td>354</td>
<td>700</td>
</tr>
<tr>
<td></td>
<td>Target (different part)</td>
<td>1,944</td>
<td>1,742</td>
<td>3,686</td>
</tr>
<tr>
<td></td>
<td>Impostor (common part)</td>
<td>4,838</td>
<td>4,952</td>
<td>9,790</td>
</tr>
<tr>
<td></td>
<td>Impostor (different part)</td>
<td>13,056</td>
<td>13,258</td>
<td>26,314</td>
</tr>
</tbody>
</table>
Equal error rate (EER)

The EER is computed from the PLDA scores as follows. Let $P_{fa}(\theta)$ and $P_{miss}(\theta)$ denote the false alarm and miss rates at threshold $\theta$:

$$P_{fa}(\theta) = \frac{\#\{\text{impostor trials with score} > \theta\}}{\#\{\text{total impostor trials}\}},$$

$$P_{miss}(\theta) = \frac{\#\{\text{target trials with score} \leq \theta\}}{\#\{\text{total target trials}\}},$$

so that $P_{fa}(\theta)$ and $P_{miss}(\theta)$ are, respectively, monotonically decreasing and increasing functions of $\theta$. EER corresponds to the threshold $\theta_{EER}$ at which the two detection error rates are equal, i.e., $EER = P_{fa}(\theta_{EER}) = P_{miss}(\theta_{EER})$.

Log-likelihood-ratio cost function ($C_{llr}$ and $C_{llr}^{\text{min}}$)

In addition to the EER, an alternative performance measure, the log-likelihood-ratio cost function ($C_{llr}$), will be estimated. It was proposed in [15] as an application-independent evaluation objective and is defined as follows:

$$C_{llr} = \frac{1}{2} \left( \frac{1}{N_{\text{tar}}} \sum_{i \in \text{targets}} \log_2 \left( 1 + e^{-\text{LLR}_i} \right) + \frac{1}{N_{\text{imp}}} \sum_{j \in \text{impostors}} \log_2 \left( 1 + e^{\text{LLR}_j} \right) \right),$$

where $N_{\text{tar}}$ and $N_{\text{imp}}$ are respectively the number of target and impostor LRR values in the evaluation set.

$C_{llr}$ can be decomposed into a discrimination loss ($C_{llr}^{\text{min}}$) and a calibration loss ($C_{llr} - C_{llr}^{\text{min}}$) [15]. The $C_{llr}^{\text{min}}$ is estimated by optimal calibration using monotonic transformation of scores to their empirical LLR values. To obtain this monotonic transformation, the pool adjacent violators (PAV) to LLR algorithm is used [15,16].

6.2 Subjective metrics

Several subjective metrics will be used in evaluation: speaker verifiability, speaker linkability, speech intelligibility, and speech naturalness. They will be evaluated using listening tests that will be carried out by the organizers as described below.

6.2.1 Subjective speaker verifiability

To evaluate subjective speaker verifiability, the following two approaches can be used. The first approach is to measure the speaker similarity between one anonymized trial utterance and one clean enrollment utterance through a large-scale crowdsourced evaluation, following the approach used in a recent study on speech anti-spoofing [17]. Given one anonymized trial utterance and one randomly selected clean enrollment utterance from the same original speaker, the subjects will be instructed

---

11In this context, application is defined as a set of prior probabilities and decision costs involved in the inferential process [15].

12How good are two classes separated for any threshold?

13In the light of domain shifts (e.g., changing speech signal quality between training and evaluation sets), scores loose the capability of correctly encoding the expected class ratio for what their value suggests; scores are not LLRs anymore and thus lose their goodness for ideal decision making. The calibration loss quantifies this gap; scores are considered to be LLRs if $C_{llr} \leq 1$, and for $C_{llr} = 1$, a coin toss provides the same information.
to imagine a scenario in which the anonymized sample is from an incoming telephone call, and the subjects need to judge whether the voice in the call is similar to the clean voice of the claimed speaker. The instruction given to the subjects is as follows:

Imagine you are working for a bank call center. Your task is to compare customer inquiries with voices recorded when the same customer made inquiries in the past. You must determine whether the voices are of the same person or another person who is impersonating the original voices. However, if you choose ‘spoofing by someone else’ more than necessary, there will be many complaints from real customers, which should be avoided. Imagine a situation in which you are working to protect bank accounts and balance convenience. Now press the ‘Sample A’ and ‘Sample B’ buttons below and listen to the samples. You can listen to them as many times as you like. Use only the audio you hear to determine if the speakers are the same or not. The content of the conversation in English is irrelevant and does not need to be heard. Please judge on the basis of the characteristics of the voice, not the content of the words.

As the instruction describes, the subject will listen to two samples (i.e., Sample A and Sample B) in one evaluation page. Sample B is always a randomly selected enrollment utterance of a given original speaker. Sample A may be an anonymized trial utterance or a clean trial utterance of the same original speaker or a different speaker. The subject will be asked to evaluate the similarity using a scale of 1 to 10, where 1 denotes ‘different speakers’ and 10 denotes ‘the same speaker’ with highest confidence. Note that, by using clean speech of the same original speaker or a different speaker as Sample A, we will have anchors in the listening test and can visualize the performance of each participant system through Detection Error Tradeoff (DET) curves.

### 6.2.2 Subjective speaker linkability

The second subjective evaluation metric is based on speaker linkage, i.e., clustering. The listeners will be asked to place a set of anonymized trial utterances and clean enrollment utterances from different speakers in a 1- or 2-dimensional space according to speaker similarity. This will be done using a specially developed graphical interface, where each utterance is represented by a point in the space. The distance between two points expresses the subjective perception of speaker dissimilarity between the corresponding utterances.

### 6.2.3 Subjective speech intelligibility

Naturalness of the anonymized speech will also be evaluated through a large-scale crowdsourced evaluation. For one subjective evaluation round, the subject will listen to a sample and evaluate its naturalness using a scale from 1 (completely unintelligible) to 10 (completely intelligible and all content is clear). The sample can be an anonymized trial utterance or a clean enrollment utterance from a randomly selected speaker. The results can be visualized through DET curves for comprehensive comparison.

### 6.2.4 Subjective speech naturalness

Similar to the subjective speech intelligibility evaluation, naturalness of the anonymized speech will also be evaluated through a large-scale crowdsourced evaluation. The subject will listen to a sample and evaluate its naturalness using a scale from 1 (completely unnatural) to 10 (completely natural). The sample can be an anonymized trial utterance or a clean enrollment utterance from a randomly
selected speaker. The performance of the participants can be visualized through DET curves for comprehensive comparison.

7 Evaluation rules

- Participants are free to develop their own anonymization system, using parts of the baseline or not. They are also free to tune the compromise between speaker verifiability and speech quality/intelligibility according to their own preference.

- Participants can use only the training datasets and the enrollment datasets specified in Section 4.1 in order to train their system and tune its hyperparameters. Using additional speech data is not allowed.

- Speaker anonymization must be done in a speaker-to-speaker manner. All enrollment utterances from a given speaker must be converted into the same pseudo-speaker, and enrollment utterances from different speakers must be converted into different pseudo-speakers. Also, the pseudo-speaker corresponding to a given speaker in the enrollment set must be different from the pseudo-speaker corresponding to that same speaker in the trial set.

- Modifications of $ASV_{eval}$ and $ASR_{eval}$ (such as changing decoder parameters) are not allowed.

- For every tested system, you should report all objective metrics ($WER$, $EER$, $C_{llr}$ and $C_{llr}^{min}$) on the two development datasets and the two evaluation datasets. You should compute these metrics yourself using the provided evaluation scripts. The organizers will be responsible for computing the subjective metrics and the post-evaluation metrics only.

8 Post-evaluation analysis

Following Deadline-2, the organizers will run additional evaluation experiments in order to further characterize the performance of the submitted systems and pave the way for the next Voice Privacy Challenge. To do so, we will ask volunteer participants to share with us the anonymized speech data obtained when running their anonymization system on the training/development datasets and on the evaluation dataset with different settings, and we will compute additional evaluation metrics on these data. The full details of the additional data to be provided by participants and the additional metrics which will be computed will be disclosed in due time.

Roughly speaking, this involves:

1. retraining the evaluation systems ($ASR_{eval}$ and $ASV_{eval}$) on anonymized training data in order to evaluate the suitability of the proposed anonymization technique for ASR training;

2. computing subjective evaluation metrics (as described in Section 6.2) on selected subsets of evaluation data;

3. computing additional metrics to assess, e.g., the fulfillment of the goal that pseudo-speakers are sufficiently different from each other and the speaker verifiability performance in the presence of an attacker with additional knowledge.
9 Registration and submission of results

9.1 General mailing list

All participants and team members are encouraged to subscribe to the general mailing list. Subscription can be done by sending an email to:

sympa@lists.voiceprivacychallenge.org

with ‘subscribe 2020’ as the subject line. Successful registrations are confirmed by return email. To post messages to the mailing list itself, emails should be addressed to:

2020@lists.voiceprivacychallenge.org

9.2 Registration

Participants/teams are requested to register for the evaluation. Registration should be performed once only for each participating entity and by sending an email to:

organisers@lists.voiceprivacychallenge.org

with ‘VoicePrivacy 2020 registration’ as the subject line. The mail body should include: (i) the name of the team; (ii) the name of the contact person; (iii) their country; (iv) their status (academic/non-academic).

9.3 Submission of results

Each single submission should include:

1. Metric values for objective assessment (WER, EER, $C_{llr}$ and $C_{min}^{llr}$) on the evaluation and development datasets according to the common protocol;

2. Corresponding PLDA (LLR) scores in Kaldi format (for development and evaluation data);

3. Corresponding anonymized speech data (wav files with the same names as in the original corpus) generated from the evaluation dataset. These data will be used by the challenge organizers to verify the submitted scores, make post-evaluation analysis with other metrics and to run listening tests for subjective evaluation\(^\text{14}\).

All the anonymized speech data should be submitted in the form of a single compressed archive. Each participant should also submit a single, detailed system description. All submissions should be made according to the schedule below. Submissions received after the deadline will be marked as ‘late’ submissions, without exception. System descriptions will be shared among all registered participants, but will not be disclosed publicly. Further details concerning the submission procedure and URI will be published via the participants mailing list and via the VoicePrivacy Challenge website.

\(^{14}\)Please note that subjective evaluation will be performed for Deadline-2 submissions only, not for Deadline-1.
10 Special session at Interspeech 2020

A special session on the VoicePrivacy 2020 Challenge will be held at Interspeech 2020 in Shanghai, China.

All participants are encouraged (but not obliged) to submit their papers, dedicated to early version of their challenge entry, to the special session of the conference by 30th March (which corresponds to the Deadline-1 in the Challenge schedule, see Section 11). The review process for papers submitted to the special session is the same as for regular papers. Scientific novelty of the proposed ideas should be clearly stated and evaluated experimentally. Subjective evaluation results will be obtained only after Deadline-2 and can be included (optionally) in the camera-ready version of the paper. Other participants may submit papers to the challenge only.

Accepted Interspeech papers and other Challenge submissions will both be presented at the special session in the form of posters. These two types of presentations will be clearly differentiated in the program.

11 Schedule

There will be two submission deadlines (early Deadline-1 and late Deadline-2) in the Challenge. Participants may submit to one or both deadlines. Interspeech paper submission is encouraged but optional. All participants will be invited to present their work at the VoicePrivacy session/event.

Table 8: Important dates

<table>
<thead>
<tr>
<th>Event</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Release of training and development data</td>
<td>8th February 2020</td>
</tr>
<tr>
<td>Release of evaluation data</td>
<td>15th February 2020</td>
</tr>
<tr>
<td>Deadline-1 for participants to submit objective evaluation results</td>
<td>30th March 2020</td>
</tr>
<tr>
<td>Interspeech 2020 paper submission deadline</td>
<td>30th March 2020</td>
</tr>
<tr>
<td>Deadline-2 for participants to submit objective evaluation results and data</td>
<td>1st June 2020</td>
</tr>
<tr>
<td>Submission of system descriptions</td>
<td>8th June 2020</td>
</tr>
<tr>
<td>Submission of additional data (optional - see Section 8)</td>
<td>15th June 2020</td>
</tr>
<tr>
<td>Organizers return subjective evaluation results to participants</td>
<td>Early/mid July 2020</td>
</tr>
<tr>
<td>VoicePrivacy special session/event at Interspeech 2020</td>
<td>14th–18th September 2020</td>
</tr>
<tr>
<td>Journal special issue submission deadline</td>
<td>Early 2021</td>
</tr>
</tbody>
</table>

12 Acknowledgement

This work was supported in part by the French National Research Agency under projects HAR-POCRATES (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. The authors would like to thank Md Sahidullah and Fuming Fang.
References


