A Comprehensive Evaluation Framework for Speaker Anonymization Systems
Carlos Franzreb\textsuperscript{1}, Tim Polzehl\textsuperscript{1}, Sebastian Möller\textsuperscript{2}
\textsuperscript{1}German Research Center for Artificial Intelligence, Germany
\textsuperscript{2}Technical University of Berlin, Germany
carlos.franzreb@dfki.de

Abstract
Speaker anonymization consists of concealing the source speaker’s identity, while keeping the linguistic and paralinguistic content intact. It is usually evaluated as a trade-off between privacy and utility. The current standard for privacy evaluation is the automatic speaker verification system (ASV) from the Voice Privacy Challenge (VPC); it involves computing speaker embeddings and comparing them with a trained PLDA algorithm. We implement this ASV system and extend the utility evaluation of the VPC, which previously consisted of automatic speech recognition performance, with emotion preservation, naturalness and performance metrics. Our framework is fast and easily customizable, facilitating the development and evaluation of new anonymization pipelines. We showcase this framework with the StarGANv2-VC, one of the most powerful voice conversion systems available.

Index Terms: speaker anonymization, privacy, automatic speaker verification, voice conversion

1. Introduction
Speaker anonymization refers to the task of removing information from the speech signal that may be used to identify the speaker [1]. This concealment must be achieved while preserving the utility of the signal, either by humans or for further processing steps. Therefore, speaker anonymization is often seen as a trade-off between privacy and utility, where the relationship between them is determined based on specific use cases.

The scope and importance of anonymization was outlined decades ago [2], motivated by public concerns regarding the advances of biometric technologies. However, the research and developments focused on face, iris and fingerprint recognition [3]. As of 2019, there were few solutions for preserving privacy in speech, although its use for biometric purposes was steadily increasing. Recent changes in legislation, especially the European data protection regulation (GDPR), have dramatically increased the academic interest on speech privacy. The VoicePrivacy initiative [4] has led these efforts by designing an evaluation framework and organizing challenges to facilitate and encourage work on this topic.

The latest version of this framework was used for the VoicePrivacy Challenge (VPC) 2022 [1]. Privacy is seen as the ability to conceal the speaker’s identity by modifying solely the voice, ignoring the information about the speaker that may be present in the linguistic channel. It is measured with two different attack scenarios, which differ in whether the attacker has access to the anonymization model. They both perform automatic speaker verification (ASV) based on x-vector speaker embeddings [5], which are compared with a probabilistic linear discriminant analysis (PLDA) algorithm [6]. The utility of the anonymized speech is measured with the word error rate (WER) computed with the predictions of an ASR model from Kaldi. Their framework also includes pitch correlation and gain of voice distinctiveness as secondary metrics. Although these are not considered for the challenge results, submissions are expected to reach a minimum pitch correlation threshold.

The entire VPC evaluation framework is based on Kaldi, a speech processing toolkit written in C++. Nowadays, most work in speech processing is done in Python, mainly because of the deep learning frameworks are implemented in Python, like PyTorch. Therefore, new researchers in speech processing usually have little to no experience with C++ and Kaldi, making the VPC framework hard to understand and use. Our contribution is a new open source evaluation framework\textsuperscript{1} that extends the one from the VPC, implemented entirely in Python and PyTorch. Popular speech processing toolkits, which are usually also written in PyTorch, can be easily integrated. We leverage pre-trained models to avoid training additional models during the evaluation procedure, which dramatically increases the runtime. The modularity and documentation facilitate researchers to use and further develop the framework.

We implement the same ASV system as the VPC, using speaker recognition models from SpeechBrain [7]. New speaker recognition models can be easily added by creating a wrapper. For utility, we use the WER computed with transcriptions from Whisper [8], arguably the state-of-the-art ASR model. We extend the utility evaluation by adding three further components, which help define the strengths and weaknesses of anonymization models, as well as understand for which use cases they are suited. We measure emotion preservation by comparing the emotion embeddings and predicted dimensions with a speech emotion recognition (SER) model based on wav2vec2.0 [9], and the naturalness of the synthesized speech with NISQA-TTS [10], a CNN-LSTM network trained on data from the Blizzard and Voice Conversion Challenges. These two metrics, together with the WER, define how well the anonymized speech resembles the original speech. Finally, we measure the inference speed of the anonymization model both on CPU and GPU, as well as its throughput on GPU. These metrics highlight the hardware requirements to train and run these models, as well as whether they are candidates for real-time processing.

Another drawback of the VPC evaluation is the lack of diversity of the evaluation data. We propose a new set of evaluation datasets which more closely resemble the real world. Alongside LibriSpeech [11], which is already part of the VPC, we include EdAcc [12], a new dataset with spontaneous speech...
recorded with a variety of microphones in real-world settings, a subset of Common Voice [13], characterized by its size and diversity, and RAVDESS [14], mainly used to evaluate the preservation of emotion. These datasets contain information about the speakers, which we use to evaluate the performance of the anonymization models for different segments of the population.

To showcase this framework, we evaluate the StarGANv2-VC [15], currently one of the best voice conversion models. We experiment with several speaker recognition models from NeMo [16] and SpeechBrain [7] to determine the best configuration for the ASV system, as well as summarizing the results of the utility evaluation.

2. Related work

The Voice Privacy Challenge offers the main evaluation framework for speech privacy. The researchers that don’t use this framework often use a simplified version where training and enrollment data are not anonymized, and the cosine distance is used instead of the PLDA algorithm (e.g. [17]). A few novel ways of evaluating anonymization pipelines have been proposed recently. We describe them in this section.

VoicePM [18] proposes a metric that combines an ASV system with a set of classifiers trained to discriminate several speaker attributes like age, gender and accent. These attributes are also sensitive information and should be concealed. Their ASV system consists of comparing SpeechBrain’s ECAPA-TDNN speaker embeddings with cosine distances. The predicted attributes for the anonymized speech samples are compared with the true attributes with the Jaccard similarity index. The indices are divided by those achieved for the baseline, where the speech is not anonymized. The resulting value is combined with the EER of the ASV system in a weighted sum. This combination describes the attacker’s ability to identify the source speaker, and also how accurately the attacker can infer the source speaker’s attributes.

The impact of the vocoder on speaker anonymization has been proven to be significant [19], resulting in a new attack model. The authors of this study show that the distance between the speaker embeddings of the target speakers and those of the anonymized utterances is close to the distance between the embeddings of the source and target speakers, meaning that the vocoder influences the identity of the anonymized speech. This result is confirmed by the EERs achieved for the different domains: when using the anonymized speaker embeddings extracted, the EERs are over three times larger than when the target embeddings are used. This means that most of the anonymization is provided by the vocoder, and not by the target selection. The authors also show that the impact of the vocoder can be reversed by learning the drift it imposes on the target embeddings with a neural network. This network ingests anonymized embeddings and returns the target embeddings, i.e. those that are picked before the synthesis. The resulting attack model is more powerful than the ignorant attack scenario, where the training and enrollment data of the ASV system are not anonymized. For some cases, it even performs better than the lazy-informed attack scenario, where the ASV system has access to the anonymization model. We plan on adding this attack to our framework in the future.

3. Evaluation framework

The evaluation of anonymization models consists of measuring the achieved privacy, which is expected to be high, and the loss in utility, which should be low. Here we discuss the components of our framework, which measure both the privacy gain and the utility loss, as well as the training and evaluation datasets, which are chosen to maximize diversity and simulate real-world scenarios.

Before running the evaluation, all evaluation samples are anonymized with consistent target speakers, meaning that each speaker is always mapped to the same target. These anonymized samples are then evaluated by all the different components, including the ASV system.

3.1. Privacy

Privacy is measured by the effectiveness of an attack with an Automatic Speaker Verification (ASV) system [20]. ASV is the task of determining whether two utterances belong to the same speaker or not. It is different from Automatic Speaker Identification (ASI), where the speaker must be identified among a pool of candidates. One of the challenges of ASV is that the utterances that are verified belong to speakers that have never been seen before, meaning that the system has to generalize what it has learned from its training data. On the other hand, ASI systems exploit the peculiarities of the candidate speakers to pick the correct one, as it has seen them during training. Therefore, the population size is independent for ASV but highly relevant for ASI.

We implement the ASV system from the Voice Privacy Challenge (VPC). It comprises a speaker recognition model, which outputs embeddings that capture speaker traits [21], and a PLDA algorithm trained to distinguish between different speakers, based on these embeddings. For each pair of trial and enrollment utterances, the PLDA algorithm outputs a likelihood that they belong to the same speaker. There are two attack scenarios, illustrated in Figure 1:

1. **Ignorant**: the attacker does not have access to the anonymization model. The training data of the ASV system and the enrollment data are not anonymized.

2. **Lazy-informed**: the attacker has access to the anonymization model. The training data of the ASV system is anonymized without consistent targets, as this yields better results [1], and the enrollment data is anonymized with consistent targets.

Probabilistic Linear Discriminant Analysis (PLDA) [6] is a probabilistic extension of the Linear Discriminant Analysis (LDA) algorithm, which projects the data into a subspace where the between-class covariance is maximal and the within-class covariance is minimal. PLDA extends this notion to unseen classes by modeling the distribution of class means and within-class covariances. [22] shows that PLDA is the best similarity metric for ASV. To train it, we first filter out samples that are shorter than 4 seconds, and afterwards speakers that comprise less than 2 utterances. If we are considering the lazy-informed
scenario, the remaining utterances are anonymized with inconsistent targets. The PLDA algorithm is trained with the centered speaker embeddings of the utterances, together with their speaker labels. We use an existing PLDA implementation\footnote{https://github.com/Ravi10oji/plda} that uses empirical Bayes for parameter estimation.

When running the ASV evaluation, the dataset is split into trial and enrollment utterances. Each speaker must have an utterance in the trial set, and these utterances are replaced with their anonymized versions from the inference run. The remaining utterances of each speaker are added to the enrollment set. If we are considering the lazy-informed scenario, the enrollment utterances are anonymized with consistent targets. The attacker does not know which target was picked for each trial utterance, so it uses the target selection algorithm to pick one. It is possible for the enrollment data to be anonymized with the same target as the trial utterance of the same source speaker.

For each pair of trial and enrollment utterances, we compute the log-likelihood ratio (LLR) of their speaker embeddings with the trained PLDA model. The LLR expresses the model’s belief that the utterances belong to the same speaker or not. All the LLRs together with their corresponding true labels are used to compute the equal error rate (EER). The EER is the point at which the false positive rate and the false negative rate are equal. It represents the balance between these two kinds of errors; a higher value means that the ASV system is worse at identifying anonymized speech.

3.2. Utility

To measure the utility of the anonymized speech, we have gathered several components that together delimit the anonymization’s behavior. Our aim is to use evaluation models that generalize well to all datasets, to avoid having to fine-tune them. This greatly simplifies the evaluation procedure and increases the fairness of the comparisons between anonymization models, as they are evaluated with exactly the same models. Fine-tuning may lead the models to exploit spurious behavior of the anonymization systems or the datasets, delivering results that don’t explain the model’s true performance. When humans are asked to evaluate a sample, they usually are not given training samples to calibrate their thinking [8]; they perform out-of-distribution evaluation. We set the same requirements for the utility components.

3.2.1. Automatic Speech Recognition (ASR)

This is the most common utility evaluation, used to assess the performance of text-to-speech and voice conversion models in many publications, and also in the VPC. It determines whether the anonymization system preserves the linguistic information [1]. We use Whisper [8], currently one of the best ASR models. It is an encoder-decoder Transformer trained on 680,000 hours of multilingual speech scraped from the Internet. The size and inherent diversity of the dataset make the resulting model robust to noise, accent and choice of words.

We use two different sizes of Whisper, small and large, to assess the utility of the anonymized speech for different use cases. Smaller ASR models, characterized by their compact architectures and lower computational requirements, can be deployed on resource-constrained devices or in real-time applications. On the other hand, larger ASR models exhibit superior recognition accuracy, making them preferable for applications demanding higher transcription quality where there are fewer resource constraints. Evaluating anonymization systems against a range of ASR model sizes helps capture the performance differences and enables a more comprehensive assessment of the system’s robustness and adaptability across different deployment scenarios.

We compute the word error rate (WER) between the reference transcriptions and the Whisper predictions. Although this metric has its drawbacks, like that it ignores the semantic correctness of the prediction [23], it is efficient to compute and gives a good estimate of the prediction’s accuracy.

3.2.2. Emotion Preservation

The emotional content of speech captures the intent of what is being said, and is often crucial to determine the meaning of an utterance. To identify the emotion from an utterance, we use a speech emotion recognition (SER) model based on wav2vec2.0 [9], which was trained on the MSP-Podcast dataset [24] and is publicly available in HuggingFace. The MSP-Podcast dataset labels emotions with three dimensions: valence, arousal and dominance. The resulting model achieves state-of-the-art performance for valence prediction, which is the most elusive dimension, given that it is heavily influenced by what is being said. The SER model generalizes to other datasets better than a convolutional baseline, but still suffers from a decrease in performance. For valence, there is a 30% drop in the Concordance Correlation Coefficient (CCC), from 0.64 to 0.49, between in-domain and cross-domain with IEMOCAP. The drop in performance is less pronounced for arousal and dominance.

To measure how much emotion is preserved by the anonymization systems, we compute the cosine distance between the emotion embeddings of the original and anonymized utterances, as well as the average differences in their scores for the three dimensions.

3.2.3. Naturalness

The concept of naturalness is oftentimes used in text-to-speech synthesis evaluation, where it involves generating speech that closely resembles human speech in terms of intonation, prosody, and pronunciation. Here, we aim at measuring natural-sounding anonymization systems in terms of intelligibility, clarity, and overall quality of the anonymized speech. For our evaluation, we use NISQA-TTS [10], a CNN-LSTM network that had previously been used for speech quality estimation. The model is first pre-trained on a speech quality database, and then fine-tuned on 16 naturalness datasets, including 10 editions of the Blizzard challenge [25], where teams build TTS systems which are evaluated by humans. The resulting model was tested on several datasets, including some that were not part of the training set. The average per-system correlation on the test set is of 0.77. Notably, NISQA-TTS outperformed MOSNet [26] on the VCC 2016 dataset.

NISQA-TTS outputs mean opinion scores (MOS), which range from 1 to 5. We compute the MOS of each anonymized utterance and then average across datasets and population segments.

3.2.4. Computational performance

Finally, we also measure how fast the anonymization models are. Besides the model’s efficiency, which is generally interesting for cost estimation, assessing the model’s speed may help determine the system’s feasibility for real-time applications, resource-constrained devices, or large-scale deployment scenar-
ios. We measure the inference speed both on CPU and GPU for different input sizes, as well as the throughput on GPU, which is relevant for assessing how fast these models can be trained, and at what cost.

### 3.3. Datasets

We have selected four different datasets, aiming to maximize the diversity in speakers, recording environment and other traits of the datasets. We have filtered samples that are shorter than 2 seconds or longer than 30 seconds for the anonymization system to run with a larger batch size, which decreases the run time. The numbers mentioned in this section refer to the filtered versions of the datasets. We gather the most important facts of these datasets in Table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gender</th>
<th>Speakers</th>
<th>Samples</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibriSpeech</td>
<td>Female</td>
<td>20</td>
<td>1383</td>
<td>2.64</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>20</td>
<td>1228</td>
<td>2.68</td>
</tr>
<tr>
<td>EdAcc</td>
<td>Female</td>
<td>29</td>
<td>2075</td>
<td>4.76</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>31</td>
<td>2625</td>
<td>6.08</td>
</tr>
<tr>
<td>Common Voice</td>
<td>Female</td>
<td>16</td>
<td>47</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>87</td>
<td>258</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>2</td>
<td>6</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Undefined</td>
<td>549</td>
<td>1642</td>
<td>2.82</td>
</tr>
<tr>
<td>RAVDESS</td>
<td>Female</td>
<td>12</td>
<td>720</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>12</td>
<td>720</td>
<td>0.73</td>
</tr>
</tbody>
</table>

#### 3.3.1. LibriSpeech

The LibriSpeech dataset [11] is a popular speech corpus derived from audiobooks in the LibriVox project. It contains approximately 1,000 hours of clean and transcribed English speech data. We use the test-clean subset, which comprises 20 male and 20 female speakers and 5 hours of audio divided into 2,611 samples, each with a duration of 7.3 seconds on average.

This dataset has the best recording conditions from those in our evaluation set, showcasing the behavior of the anonymization models in the absence of noise. It also does not contain speakers with different accents; its comparison with accented datasets highlights how the models can cope with accents and other speaker traits.

#### 3.3.2. Edinburgh Accented dataset (EdAcc)

The Edinburgh International Accents of English Corpus [12] is a new dataset from the University of Edinburgh comprising dyadic conversations between 60 pairs of speakers with different backgrounds and accents. The test set, which we use here, comprises 60 speakers, 31 male and 29 female, and 10.8 hours of speech. On average, speakers are 36 years old.

EdAcc is the only dataset from our evaluation set that contains spontaneous speech, which differs from read speech in terms of prosody and emotion. This fact, together with the diverse speakers, comprising both native and foreign speakers, closely resemble the real world. The major difference in this regard is that speakers are predominantly highly educated: 49 of the 60 speakers have a university degree, and they speak almost 4 languages on average. Another advantage of this dataset is that the linguistic background of the participants has been thoroughly documented. This information helps determine their English proficiency, which defines how salient their accent is.

#### 3.3.3. Common Voice

Common Voice [13] is a crowd sourced dataset where the participants record the audio on their phone or computer. We have filtered the test set, keeping only the samples from speakers which comprise at least three utterances, as we require several utterances per speaker for the privacy evaluation. The resulting dataset comprises 654 speakers, by far the most from our evaluation datasets, and 3.34 hours of speech divided into 1953 samples. On average, samples have a duration of 6.1 seconds. 87% of the speakers have not reported their gender. From the remaining 105 speakers, 87 of them are male, making this dataset uneven in this regard.

Common Voice is the most challenging dataset, as it includes more noise and has a lower recording quality. It is also the most noisy in terms of labeling, as other participants validate the transcripts; the other three datasets have fewer errors in the transcripts, as these were made by experts (EdAcc) or the dataset arises from read speech by professionals (LibriSpeech and RAVDESS). Still, the large number of speakers in this dataset makes it very valuable. It is also the only dataset with older speakers, with 45 speakers over 40 years old.

#### 3.3.4. RAVDESS

RAVDESS [14] is an emotional dataset. We use the speech subset, where 24 professional actors are asked to read a sentence with a certain emotion and intensity. 12 actors are male and 12 are female, and all speak with a neutral North American accent. The dataset comprises two different sentences, two levels of intensity (normal and strong), and 7 emotions: calm, happy, sad, angry, fearful, surprise, and disgust. It contains 1,440 samples, 60 per actor.

This dataset is useful to measure the emotion preservation of the anonymization models. All actors read the same sentences aloud, enabling us to focus on the elicited emotions.

### 4. StarGANv2-VC evaluation

We demonstrate our evaluation framework with the StarGANv2-VC. We experiment with several ASV configurations and summarize the results of the utility evaluation, highlighting the most pronounced differences among population segments for each component.

We run our experiments on an NVIDIA RTX6000 GPU, which has 48 GB of memory. Each experiment takes between 2 and 4 hours, depending on the evaluation components chosen.

#### 4.1. StarGANv2-VC

The StarGANv2-VC [15] is an unsupervised voice conversion (VC) model where the input speech is adapted to conform with a given style vector, which we extract from the mapping network with the target ID. The generator receives the style vector of the target, as well as the spectrogram and the F0 contour. It converts the input spectrogram to the given style vector, preserving the F0 contour, which is then synthesized with the Parallel WaveGAN [27]. The generator is trained adversarially with two models: one which discriminates real samples from fake ones, and another which recognizes the converted sample’s speaker. Six additional losses are included to ensure that the style vector is used appropriately, and that
the converted speech is consistent with the input speech. The model was trained on 20 speakers of VCTK [28], a corpus of read speech recorded in a professional studio.

In our experiments, each source speaker is assigned a target speaker randomly. The random number generator is seeded differently between the inference and evaluation runs, so that the target speakers of the enrollment set in the lazy-informed ASV scenario are not the same as those picked during inference. Given that the choice of the targets is crucial to the performance of the anonymization, we repeat each experiment three times, each time with different inference and evaluation seeds.

4.2. ASV configuration

We experiment with 4 different configurations for the ASV system, to determine which one is the most effective. They differ in the kind of speaker embeddings. The x-vector speaker recognition model [5] is a TDNN neural network trained to discriminate between speakers. Its architecture has been enhanced to create more robust speaker embeddings, resulting in the ECAPA-TDNN model [21]. We use these two models, both provided by SpeechBrain [7]. They were trained on VoxCeleb 1 & 2, and their reported EERs for the VoxCeleb 1 test set are 3.2 for the x-vector model and 0.8 for the ECAPA model. The speaker embeddings may be reduced with the LDA algorithm prior to being fed to the PLDA algorithm. Here are the 4 configurations:

1. **ECAPA**: ECAPA-TDNN speaker embeddings are used, without LDA dimensionality reduction.
2. **ECAPA + LDA**: ECAPA-TDNN embeddings are computed and reduced to 100 dimensions with LDA.
3. **x-vectors + LDA**: the x-vector embeddings are computed and reduced to 200 dimensions with LDA. This is the configuration of the Voice Privacy Challenge (VPC).
4. **x-vectors + ECAPA + LDA**: x-vector and ECAPA embeddings are concatenated and reduced to 200 dimensions with LDA.

We experiment with ECAPA and LDA, as this was done by the VPC to reduce the x-vector embeddings. ECAPA embeddings are already smaller than the reduced x-vector embeddings, but maybe LDA is a useful feature extractor. We therefore try both variants for the ECAPA embeddings. We also try concatenating both embeddings, as previous work shows that they encode different speaker traits [29], meaning that they could complement each other.

We use the development set of EdAcc and 100 hours of LibriSpeech’s training set to train the ASV systems; both datasets are described in Table 2. The genders displayed in the table are only to showcase the dataset’s distribution; they are not considered during training. The speaker recognition models are not fine-tuned; the training data is used to train the PLDA and (optionally) the LDA algorithms.

Table 2: Training data for the ASV system

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Gender</th>
<th>Speakers</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibriSpeech</td>
<td>Female</td>
<td>125</td>
<td>50.38</td>
</tr>
<tr>
<td>train-clean-100</td>
<td>Male</td>
<td>126</td>
<td>50.20</td>
</tr>
<tr>
<td>EdAcc-dev</td>
<td>Female</td>
<td>33</td>
<td>5.49</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>28</td>
<td>5.49</td>
</tr>
<tr>
<td>Demiboy</td>
<td></td>
<td>1</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 3 shows the Equal Error Rates (EERs) of all 4 ASV configurations, for both attack scenarios. A lower EER means that the ASV system more accurately identifies the source speakers. The averages shown in the bottom row are weighted across datasets. All configurations yield similar results for the ignorant scenario, whereas the configurations that use x-vectors perform better in the lazy-informed scenario. The performance of the configurations that only use ECAPA embeddings barely changes between the scenarios. Concatenating x-vector and ECAPA embeddings does not yield better accuracy than using only x-vector embeddings.

Previous work has already shown that the features extracted by the x-vector and ECAPA models contain different information about the speakers [29]. Apparently, x-vector features are more useful to detect anonymized speech than ECAPA features. X-vectors may perform better because of their larger size: they comprise 512 dimensions, against the 192 of the ECAPA vectors, giving x-vectors more room to encode features that are not modified by the anonymization process. Another possible cause of this performance disparity is that ECAPA models are more likely to overfit the training data, given the model’s larger size: they comprise 22.3 million parameters, whereas the x-vector model solely has 4.2 million parameters. Fine-tuning the ECAPA model on the anonymized data should improve its performance; we will try this in the future.

Regarding the individual datasets, the EERs for RAVDESS are usually the largest, meaning that the speakers in this dataset are hard to identify when anonymized. The Common Voice dataset represents the lower bound for EER. The two remaining datasets, LibriSpeech and EdAcc, were partly used to train the ASV system, so we expect their EERs to be smaller than those for Common Voice. Common Voice comprises by far the most speakers, which should result in a higher EER, as there are more enrollment speakers that can be confused with each trial speaker. However, they seem to be easy to tell apart, maybe because of Common Voice’s diversity in terms of recording environment and speaker attributes.

4.3. Utility

We measure the utility of speech anonymized with the StarGANv2-VC with all four utility components. The results are shown in Table 4, together with the privacy results of the best configuration from the previous experiment (x-vectors reduced with LDA to 200 dimensions). The results for original speech, which serve as a baseline, are also included in the table. The SER evaluation does not include original results, as the results shown in the table are the cosine similarities between the emotion embeddings of the original and anonymized samples. Similarly, it does not make sense to compute a baseline for the lazy-informed ASV scenario, as it involves anonymizing data.

The baseline results show that LibriSpeech speakers are easily identified, possibly because LibriSpeech data is used to train the ASV system, the number of speakers is low and there are many enrollment samples per speaker. The baseline ASR results are lowest for RAVDESS, presumably because the two sentences uttered by the actors consist of common words that are easy to identify. LibriSpeech’s baseline WER is also very low, given its clean recordings, whereas EdAcc and Common Voice are the most challenging datasets in this regard. The MOS scores for the original speech rank the datasets in the same way as the WERs.

According to the ASR evaluation, LibriSpeech samples can be anonymized with little loss in intelligibility. In contrast, the other datasets yield anonymized samples that are consid-
Table 3: EERs for the 4 ASV configurations, which range from 1 to 100, lower being better. The left number corresponds to the EER for the ignorant attack scenario, and the right one to the lazy-informed scenario. The lowest EER for each dataset across the different scenarios is marked in bold. The averages are unweighted.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ECAPA</th>
<th>ECAPA+LDA</th>
<th>Xvect+LDA</th>
<th>Xvect+ECAPA+LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibriSpeech</td>
<td>32.18 / 36.52</td>
<td>34.34 / 36.67</td>
<td>32.46 / 26.77</td>
<td>38.8 / 26.18</td>
</tr>
<tr>
<td>EdAcc</td>
<td>39.91 / 41.43</td>
<td>39.8 / 41.17</td>
<td>37.89 / 28.4</td>
<td>39.99 / 30.15</td>
</tr>
<tr>
<td>Common Voice</td>
<td>34.18 / 33.26</td>
<td>36.49 / 33.4</td>
<td>27.22 / 21.68</td>
<td>33.51 / 21.97</td>
</tr>
<tr>
<td>RAVDESS</td>
<td>43.42 / 43.84</td>
<td>46.8 / 42.75</td>
<td>43.66 / 35.81</td>
<td>46.62 / 34.48</td>
</tr>
<tr>
<td>Average</td>
<td>37.42 / 38.76</td>
<td>39.36 / 38.5</td>
<td>35.31 / 28.17</td>
<td>39.73 / 28.19</td>
</tr>
</tbody>
</table>

Table 4: Privacy and utility results for the StarGANv2-VC. The ASV results are those for the VPC configuration (x-vectors + LDA). The baseline results are computed with the original samples, skipping the anonymization step.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ASV ignorant (EER)</td>
<td>0.96</td>
<td>32.46</td>
<td>4.69</td>
<td>37.89</td>
<td>1.83</td>
<td>27.22</td>
<td>4.35</td>
<td>43.66</td>
</tr>
<tr>
<td>ASV lazy-informed (EER)</td>
<td>-</td>
<td>26.77</td>
<td>-</td>
<td>28.4</td>
<td>-</td>
<td>21.68</td>
<td>-</td>
<td>35.81</td>
</tr>
<tr>
<td>ASR-small (WER)</td>
<td>0.06</td>
<td>0.14</td>
<td>0.32</td>
<td>0.75</td>
<td>0.18</td>
<td>0.71</td>
<td>0.01</td>
<td>0.57</td>
</tr>
<tr>
<td>ASR-large (WER)</td>
<td>0.07</td>
<td>0.09</td>
<td>0.29</td>
<td>0.6</td>
<td>0.13</td>
<td>0.51</td>
<td>0.0</td>
<td>0.35</td>
</tr>
<tr>
<td>SER (cos. similarity)</td>
<td>-</td>
<td>1.0</td>
<td>-</td>
<td>0.99</td>
<td>-</td>
<td>0.99</td>
<td>-</td>
<td>0.98</td>
</tr>
<tr>
<td>Naturalness (MOS)</td>
<td>3.88</td>
<td>3.12</td>
<td>2.63</td>
<td>2.95</td>
<td>2.96</td>
<td>3.41</td>
<td>3.37</td>
<td>3.32</td>
</tr>
</tbody>
</table>

The MOS scores of the anonymized samples, which measure naturalness, disagree with the SER and ASR results, as they state that the most natural speech comes from Common Voice. Even RAVDESS yields more natural-sounding speech than LibriSpeech, which was the most intelligible according to the ASR evaluation, and also achieves the best MOS score in the baseline. These results differ greatly to those of LibriSpeech, whose recording conditions are similar to those of LibriSpeech. The difference between the two ASR model sizes is largest for Common Voice and RAVDESS, where the WER increases by 0.4. The large ASR model understands the anonymized speech better than the smaller one; its WER increases by 0.26 on average across datasets, whereas the small ASR model’s WER increases by 0.56 when the speech is anonymized.

All differences between original and anonymized speech are small according to the SER evaluation, meaning that the StarGANv2-VC can successfully preserve emotion. The difference is largest for the RAVDESS, which is the most challenging in this regard, as it contains the most emotion. The anonymized samples of LibriSpeech yield practically the same emotion embeddings as the original ones; similarly to ASR, this may be due to the dataset’s similarity to the training data.

The MOS scores of the anonymized samples, which measure naturalness, disagree with the SER and ASR results, as they state that the most natural speech comes from Common Voice. Even RAVDESS yields more natural-sounding speech than LibriSpeech, which was the most intelligible according to the ASR evaluation, and also achieves the best MOS score in the baseline. These results differ greatly to those of the baseline, where LibriSpeech was given the largest MOS scores. The synthesized speech NISQA-TTS was trained with probably differs from that of the EdAcc, Common Voice and RAVDESS, where the speech is barely intelligible according to the ASR results. Therefore, the MOS scores for these datasets may be unreliable.

4.4. Performance

The StarGANv2-VC’s inference times are depicted in Table 5. On GPU, inference takes 0.08 seconds for every 5 seconds of input. For an input duration of 1 minute, the model still computes the anonymized speech in less than 1 second, making this model suitable for real-time applications.

On CPU, the inference time varies more depending on the input duration. We run our experiments on 10 cores of an AMD EPYC 7443 24-Core Processor, where the inference time for 5 seconds of input is about 2 seconds. For inputs of 60 seconds, the StarGANv2-VC, needs 44 seconds before outputting a result.

The throughput on GPU is shown in Table 5. For an input duration of 10 seconds, the StarGANv2-VC can process over 6 batches of size 4 per second, meaning that it can process 1 hour of input in less than 10 minutes. The maximum batch size decreases for larger input durations, and is 1 for input durations of 30 seconds. As we used samples of up to 30 seconds in our experiments, we also used batch size 1.

Overall, the StarGANv2-VC is relatively small and efficient, with 50M parameters. It’s cheap to run and train, and its inference speed on GPU enables real-time applications. Even on CPU, the inference speed is reasonably low. For comparison, FastPitch [30] is roughly the same size and has been shown to be capable of handling real-time requirements. This model transforms text into mel-spectrograms and has already been used for anonymization [31].

Table 5: GPU throughput of the StarGANv2-VC for different input durations.

<table>
<thead>
<tr>
<th>Input dur.</th>
<th>Inference time (s)</th>
<th>GPU throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CPU</td>
<td>GPU</td>
</tr>
<tr>
<td>5</td>
<td>2.102</td>
<td>0.084</td>
</tr>
<tr>
<td>10</td>
<td>6.345</td>
<td>0.162</td>
</tr>
<tr>
<td>20</td>
<td>14.107</td>
<td>0.315</td>
</tr>
<tr>
<td>30</td>
<td>43.67</td>
<td>0.927</td>
</tr>
</tbody>
</table>
4.5. Population segments

As mentioned above, the datasets we have chosen provide a lot of information about the speakers, which we can use to evaluate the model’s performance for different segments of the population. In this section, we highlight some of the largest differences amongst groups. Tables with the utility and privacy results for all these groups can be found in the GitHub repository.

Regarding gender, the results differ depending on the dataset. In Common Voice, more privacy and utility is achieved for women, whereas in LibriSpeech it is the other way around. For EdAcc, the results are similar for both genders. In Common Voice, the privacy is similar across genders, but the utility is considerably higher for females. For example, the WER of the small ASR model is 0.52 for females and 0.75 for males. In RAVDESS, more utility is achieved for males, both in WER and MOS, and more privacy for females in the lazy-informed scenario. The differences regarding gender are negligible in EdAcc and LibriSpeech.

EdAcc also provides the ethnicity of its speakers (self-identified). Privacy is similar across ethnic groups, except for Asians in the lazy-informed scenario, where their EER is 0.05, whereas the other groups are around 0.07. The difference is more pronounced when looking at the achieved utility, where white speakers achieve lower WERs and higher MOS scores. This is expected, as the training data of the StarGANv2-VC is likely to be predominantly white. The anonymized samples have the worst naturalness scores for South Asian people, although the differences in this area are small: 2.84 for South Asian people is the lowest, whereas 3.02 for White people is the highest. The WER is highest for black people, with an 0.88 WER for the small ASR model; in contrast, white people have the lowest WER: 0.61.

With the RAVDESS samples, we can evaluate the utterances belonging to each emotion category separately, across speakers. Doing so shows that utility is highest for the “calm” and “neutral” categories. The “fearful” category has the worst utility overall, with the highest WER (1.14 for the small ASR model) and the second-lowest naturalness score (3.19). In contrast, the lowest WER belongs to the “calm” class (0.26 for the small ASR model), as well as the highest naturalness score (3.65). The emotion preservation is highest for the “calm” and “neutral” categories, reaching a cosine similarity of 0.99 with the original speech. For all other categories, the cosine similarity is 0.98.

5. Conclusion

We propose a new evaluation framework for speaker anonymization, where the utility assessment is extended with more components. These components help assess for which use cases are suitable the anonymization systems. We only train the PLDA algorithm of the ASV system on the anonymized data; the rest of the components were trained beforehand on original data. We pick robust components that can cope with the differences between original and anonymized speech, and deliver accurate results for various datasets. The datasets we have chosen are diverse and comprise a lot of information about the speakers, allowing us to compare the anonymization system’s performance for different population segments. Although the evaluation components would all benefit from fine-tuning, using pre-trained models makes the evaluation much faster, which is better for development purposes.

Our evaluation of the StarGANv2-VC shows that it does not achieve perfect anonymization with random target selection, but that the utility of the anonymized speech is high for clean speech. For speech recorded in worse conditions, the utility is lower, as the StarGANv2-VC’s training data didn’t include noise. Also, the anonymization works best for speech samples with less emotional content, and from white speakers. Making the StarGANv2-VC’s training data more diverse should alleviate the differences across population segments.

6. Acknowledgements

This research has been partly funded by the Federal Ministry of Education and Research of Germany in the project EMMONY (project number 01IS21060B) and partly funded by the Volkswagen Foundation in the project AnonymPrevent (AI-based Improvement of Anonymity for Remote Assessment, Treatment and Prevention against Child Sexual Abuse).

7. References


