

On the Effect of Dataset Size and Composition for Privacy Evaluation

Danai Georgiou, Carlos Franzreb^a and Tim Polzehl^b

German Research Center for Artificial Intelligence, Germany
danai.georgiou@campus.tu-berlin.de, {carlos.franzreb, tim.polzehl}@dfki.de

Keywords: Speaker Anonymization, Privacy, Automatic Speaker Verification, Voice Conversion.

Abstract: Speaker anonymization is the practice of concealing a speaker's identity and is commonly used for privacy protection in voice biometrics. As proposed by the Voice Privacy Challenge (VPC), Automatic Speaker Verification (ASV) currently represents the de facto standard for privacy evaluation; it includes extracting speaker embeddings from speech samples, which are compared with a trained PLDA back-end model. We implement this ASV system to systematically explore the influence of two factors on the ASV performance: a) the amount of speakers to be evaluated, and b) the amount of utterances per speaker to be compared. The experimentation encompasses the privacy evaluation of the StarGANv2-VC and the kNN-VC on the LibriSpeech dataset. The experimental results indicate that the validity and reliability of privacy scores inherently depend on the evaluation dataset. It is, furthermore, demonstrated that limiting the number of speakers and utterances per speaker can reduce the evaluation time by 99%, while maintaining the reliability of the scores at a comparative level.

1 INTRODUCTION

Speaker anonymization refers to the task of eliminating any distinguishable features from speech recordings that could be used to reveal the original speaker's identity (Tomashenko et al., 2022). This process usually involves transforming the voice of the original (*source*) speaker into that of another, commonly arbitrary, *target speaker* (Mohammadi and Kain, 2017). The goal of speaker anonymization is to protect the speaker's privacy to the greatest possible extent, while preserving all linguistic and para-linguistic content of the speech signal (Tomashenko et al., 2022).


Due to the inherent sensitivity of speech data, the continuous advances in voice biometrics have raised several public concerns regarding privacy, thereby highlighting the demand for effective privacy protection technologies. The recent legislation changes by the European data protection regulation (GDPR) lead to a drastically increased interest in privacy protection within the speech community, most significantly with the introduction of the Voice Privacy Challenge (VPC), which proactively aims at evaluating research and encouraging further advancements on privacy preservation solutions.


The VPC has spearheaded initiatives to design an evaluation framework for speaker anonymization, by

measuring effectiveness in concealing the speaker's voice identity, while maintaining intelligibility. It measures privacy by assuming an attack scenario, where the attacker has access to one or more 'public' anonymized *trial* utterances and several *enrollment* utterances of a speaker (Tomashenko et al., 2022). Evaluation is performed with an Automatic Speaker Verification (ASV) system that extracts speaker embeddings and then analyzes them through a Probabilistic Linear Discriminant Analysis (PLDA) algorithm (Ioffe, 2006).

Anonymization models are assessed by measuring the ASV system's ability to accurately match the anonymized samples to the correct speaker identity. By running numerous samples by multiple speakers through the ASV system, it is possible to quantify how often it fails to identify the speakers and thus how much privacy was gained. Therefore, privacy assessment may intrinsically depend on the number of speakers and speech samples used for evaluation, which highlights the importance of sample quantity in understanding privacy outcomes. However, the studies that comprehensively explore different compositions of the evaluation data are fairly limited. In light of the importance of voice privacy research, there arises a critical necessity of understanding the impact of the test dataset's size on the ASV system's performance.

We seek to find a reliable methodology for quanti-

^a  <https://orcid.org/0000-0002-1188-7861>

^b  <https://orcid.org/0000-0001-9592-0296>

fyng the privacy gained through speaker anonymization. On that account, the following research questions are derived:

- RQ1. *How does the size of speaker population in the evaluation dataset influence the performance of the ASV system?*
- RQ2. *How does the number of speech samples, i.e. utterances per speaker, influence the performance of the ASV system, particularly when varying the amount of trial and enrollment utterances?*

We expect a significant relationship between the ASV system’s performance and the number of speaker pairs and utterance pairs to be examined by the system. Therefore, we hypothesize:

- H1. The error rates of the ASV system will increase, as the number of speakers increases, due to the necessity of detecting a greater number of non-matching speaker pairs.
- H2. The error rates of the ASV system will decrease, as the number of utterances per speaker increases, as more speech data lead to less variance for each speaker.

We seek to establish a robust ASV configuration, providing results that generalize to other datasets and use cases. Here; we evaluate two powerful Voice Conversion (VC) models, the StarGANv2-VC (Li et al., 2021) and the kNN-VC (Baas et al., 2023), on the LibriSpeech dataset with the help of the evaluation framework proposed by Franzreb et al. (2023), which implements a similar ASV system from the VPC 2022. We experiment with different numbers of speakers and different speech samples per speaker, focusing on the split between trial and enrollment utterances, while taking into account the variance introduced by different target selections. Thereby, we assess the role these two factors play when measuring privacy. The results of the experimentation pose as a recommendation for conducting privacy evaluation in a way that ensures a balanced trade-off between the reliability of the ASV results and the computation time.

2 RELATED WORK

A few previous studies have made an effort towards investigating the impact of the speaker population and utterance amount on the privacy evaluation.

Sholokhov et al. (2020) study the factor that the number of speakers play from a voice spoofing perspective, where an attacker aims at spoofing the ASV

system by finding the closest (trial) impostor for a given target speaker (enrolled) from a large population. They show that an increasing number of speakers results in an increasing probability of false alarm, thereby demonstrating that a large speaker corpus will eventually result in a high speaker confusability. They further examine the effect of the number of utterances, showing that a low number of available utterances for impostor search results in lower false acceptance rates.

Building upon this, Srivastava et al. (2022) assess the quality of speaker anonymization by studying the effect of the speakers’ population size on the performance of speaker recognition. Their experiments range from a population of 20 to approximately 25,000 speakers. They show that the ability of both systems to correctly verify the true speaker decreases very rapidly as the number of enrolled speakers increases. Their experiments further show that an anonymization system with an optimal pseudo-speaker design strategy provides the same level of protection against re-identification among 50 speakers as non-anonymized speech among a vastly larger population of 20,500 speakers.

Exploring a distinct objective, Meyer et al. (2023) also investigate varying numbers of speakers and utterances per speaker in their experiments. However, their focus is to reduce the training time of the ASV system by restricting the size of the training dataset. Thus, they explore the effect of fine-tuning the already pre-trained ASV system with two different data reduction techniques of the dataset, i.e. (1) by limiting the number of utterances per speaker, and (2) by selecting all utterances from a limited number of speakers. We propose an extensive experimentation using the same variables to comprehensively assess the robustness of the ASV system specifically during the evaluation phase.

3 EXPERIMENTAL SETUP

The experimentation process is performed with the help of the open-source evaluation framework by Franzreb et al. (2023)¹. Here, we discuss the components used for privacy evaluation, including the ASV system deployed within the framework and attack models to be considered, as well as the selected VC-models and the evaluation datasets.

¹https://github.com/carlosfranzreb/spkanon_eval

3.1 Privacy Evaluation

To measure privacy, an attack scenario must be considered, where an attacker attempts to reveal a speaker’s identity by comparing anonymized samples publicly shared by the speaker with found samples spoken by the same individual. The relationship between privacy and the effectiveness of an attack with the ASV system is inverse analogue. Therefore, resilience to a strong attack by the ASV signifies a strong speaker anonymization, i.e. high privacy preservation.

3.1.1 The ASV System

The selected framework implements the ASV architecture proposed by the VPC 2022. The ASV consists of a speaker recognition model for extracting speaker embeddings from the speech samples and a PLDA-based backend algorithm. PLDA is a probabilistic extension of the Linear Discriminant Analysis (LDA); it is a common technique for reducing dimensionality by projecting data into a lower-dimensional subspace, where the between-class covariance is maximized and the within-class covariance is minimized (Ioffe, 2006). To generalize to unseen classes, the PLDA extends the LDA by assuming that the speaker embeddings follow Gaussian distributions, with which the between-speaker and within-speaker variability is modelled. The ASV computes the standard x-vector embeddings, which are then reduced to 200 dimensions with the LDA. The extracted speaker embeddings and their corresponding speaker labels are used to train the PLDA back-end, in which parameters are estimated with empirical Bayes ².

Two attack scenarios are considered according to the amount of knowledge of the attacker, as illustrated in Figure 1:

1. *Ignorant Attack*: The attacker is unaware of the anonymization step. The training data of ASV_{eval} and the enrollment data are not anonymized.
2. *Semi-Informed Attack*: The attacker is aware of the anonymization and has access to the anonymization model, but is oblivious to the particular parameters for target selection. Therefore, the attacker anonymizes enrollment data with the same anonymization model, but with different targets from the trial data. The ASV system, denoted ASV_{eval}^{anon} , is trained with an anonymized training set.

When considering the ignorant scenario, ASV_{eval} is trained on the original training set. For the stronger

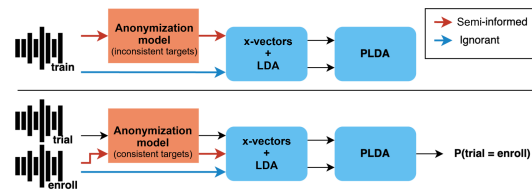


Figure 1: The x-vector-based ASV workflows for training (above) and evaluation (bottom) (Franzreb et al., 2023).

semi-informed scenario, the training data of ASV_{eval}^{anon} is anonymized with inconsistent targets for a better accuracy.

During evaluation, the dataset is split into two subsets: (a) trial set and (b) enrollment set. Each speaker must have at least one trial utterance, which is transformed with consistent target speakers by the VC-model during inference. For the semi-informed scenario, the enrollment set is anonymized with consistent targets, meaning that all the enrollment utterances of the same speaker are anonymized with the same pseudo-speaker. As the attacker is unaware of the target selection used for the trial utterances, a random target selection algorithm is employed for anonymizing the enrollment set. There is, thereby, a possibility that the trial and the enrollment utterances of a speaker are anonymized with the same target.

For each pair of trial and enrollment utterances, the log-likelihood ratio (LLR) of their speaker embeddings is computed with the trained PLDA model to determine how likely it is that the pair belongs to the same or a different speaker. The same-speaker vs. different-speaker decisions for each pair is made by using a threshold. Privacy is measured from the errors the ASV system makes with regard to the true speaker labels. Based on the validity of the same- and different-speaker decisions, two types of detection errors occur: false alarms and misses. The Equal Error Rate (EER) corresponds to the threshold for which the two detection errors are equally likely or “balanced”. A higher EER indicates that the ASV system, i.e. the attacker, fails at verifying the correct speaker identities.

3.2 Evaluated VC-Models

With the intention of drawing more general conclusions, experiments are performed for two fairly different VC-models, a non-parallel many-to-many GAN-based model and a self-supervised any-to-any model. Both VC-models are evaluated under comparable conditions with the same ASV system and on the same datasets.

²<https://github.com/RaviSoji/plda>

3.2.1 StarGANv2-VC

The StarGANv2-VC Li et al. (2021) is an unsupervised non-parallel many-to-many VC-model where an input speech is transformed to conform to a given style vector of a target. A generator receives as input the Mel-spectrogram and the extracted F0 contour of a source speaker, as well as the style vector of a target from the mapping network. The generator converts the Mel-spectrogram of the input sample to the given target style vector, while maintaining F0-consistency. The synthesis is performed with the Parallel WaveGAN (Yamamoto et al., 2020). The VC-model was originally trained on the VCTK corpus with 20 selected English speakers.

3.2.2 kNN-VC

The kNN-VC (Baas et al., 2023) uses a more straightforward approach with the k-nearest neighbors regression. Contrary to StarGANv2-VC, it performs an any-to-any conversion that transforms source speech samples into the voice of any target speaker, given that some reference utterances of the speaker are provided. The model utilizes WavLM-Large encoder (Chen et al., 2022) to extract feature sequences from both the source speaker’s input and the target speaker’s reference utterances. kNN-VC converts the input sequences by replacing them with the mean of the closest matching segments from the reference. The converted feature sequences are then transformed into audio waveforms using a HiFi-GAN-based vocoder (Kong et al., 2020). kNN-VC is not pretrained on target speakers, therefore the LibriSpeech *dev-clean* subset consisting of 40 speakers is selected for targets.

3.3 Datasets

The LibriSpeech is an openly available data set, which contains approximately 1,000 hours of clean read English speech sampled at 16 kHz. The corpus is derived from a large collection of public domain audiobooks by the LibriVox project.

The ASV system, i.e. the LDA and PLDA algorithms are trained on the LibriSpeech *train-clean-360* subset (see Table 1, either on the original corpus for the ignorant scenario (ASV_{eval}) or on the anonymized corpus for the semi-informed scenario (ASV_{eval}^{anon}). The ASV is trained once for each VC-model.

An overview of the selected evaluation datasets are presented in Table 2. In our experiments, the genders are not considered for evaluation; they are only

Table 1: Training data for the ASV system.

| Subset | Gender | Spkrs | Hours |
|-----------------|--------|-------|-------|
| train-clean-360 | Male | 482 | 363.6 |
| | Female | 439 | |

Table 2: Selected LibriSpeech Subsets.

| Subset | Gender | Speakers | Hours |
|-----------------|--------|----------|-------|
| test-clean | Male | 20 | 5.4 |
| | Female | 20 | |
| test-other | Male | 16 | 5.1 |
| | Female | 17 | |
| train-other-500 | Male | 602 | 496.7 |
| | Female | 564 | |

displayed to highlight the gender balance of the corpus in terms of the number of speakers. To avoid imbalances and optimize computation, samples that are shorter than 2 seconds or longer than 25 seconds are filtered from the datasets.

4 RESULTS

The experimentation is performed on subsets of the evaluation datasets by considering two data reduction strategies:

- (1) selecting different amounts of speakers to be evaluated, and
- (2) selecting different amounts of utterances per speaker to be evaluated.

With regard to the second experimentation technique, two sub-strategies are to be considered: (2a) selecting different amount of enrollment utterances with a fixed number of trials per speaker, and (2b) selecting different amount of trial utterances, while keeping the number of enrollments per speaker constant.

The relevant statistics of the evaluation datasets are illustrated in Table 3. The test-clean and test-other are combined into one dataset, i.e. the test set, to moderately increase their range of speakers and utterances. For a more challenging evaluation, the train-other-500 is also used, which allows for a larger-scale experimentation. On that account, the results of the experiments will provide insight on the role the speaker and sample size play in the ASV performance.

We run our experiments on an NVIDIA RTX6000 GPU, which has 48 GB of memory. For both VC-models, the datasets are anonymized five times with different seeding to avoid bias stemming from target selections. For each experiment, the

Table 3: Evaluation datasets. *Spk.* stands for speakers and *Utt.* for utterances.

| Dataset | Spk. | Utt. | Utt. / Spk. |
|-----------------|-------|---------|-------------|
| test set | 73 | 5,454 | 73 |
| train-other-500 | 1,166 | 148,182 | 127 |

speakers/utterances are randomly sampled five times. This process is repeated for the five different target selections, resulting in a total of 25 repetitions per experiment. The results consistently report the score averaged over all seeded target selections for each attack scenario.

4.1 Effect of Speaker Population Size

In this section, we explore the effect of the speaker population size on the performance of the ASV system. Following the standard evaluation method defined in the framework Franzreb et al. (2023), each speaker has one anonymized utterance in the trial set and the remaining utterances are included either original for ASV_{eval} or anonymized for ASV_{eval}^{anon} in the enrollment set. Within the following experiments of this section, all utterances from a specific subset of speakers are used for evaluation.

Table 4 lists the mean EERs for the two VC-models for various speaker subsets. The speaker selection is increased gradually and randomly sampled from all speakers of the subset. Looking at the results for the semi-informed scenario ASV_{eval}^{anon} , it can be observed that the more we increase the amount of speakers, the more the EER increases. Comparing the results for $\#spkrs = 10$ and $\#spkrs = 50$ of the test set, the EERs show an increase of 14% for StarGANv2-VC and of 16% for kNN-VC, whereas the standard deviation decreases. A similar increasing trend of the EER is also displayed in the results for the larger train-clean-500 subset, where the EER changes from 14.47% to 15.17% for StarGANv2-VC and from 4.15 % to 5.20 % for kNN-VC, when increasing the speakers from 50 to 1000.

4.2 Effect of Sample Amount per Speaker

Moreover, we measure the impact of the sample amount per speaker on the performance of the ASV system, by evaluating with different splits between trial and enrollment utterances. To eliminate bias, both the enrollment utterances and the trial utterance are randomly selected for each speaker, ensuring that no utterance is included in both sets.

4.2.1 Impact of Enrollment Utterances

In this section, we present the results for investigating the effect of different sample amounts in the enrollment set. On that account, an increasing amount of enrollment utterances per speaker is chosen for each experiment, while the number of trial utterances is consistently limited to one per speaker. Figure 2a illustrates the results for both VC-models on the test set, where the three speakers with less than 31 speech samples are excluded. It can be observed from the results that the EER follows a decreasing trend for ASV_{eval}^{anon} , as the number of enrollment utterances per speaker increases. The EER curve shows a plateau after 20 enrollment utterances. A comparable behavior of the EER is noted for the train-other-500, for which speakers with less than 120 utterances are excluded from evaluation, resulting in a dataset of 796 out of 1166 speakers. As shown in Table 5, the ASV_{eval}^{anon} performance is nearly equivalent for $\#enroll \text{ per spkr} = 20$ and $\#enroll \text{ per spkr} = 100$; it shows only a 2.5% decrease for StarGANv2-VC and a 5.25% decrease for kNN-VC, whereas σ is relatively low.

4.2.2 Impact of Trial Utterances

Figure 2b shows the influence of the trial utterance amount on the EERs. Here, the VC performance is assessed for an increasing number of trial utterances per speaker, while the number of enrollment utterances remains consistent. Based on the previous results, 20 enrollment utterances are used for each speaker for every evaluation, to limit the computation time.

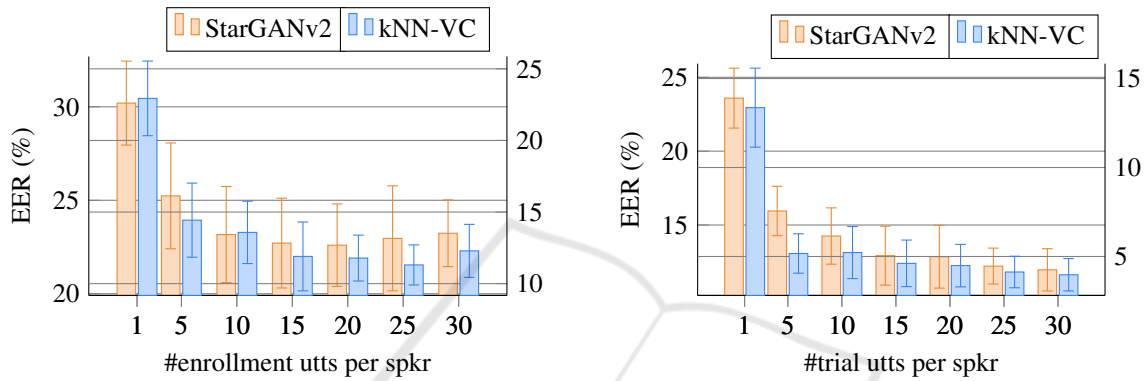
Comparing the standard of $\#trial \text{ per spkr} = 1$ and $\#trial \text{ per spkr} = 5$, the EER of ASV_{eval}^{anon} decreases by 32% for StarGANv2-VC, from 23.59% to 15.94%. Similarly, there is an EER decrease of 61% for kNN-VC from 13.36% to 5.16%, when 5 utterances are included in the trial set for each speaker. Moreover, the EER approaches a 49% reduction for StarGANv2-VC and 70% for kNN-VC with 30 trial utterances per speaker. An analogous EER trend decrease is observed, when experimenting with the train-other-500. Looking at the StaGANv2-VC results in Table 6, the effect of increasing the trial utterances achieves an EER decrease of 51% for ($\#trial \text{ per spkr} = 50$), with a very low σ of 0.51. For kNN-VC, the EER shows 86% decrease for 50 trial utterances, with a variability close to zero.

4.3 Computation Time

During evaluation, each utterance of the trial set is compared with each utterance on the enrollment set, therefore the number of evaluated utterance pairs in-

Table 4: EERs (%) for different amounts of speakers. The table shows the mean EER and standard deviation σ for ASV_{eval} and ASV_{eval}^{anon} . The results compare the behavior of EER with an increasing amount of speakers.

| Subset | #spkr | StarGANv2-VC | | | | kNN-VC | | | |
|-----------------|-------|--------------|----------|---------------------|----------|--------------|----------|---------------------|----------|
| | | ASV_{eval} | σ | ASV_{eval}^{anon} | σ | ASV_{eval} | σ | ASV_{eval}^{anon} | σ |
| test set | 10 | 35.33 | 7.67 | 19.91 | 7.66 | 43.36 | 7.60 | 9.51 | 4.5 |
| | 30 | 35.18 | 3.43 | 22.55 | 3.45 | 43.82 | 4.22 | 10.97 | 3.13 |
| | 50 | 35.98 | 2.68 | 22.71 | 2.20 | 44.02 | 2.25 | 11.02 | 2.49 |
| train-clean-500 | 50 | 28.05 | 3.21 | 14.47 | 1.86 | 40.07 | 3.57 | 4.15 | 1.24 |
| | 200 | 26.59 | 1.76 | 15.52 | 1.45 | 40.97 | 1.81 | 5.22 | 0.68 |
| | 500 | 27.18 | 0.59 | 15.04 | 0.76 | 40.49 | 1.04 | 5.14 | 0.38 |
| | 1000 | 26.98 | 0.60 | 15.17 | 0.47 | 40.96 | 0.67 | 5.20 | 0.19 |



(a) EERs for increasing enrollment utterances per speaker.

(b) EERs for increasing trial utterances per speaker.

 Figure 2: Comparison of the effect of increasing a) enrollment utterances and b) trial utterances per speaker on ASV_{eval}^{anon} , for both StarGANv2-VC (left y-axis) and kNN-VC (right x-axis) on the test set. For increasing enrollment utterances, each speaker has one utterance in the trial set. On the other hand, each speaker is assigned 20 enrollment utterances, for increasing trial utterances.

Table 5: EERs (%) for different amounts of enrollments per speaker from train-other-500.

| VC | #enroll | ASV_{eval} | σ | ASV_{eval}^{anon} | σ |
|-----------|---------|--------------|----------|---------------------|----------|
| StarGANv2 | 20 | 27.71 | 0.59 | 15.20 | 0.72 |
| | 100 | 27.57 | 0.74 | 14.82 | 0.61 |
| kNN-VC | 20 | 40.35 | 0.69 | 4.57 | 0.37 |
| | 100 | 40.29 | 0.59 | 4.33 | 0.39 |

Table 6: EERs (%) for different amounts of trials per speaker from train-other-500.

| VC | #trial | ASV_{eval} | σ | ASV_{eval}^{anon} | σ |
|-----------|--------|--------------|----------|---------------------|----------|
| StarGANv2 | 1 | 27.72 | 0.76 | 15.05 | 0.64 |
| | 50 | 25.71 | 0.70 | 7.39 | 0.51 |
| kNN-VC | 1 | 40.25 | 0.49 | 4.69 | 0.36 |
| | 50 | 39.88 | 0.38 | 0.64 | 0.09 |

creases prominently, as demonstrated in Figure 3. The average time cost for ASV_{eval}^{anon} is 0.5 hours on average for 796 speakers, with #enroll per spkr=20 and #trial per spkr=5. Increasing the trial utterances to 50 per speaker results in a time cost of approximately 18 hours for ASV evaluation.

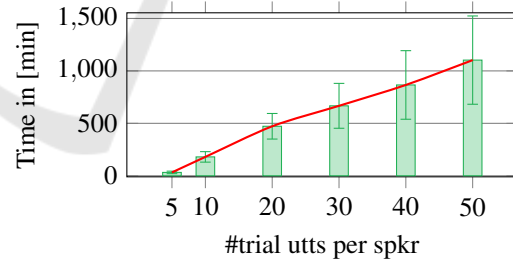


Figure 3: Time Cost for increasing trial utterances for both StarGANv2-VC and kNN-VC. 796 speakers are filtered from train-other-500 and 20 enrollment utterances are selected for each speaker.

We repeat the experimentation based on the above results to investigate how the EER and the time cost change. Both the test set and train-other-500 are limited to 50 speakers, where only 20 enrollment utterances and 5 trial utterances are selected per speaker. The proposed ASV_{eval}^{anon} configuration results are compared to the results of evaluating on the whole datasets, as presented in Table 7. The proposed data reduction decreases the computation time to one

Table 7: Comparison of ASV_{eval}^{anon} EERs obtained by evaluating the whole dataset (*all*) vs the *proposed* configuration of 50 speakers with 5 trial and 20 enrollment utterances per speaker.

| Subset | #spkrs | StarGANv2-VC | | kNN-VC | |
|-----------------|----------|---------------------|----------|---------------------|----------|
| | | ASV_{eval}^{anon} | σ | ASV_{eval}^{anon} | σ |
| test set | all | 22.57 | 1.85 | 10.47 | 1.57 |
| | proposed | 14.50 | 1.78 | 4.34 | 1.62 |
| train-clean-500 | all | 15.10 | 0.42 | 5.18 | 0.18 |
| | proposed | 9.18 | 2.02 | 1.24 | 0.93 |

minute on average. Compared to 5.5 hours needed for the whole train-other-500, we decrease the time cost by 99%.

5 DISCUSSION

In the above section, the results of different evaluation strategies were presented with regard to their impact on the ASV performance:

The first objective we explored was the impact of the speaker population size on the EERs. We hypothesized that the EER of the ASV system will increase, as the number of source speakers increases. This was expected and confirmed by the experimentation, as a larger number of speakers will result in more non-matching speaker pairs, i.e. a larger amount of enrollment speakers that may be confused with each trial speaker. Moreover, the above results demonstrate that a very small speaker subset introduces a high variability and unreliability, i.e. high standard deviation. This is especially problematic, as the EER might suggest a better privacy protection than what is actually provided by the evaluated VC-model. Although evaluating on a very large number of speakers seems to output more reliable results, the time cost should also be considered.

With regard to the sample quantity, our results confirm our second research hypothesis: an increase in the amount of samples per speaker, i.e. trial or enrollment utterances, resulted in an EER decrease. On that account, it must be acknowledged that the computation time is inherently dependent on the amount of trial-enrollment-comparisons. As each trial utterance must be compared with every utterance in the enrollment set, the time cost increases significantly with numerous trial utterances. Although evaluating the full dataset with several trial utterances per speaker leads to the strongest attack by the ASV, it is important to consider a balanced trade-off between EER and computation time.

6 LIMITATIONS

For privacy evaluation, the LibriSpeech dataset was selected for evaluation, as it contains a wide range of speakers in noise-free ambient conditions. It should be considered that kNN-VC was originally trained on LibriSpeech. Therefore, it was expected that the kNN-vc might output overall lower EER in contrast to StarGANv2-VC, which was trained on an external dataset. However, the interest of this research centers on the trend changes of the EER with regard to different data reduction strategies, rather than the results per sé. It was, therefore, essential that both VC-models were investigated under comparable conditions for generalization purposes.

The emphasis of sample amount experimentation shifted to the number of utterances per speaker. For evaluation, samples between 2 and 25 seconds were included from the selected datasets. This diversity of duration between the samples might contribute to the variability of the EER. Shorter utterances are known to degrade the ASV performance, as they are unable to capture all the variations of a speaker’s voice (Park et al., 2017). Further experimentation is needed to provide more insight into the impact of utterance duration on the reliability of ASV evaluation.

It should also be considered that we leverage a pre-trained speaker embedding model to circumvent the need for training additional models during the evaluation process, thereby minimizing the computation time. Future work could address this by training the speaker embedding model tailored to our data, potentially improving performance and the overall accuracy of the results.

Since we performed these experiments, the VPC has released a new framework. Regarding the privacy evaluation, this new edition removes the consistency constraint when selecting the target speakers for the enrollment utterances. In previous editions, all the utterances of each enrollment speaker were anonymized with the same target speaker. The effect of this change on the behavior of the ASV system is unknown, and we leave it for future work.

7 CONCLUSION

We examined the influence of varying number of speakers and speech samples in privacy evaluation with ASV. The experimental results demonstrated that a very small subset of speakers or utterances per speaker might produce unreliable EER that create a misleading impression of high privacy protection by the evaluated VC-model. We further showed that we could decrease the computation time needed for evaluation by 99% by reducing the number of speakers and samples per speaker, while still upholding the reliability of the results. We anticipate that our research could offer insights for conducting privacy evaluation in a way that ensures the validity of the results and their applicability to the greatest possible number of scenarios. Experimentation on further VC-models and with more challenging datasets would provide additional contributions to generalizability.

ACKNOWLEDGEMENTS

This research has been partly funded by the Federal Ministry of Education and Research of Germany in the project Medinym and partly funded by the Volkswagen Foundation in the project AnonymPrevent.

REFERENCES

- Baas, M., van Niekerk, B., and Kamper, H. (2023). Voice conversion with just nearest neighbors. *arXiv preprint arXiv:2305.18975*.
- Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., and Wei, F. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Franzreb, C., Polzehl, T., and Möller, S. (2023). A comprehensive evaluation framework for speaker anonymization systems. In *3rd Symposium on Security and Privacy in Speech Communication*, pages 65–72, ISCA. ISCA.
- Ioffe, S. (2006). Probabilistic linear discriminant analysis. In *Computer Vision—ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006, Proceedings, Part IV 9*, pages 531–542. Springer.
- Kong, J., Kim, J., and Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis.
- Li, Y. A., Zare, A., and Mesgarani, N. (2021). Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion.
- Meyer, S., Miao, X., and Vu, N. T. (2023). Voicepat: An efficient open-source evaluation toolkit for voice privacy research.
- Mohammadi, S. H. and Kain, A. (2017). An overview of voice conversion systems. *Speech Communication*, 88:65–82.
- Park, S. J., Yeung, G., Kreiman, J., Keating, P. A., and Alwan, A. (2017). Using voice quality features to improve short-utterance, text-independent speaker verification systems. In *Interspeech*, pages 1522–1526.
- Sholokhov, A., Kinnunen, T., Vestman, V., and Lee, K. A. (2020). Voice biometrics security: Extrapolating false alarm rate via hierarchical bayesian modeling of speaker verification scores. *Computer Speech & Language*, 60:101024.
- Srivastava, B. M. L., Maouche, M., Sahidullah, M., Vincent, E., Bellet, A., Tommasi, M., Tomashenko, N., Wang, X., and Yamagishi, J. (2022). Privacy and utility of x-vector based speaker anonymization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2383–2395.
- Tomashenko, N., Wang, X., Miao, X., Nourtel, H., Champion, P., Todisco, M., Vincent, E., Evans, N., Yamagishi, J., and Bonastre, J.-F. (2022). The voiceprivacy 2022 challenge evaluation plan.
- Yamamoto, R., Song, E., and Kim, J.-M. (2020). Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram.