Generating Identities with Mixture Models for Speaker Anonymization

Henry Turner*, Giulio Lovisotto, Ivan Martinovic

Department of Computer Science, University of Oxford, United Kingdom

Abstract

Speaker anonymization methods are a growing research area, due to the common use of voice interfaces coupled with growing privacy requirements. However, existing systems suffer from several issues, in particular a reduction in the entropy space of the newly created voices. This is problematic as it reduces the diversity of the produced anonymous voices, thus making distinguishing between anonymized voices more difficult, and limiting the number of anonymous voices that can be generated.

In this work we propose a method for creating the new identity component for anonymized voices, termed an x-vector, which aims to better reflect the natural diversity of voices, in turn increasing the diversity of the voices of anonymized speakers. We combine this identity generation method with existing anonymization schemes, to produce an overall anonymization system, which we evaluate. Our results demonstrate that our scheme creates more diverse anonymized voices than the existing baseline method.

Furthermore, our results show that the assumption of perfect de-coupling between identity and non-identity voice components used in existing speaker anonymization frameworks does not hold, highlighting a clear avenue for future work.

Keywords: Speaker anonymization, Speech, Speaker, Anonymization, Privacy

1. Introduction

As the use of voice as an interface proliferates, combined with people generating more vocal content which is often widely shared online, it is becoming

*Author pre-print.

DOI: 10.1016/j.csl.2021.101318

*Corresponding author

Email addresses: henry.turner@cs.ox.ac.uk (Henry Turner), giulio.lovisotto@cs.ox.ac.uk (Giulio Lovisotto), ivan.martinovic@cs.ox.ac.uk (Ivan Martinovic)

increasingly clear that solutions for anonymisation of individual voices are necessary. Voice cloning techniques have been rapidly advancing, with systems now able to generate realistic synthetic voices (Shen et al., 2018; Liu et al., 2018). At the same time further works demonstrated that few voice samples are required to bypass voice authentication systems and clone users voices (Arik et al., 2018; Turner et al., 2019).

Voice data, whether captured live or leaked from remote servers, not only constitutes personally identifiable information but can also contain user-sensitive information. Together with the introduction of new regulations, such as the General Data Protection Regulation in Europe, it has become increasingly important to develop techniques and methods to protect the privacy of voice data from adversaries.

Speaker anonymization techniques have been proposed to fulfill these protection requirements. These techniques process audio, so that the user-identifiable components of speech (i.e., those that link speech to user identity) have been removed, whilst retaining speech content and its other characteristics, such as tone and delivery. The VoicePrivacy Challenge 2020 (Tomashenko et al., 2020b) (VPC) was one of the early efforts to provide a common ground for the speaker anonymization problem. The challenge established datasets and metrics to evaluate speaker anonymization methods. In this work we develop a system within the parameters of the VPC, by focusing on improving the anonymous identity generation mechanism of the x-vector baseline challenge system. The baseline system works on the principle of decomposing the audio into the identity component, x-vectors (Snyder et al., 2018), and non-identifying components. The x-vectors can then be replaced with a pseudo x-vector, generated according to some strategy, and the audio re-synthesized. This paper is an extension to our system description paper (Turner et al., 2020) which was initially submitted to the 2020 VPC.

In this work, we highlight how the VPC baseline anonymous identity generation methods leads to pseudo x-vectors which are largely similar to each other. We find this to be a consequence of averaging large subset of original x-vectors in the baseline technique. To overcome this problem, we instead train Gaussian Mixture Models (GMMs) on a reduced version of the x-vector space; we then can sample from these GMMs to generate new pseudo x-vectors. We provide a general pipeline to help choose parameters to optimize the GMM performance. For this we make use of two relevant metrics which can be applied in x-vector space to estimate how well the GMM approximates the original x-vector distribution and its properties. We additionally show that the assumption of perfect separation between identity and speech content does not hold in the underlying pseudo-xvector anonymisation system. In fact, we find that anonymous voices are biased towards the original voice, highlighting a potential avenue for improvement.

Our main contributions are as follows:

1. We identify the shortcomings of the baseline anonymous x-vector generation method, demonstrating that it creates pseudo x-vectors which have

different similarity distributions to legitimate x-vectors.

2. We present a general method to generate new identities for anonymous voices and provide a framework for determining optimal parameters for this method.

3. We evaluate the overall anonymization system within the VPC framework, as well as further experiments, to validate its anonymization performance.

4. We investigate the differences between supplied x-vectors and resulting audio, highlighting a potential weakness in systems using the same pseudo-xvector anonymization architecture.

2. Related Work

2.1. Speaker Anonymization

Speaker anonymization has its origins in analog processing systems, when methods for securing and encrypting voices were first developed (such as Cox et al. (1987)). Recently, analog methods have become less relevant, as modern machine learning based voice systems take place on data in the digital domain.

Jin et al. (2009) presented a speaker anonymization system that uses voice conversion to transform speaker's voices to a new special speaker identity. The approach required parallel training data for each speaker, restricting its use somewhat.

A GMM based approach was proposed in Pobar and Ipšić (2014), which transformed user voices to a synthetic target voice using a combination of GMM mapping and harmonic plus stochastic models. This method achieved de-identification on 87.4% of samples, albeit with a limited database size of 10 speakers. The generated audio also lacked naturalness, due to the synthetic target speaker. Abou-Zleikha et al. (2015) improved on this by transforming the speaker to one of several voices from a pool of speakers, where the target speaker to be transformed to is selected to maximize de-identification performance.

Magarinos et al. (2017) improves on these works by using a cepstral frequency warping transformation based approach. A transformation function is applied in the spectral domain, de-identifying the voice. The inverse transform can later be applied to recover the original voice. Target speakers are selected to be the most dissimilar speaker to the original (based on PLDA distance between i-vectors).

Bahmaninezhad et al. (2018) took a similar approach, using a Convolutional Neural Network (CNN) to transform to a new voice, created from a set of transformation features from a source voice and a voice database.

Fang et al. (2019) advanced the area further, presenting an approach based on decomposing the audio into identifying (x-vector) and non-identifying components, before replacing the identity component and re-creating the audio. This work is based on this system, and as such further details are given in section 3.1. Srivastava et al. (2020) further explore this system, examining the anonymization impacts of different parameters on x-vector selection.

2.2. The VoicePrivacy Challenge 2020

The VoicePrivacy Challenge 2020 (Tomashenko et al., 2020b) provides the setting for this paper, and defines specific goals, selection of datasets, and set of metrics for the evaluation and comparison of voice anonymization systems. The challenge seeks solutions for a scenario where speakers want to hide their identity whilst still allowing all other downstream goals to be achieved (Tomashenko et al., 2020a). This is done by converting a speaker to a *pseudo-speaker* with a different voice: the new anonymous identity of the original speaker.

In order to accomplish downstream goals the following system requirements are given for the system: (a) to output a speech waveform, (b) to hide speaker identity as much as possible, (c) to distort other speech characteristics as little as possible, (d) to 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.

Throughout this work we use the evaluation dataset subsets specified by the VPC for evaluation of our models: (i) the Librispeech (Panayotov et al., 2015) clean development and test sets, and (ii) the VCTK (Yamagishi et al., 2019) development and test sets. We also evaluate our work using the objective metrics proposed by the VPC, using the models for this trained by the VPC, namely Equal Error Rate (EER), log-likelihood-ratio cost function C_{llr} and the discrimination loss component of this, C_{llr}^{min}, when analyzing the privacy performance of the system, and Word Error Rate (WER) for evaluating the speech recognition performance.

Several additional metrics were proposed as part of the VPC for evaluating speaker anonymization systems, the results of which we also evaluate in this work.

Specifically we also consider:

- Expected privacy disclosure at a population level, D_{ECE}, and worst case privacy disclose for an individual, $\log_{10}(l)$, from the Zero Evidence Biometrics Recognition Assessment (ZEBRA) framework (Nautsch et al., 2020).
- Linkability, specifically the global linkability $D_{\leftrightarrow}^{sys}$ (Maouche et al., 2020), which examines the (differences in) mated and non-mated score distributions.
- Global de-identification De_{ID}, and global voice distinctiveness G_{VD}, both of which are derived from voice similarity matrices (Noé et al., 2020).

The VPC also stipulated which datasets may be used in training anonymization systems in order to ensure that results are comparable. As such we make use of these same datasets for training our system: VoxCeleb 1 and 2 (Nagrani et al., 2017; Chung et al., 2018), LibriSpeech train-clean-100 and train-other-500 and LibriTTS train-clean-100 and train-other-500(Zen et al., 2019).

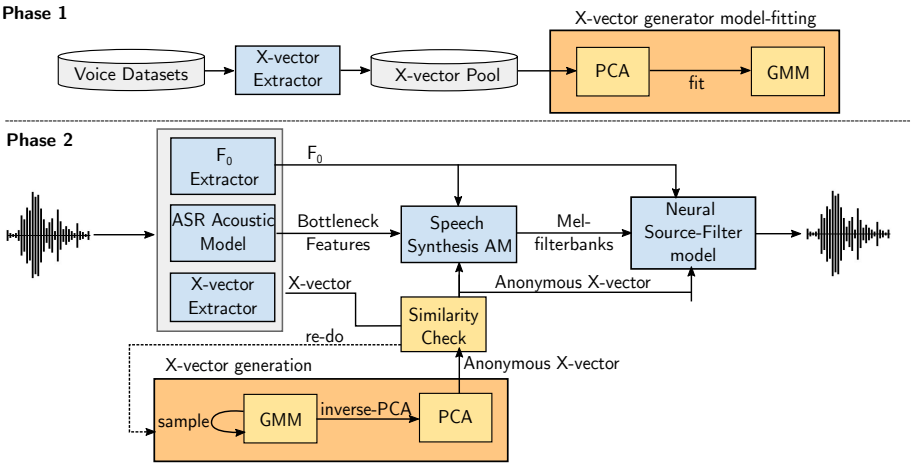

Figure 1: Voice Anonymization system diagram. We replace the sub system for generating pseudo X-vector's with a a combination of PCA and GMM (shown in orange in the diagram).

3. System Design

3.1. Overview

Our system design follows the same approach as the x-vector baseline system used in the VPC (Tomashenko et al., 2020a) and is inspired by Fang et al. (2019). This system takes the audio to be anonymized and derives three components from it, the x-vector, the bottleneck features (BN) , and the pitch information (F_0). The BN features are obtained by applying an Automatic Speech Recognition (ASR) acoustic model, which is a factorized time delay neural network (TDNN-F), with 40 MFCCs and a 100 dimension i-vector as input, and produce output BN features of dimension 256. This ASR model is trained using Librispeech train-clean-100 and train-other-500.

The x-vector extractor is a TDNN, using 30 MFCCs as input features, and outputting a 512 dimension speaker x-vector. It is trained using Voxceleb 1 and 2. The system assumes that these components decouple the speech content (BN and F_0) and the speaker identity x-vector.

Following this the x-vector is modified or replaced according to a generation technique; the modified x-vector is termed a pseudo or fake x-vector. The pseudo x-vector represents the new identity of the speaker, and is used for all utterances intended to be spoken by that identity. Subsequently a speech synthesis module uses the F_0, BN features and pseudo-x-vector to generate mel spectrograms. This speech synthesis model is an autoregressive network, outputs Mel-filterbanks of dimension 80, and is trained using LibriTTS train-clean-100.

A Neural Source-Filter (NSF) model then processes these filterbanks, along with the F_0 and the pseudo x-vector, generating the anonymized audio. This NSF model is trained with LibriTTS train-clean-100.

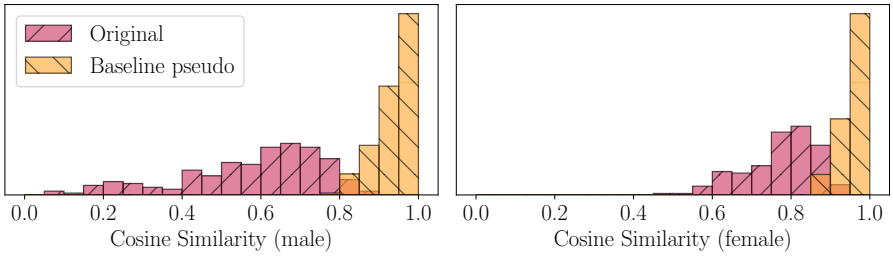

Figure 2: Distribution of cross-cosine similarities between pairs of x-vectors from original voices and from the baseline fakes. The baseline fake x-vectors do not follow the same distribution of cosine similarities as the original x-vectors: these fake x-vectors are much more similar to one another than x-vectors extracted from organic speakers.

A diagram showing the system can be seen in Figure 1. We use the models provided during the VPC, see Tomashenko et al. (2020a); Fang et al. (2019) for further details on the training of these models.

Fang et al. (2019) proposed three techniques for generating fake X-vectors: (i) nearest speakers, (ii) random selection and (iii) range selection. The VPC x-vector baseline system uses a variant of the last of these techniques, selecting the 200 furthest away x-vectors from the original speaker, and then averaging a random selection of 100 within these to produce the new fake x-vector. The LibriTTS (Zen et al., 2019) train-other-500 dataset is used in the baseline for this pool, with 600 users in the male pool and 560 in the female pool.

3.2. Rationale

We examine the pseudo x-vectors that are created by the baseline generation technique. Figure 2, shows the cosine similarities of the x-vectors extracted from the original voices, and pseudo x-vectors supplied to the later stages of the baseline system. We see that that the cross-similarity distribution between original voices and the pseudo x-vectors differs: pairs of pseudo x-vectors are more similar to one another than pairs of original voices.

This reduction in entropy leads to anonymized voices which are less distinct from one another, increasing the difficulty of distinguishing between anonymized voices. Furthermore, this also means the x-vector space is being underutilized, meaning the total number of distinct anonymous voice available will be reduced compared to organic voices.

We postulate that the reduction in entropy, and consequent reduced diversity of voices, occurs due to the averaging of several x-vectors in the pseudo x-vector generation process. Intuitively, similarly to what happens when sampling the mean of random samples of a normal distribution (which leads to a reduction in variance of a factor n, with n sample size), averaging subsets of 100 x-vectors will reduce the variance of the sampled x-vector means. In other words, the set of subset means will be more central in the complete x-vector space than the original population. While this can not be modeled exactly since the subsets

are not sampled at random (but are biased depending on the current user), its effects are clear. Likewise, the alternative selection methods proposed by Fang et al. (2019) and Srivastava et al. (2020) suffer from the same problem, due to their use of averaging x-vectors.

Figure 2 also reveals that the distribution of cosine similarities for females differs significantly from that of males, with female voices having increased similarity compared to males. This performance may be due to an imbalance in the number of voices for males and females in the dataset used to train the xvector extractor (2912 females vs 4451 males). Alternatively it could be due to female voices being higher pitch, and contain less spectral information, resulting in increased difficulty in discriminating between them than male voices.

3.3. Method

We improve the x-vector generation in two steps. At first, we learn the properties of the 512-dimensional x-vector space by using principal component analysis (PCA) on a large x-vector dataset. Secondly, we fit a generative model on the PCA-reduced space, in order to sample from it. By using a generative model we can retain as much of the diversity of the original space as possible, and avoid removing entropy with averaging operations. To generate a new pseudo x-vector using our method, we sample from the GMM in the PCA reduced space and then apply the PCA inverse transform.

As in the baseline, in the later stages of the anonymization a Speech Synthesis acoustic model is used to generate Mel-filterbank features, which are fed with the F0 and pseudo x-vector to a Neural source-filter model to generate audio. We train and reuse the models in the same way as the baseline, with the exception that we use the VoxCeleb1 (Nagrani et al., 2017) and VoxCeleb2 (Chung et al., 2018) datasets in our pool of speaker x-vectors, in addition to the LibriTTS (Zen et al., 2019) train-other-500 dataset. Figure 1 gives a full overview of how the system components fit together.

3.4. Determining Hyper-Parameters

Here we describe how we optimized the GMM parameters.

3.4.1. Preliminaries

In order to choose the best parameters for the system, we focus on analyzing the performance of two key metrics: (i) the cross-similarity distribution between the fake and original x-vectors and the (ii) resulting differential entropy of the trained GMM. High match between the cross-similarity distributions of original and anonymous x-vectors indicates that the generated anonymous x-vectors retain the similarity properties that are expected in the original set. Differently, the differential entropy of the Gaussian mixture model is an indicator of how much information is retained by the mixture (akin to its discrete counterpart being a measure of how many bits are necessary to encode the information). In this case, GMM with high entropy are preferable, as they more closely approximate the underlying mixture distribution.

We measure the cross-similarity distribution with the Kolmogorov–Smirnov (KS) statistic:

$$D_{KS} = \sup_x |F(x) - O(x)|, \tag{1}$$

where F and O are the cumulative distributions of the empirical cross-similarity distributions among a set of X-vectors:

$$f(x) = \{\cos \text{sim}(x_i, x_j), \} \quad \forall i, j \in x \ ,$$

O refers to the original X-vectors and F refers to fake (pseudo) X-vectors. We use the KS statistic for this as it is non parametric and can be easily applied to two empirical distributions.

The differential entropy of the GMM does not have a closed-form so we instead measure it by repeated Monte-Carlo sampling of the log-likelihood of the GMM-generated X-vectors:

$$\hat{H}(X) = \mathbb{E}_i[\log \sum_{k=1}^{K} \pi_k \frac{\exp(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu))}{\sqrt{(2\pi)^k |\Sigma|}}], \tag{2}$$

with x_i the generated samples, μ the component means, Σ the covariance matrix and π_k the weight of the k-th component. As this way of estimating the differential entropy (hereafter entropy) may be imprecise for insufficiently large samples, we also report the mixture entropy estimators introduced by Kolchinsky and Tracey (2017), which provide a tight estimation and are extremely fast to compute. See Appendix A for a more in-depth description of these estimators.

3.4.2. Setup

In the analysis, we vary the number of PCA and GMM components. We setup the evaluation as follows. Firstly we extract all the development and test X-vectors from VoxCeleb1, VoxCeleb2 (4,451 and 2,912 for male and female), and for each gender we perform a 50% train-test split, training the PCA+GMM models with only the training split. 2,000 samples are taken from the GMM and used to compute the entropy of Equation 2. We then apply the PCA inverse transform to obtain 512-dimensional pseudo x-vectors. To compute the KS statistic of Equation 1, we compute the cross-similarity among these pseudo x-vectors (obtaining F, Eq. 1) and we do the same among the testing part of the initial 50% split (obtaining O, Eq. 1). For the GMM we learn a diagonal covariance matrix on the PCA features, as these are de-correlated from one another, and set the maximum number of Expectation-Maximization iterations to 500 and the convergence tolerance to 10^{-15}, these are linearly increased if the EM does not converge. Additionally it is easier to compute the entropy using a diagonal co-variance matrix, and the reduced dimensionality of the matrix makes parameter estimation easier.

As the GMM fitting may vary, we repeat the train-test split two times, and for each split we also repeat the training twice and average the results.

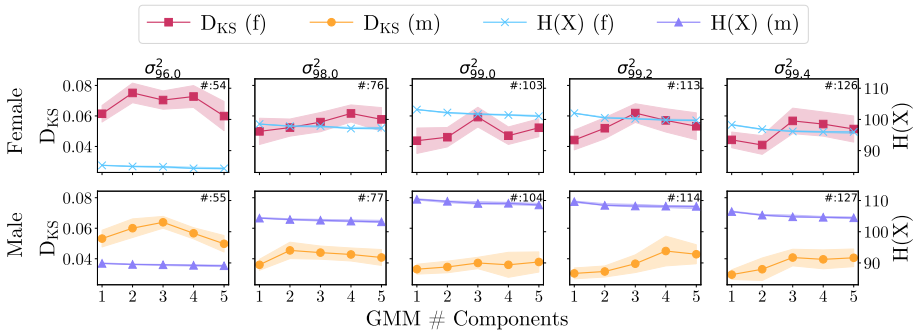

Figure 3: KS statistic (D_{KS}, Eq. 1) and entropy (H(X), Eq. 2) results across number of GMM components, amount of retained PCA variance (σ_p^2). Number of PCA components are given in the top right of each plot. Shaded areas for D_{KS} are 90% confidence intervals for D_{KS}. Shaded areas for H(X) are the lower and upper bounds measured with the estimators of Section Appendix A. The figure shows that increasing number of GMM components has a negligible effect on the two metrics.

3.4.3. Results

We report in Figure 3 and 4 the resulting entropy and KS statistic for varying the number of GMM components and for varying PCA retained variance. Figure 3 shows that the number of GMM components does not significantly affect the result (either the similarity or the entropy), showing some fluctuations but no statistically significant improvement for increasing number of components. Instead, the amount of retained variance more directly affects the result: increasing it to 99% bring significant improvement in the similarity and also corresponds to the configuration which gives the maximal entropy. Figure 4 confirms the same insights, also highlighting how retaining more than 99% of variance leads to a slight degradation in entropy and no visible benefit in the distribution similarity. This can be explained with the fact that increasing the retained variance has diminishing beneficial returns, as the number of PCA extracted features has to grow significantly as we retain more variance. For example, for male x-vectors, going from 96% to 98% increases the number of features by 22, while going from 99.2% to 99.4% (a ten-time smaller increment) increases the same number by 13. Figure 4 shows that retaining more of these features is only beneficial up to 99%, after which point the co-variance matrix needs to include increasingly small elements which bring a reduction to the overall entropy.

For the remaining experiments we choose to use one GMM component and 99% of variance retained by the PCA transformation, which corresponds to 103 components for females and 104 for males. Using these parameters, we report in Figure 5 a comparison of the cross-cosine similarity distributions across various x-vectors subsets, pseudo and original. The figure shows that our generated pseudo x-vectors more closely match the expected similarity distribution found among VoxCeleb x-vectors compared to the baseline-generated anonymous x-vectors.

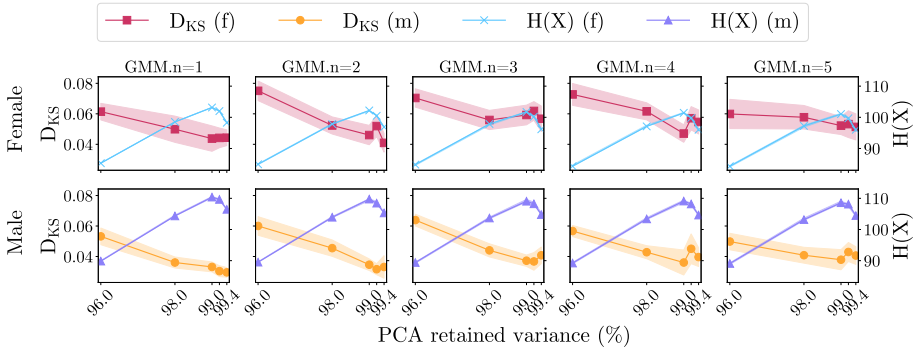

Figure 4: KS statistic (D_{KS}, Eq. 1) and entropy (H(X), Eq. 2) results across the amount of retained PCA variance (σ_p^2) and number of GMM components. Shaded areas for D_{KS} are 90% confidence intervals for D_{KS}. Shaded areas for H(X) are the lower and upper bounds measured with the estimators of Appendix A. The figure shows that the best entropy-distribution similarity trade-off is found at 99% retained variance.

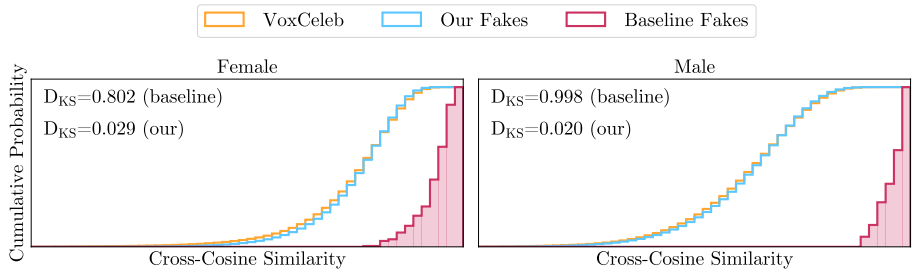

Figure 5: Example empirical cumulative distribution functions computed during the hyper-parameter search: the plots use one GMM component and 99% retained PCA variance. We do a 50% split on VoxCeleb data and use the first part to train our PCA+GMM and the second part to plot the VoxCeleb data series. We plot the baseline fake X-vectors computed with the VPC technique.

3.5. Forced Dissimilarity

As the GMM pseudo x-vectors are extracted randomly, these might occasionally be relatively close to the user's x-vector (this does not happen in the baseline as pseudo x-vectors are generated by selecting the n-furthest away x-vectors). This is detrimental to the quality of anonymization, so to avoid it we introduce a similarity check, termed *forced dissimilarity*, between the speaker's x-vector and the generated pseudo x-vector. For a pseudo x-vector X_p, we repeat the generation as long as X_p and the original x-vector X_o are too similar based on the θ_{FD} parameter:

$$\cos\,\mathrm{sim}(X_o, X_p) > \theta_{FD}.$$

We study the effects of this mitigation on the anonymization process in Section 4.3.

4. System Evaluation

In this section we evaluate the performance of our anonymization system, firstly by applying the VPC framework and extended metrics. Subsequently we also examine performance when using an alternate ASV system, and under our forced dissimilarity measure. Finally we examine the resulting x-vectors from our produced audio, to glean further insight into the systems performance.

4.1. Evaluation with Voice Privacy Challenge Framework

4.1.1. Setup

The VPC evaluation framework uses two datasets for evaluation, with two sub splits of each: (i) the Librispeech clean development and test sets, and (ii) the VCTK development and test sets. We focus on the diff split of the VCTK dataset, for brevity, and following the practice of most systems submitted for the VPC.

The VPC evaluates speaker verifiability, using metrics derived from speaker verification scores, and word error rate. Two scenarios are studied for speaker verifiability. Firstly original enrollment and anonymized trial (O-A), which examines the differences between the original voices and an anonymized version of them. In this case scores are computed between the clean audio, and the anonymized audio of the same speaker for a target trial, and the anonymized audio of a different speaker for a non-target trial.

Secondly anonymized enrollment and anonymized trial (A-A) is examined, where each of the enrollment and trial utterances are anonymized but to different identities. In this case a target trial is the anonymized enrollment and trial of the same original speaker, but anonymized to two different identities. Non-target trials are anonymized enrollments and anonymized trials from different speaker. In both cases the same identity is used for all utterances within that set (i.e., all the enrollment utterances are anonymized to voice A, all the trial utterances to voice B, where $A \neq B$. Speaker verifiability scores are calculated using an x-vector based system (Snyder et al., 2018) trained using the LibriSpeech train-clean-360 dataset, using a probabilistic linear discriminant analysis (PLDA) backend.

We focus on EER and C_{llr}^{min}in our analysis, as these were utilized in the VPC and have been demonstrated to be robust in Maouche et al. (2020). These are both computed from the set of scores between target and non-target utterances. Both the EER and C_{llr}^{min}minimize the discriminating power of the classifier against a dataset. Due to the anonymization applied to the audio, an EER of 50% and a $C_{llr}^{min}$of 1.00 are optimal, as target trial utterances are always anonymized compared to the enrollment utterance, and thus we hope to see that classifier unable to identify the two utterances as being spoken by the same (original) speaker.

Word error rates are calculated using a TDNN-F acoustic model with a trigram language model, based on the Kaldi recipe for Librispeech. Both evaluation systems are trained with the LibriSpeech train-clean-360 dataset. Further details can be found in the VPC evaluation plan (Tomashenko et al., 2020b).

Table 1: Speaker verifiability results for the pretrained ASV_{eval} model. Results for our anonymization method with 1 GMM component and σ^2_{99} PCA, without forced dissimilarity. In parenthesis we report the difference with the baseline system.

Dataset	Gender	Scenario	Development		Test	
			EER (%)	**C_{llr}^{min}**	**EER (%)**	**C_{llr}^{min}**
LibriSpeech	F	O-O	8.7	0.30	7.7	0.18
		O-A	46.9(-3.3)	0.97(-0.03)	40.3(-6.9)	0.95(-0.04)
		A-A	45.0(+8.2)	0.97(+0.07)	44.2(+12.0)	0.97(+0.13)
	M	O-O	1.2	0.03	1.1	0.04
		O-A	53.3(-4.5)	0.99(-0.01)	48.1(-4.0)	1.00(-0.00)
		A-A	46.7(+12.6)	0.97(+0.10)	43.4(+6.7)	0.98(+0.07)
VCTK (diff)	F	O-O	2.9	0.10	4.9	0.17
		O-A	46.0(-4.0)	0.96(-0.03)	44.6(-3.4)	0.98(-0.02)
		A-A	35.3(+9.1)	0.87(+0.11)	33.7(+2.0)	0.89(+0.04)
	M	O-O	1.4	0.05	2.1	0.07
		O-A	53.0(-0.9)	1.00(-0.00)	47.6(-6.2)	0.99(-0.01)
		A-A	36.0(+5.1)	0.92(+0.08)	40.2(+9.2)	0.93(+0.09)

Table 2: Details of number of trials for each of the datasets and their respective splits.

Dataset	**Split**	**Trials**	**Female**	**Male**	**Total**
Librispeech	Dev.	Target	704	644	1348
		Non-target	14566	12796	27362
	Test	Target	548	449	997
		Non-target	11196	9457	20653
VCTK (Diff.)	Dev.	Target	1781	2015	3796
		Non-target	13219	12985	26204
	Test	Target	1944	1742	3686
		Non-target	13056	13258	26314

4.1.2. Initial Framework

The full results for EER and C_{llr}^{min} are presented in Table 1, using the parameters determined previously in Section 3.4.3 (99% variance retained, one GMM component, one model per gender). The number of target and non-target trials for each of the datasets are given in Table 2.

For the O-A scenario, we experience a small performance degradation when compared to the baseline in most cases, with EER decreasing by up to 6.69% (female LibriSpeech test). The values of C_{llr}^{min} also feature a small drop, of a maximum of 0.04, however the values still remain generally close to 1, and as such the performance decrease is fairly limited. The degradation in results is more pronounced in females, with a C_{llr}^{min} decrease averaging 0.03, whereas for males the average decrease is 0.01. This could be because the baseline creates its pseudo x-vector by averaging x-vectors far away from the original, which in most cases will yield an x-vector that is also dissimilar from the original, acting as a dissimilarity constraint.

For the A-A scenario, we observe increases in the EER compared to the baseline for all settings, varying from an increase of 2% (Female VCTK Test) to 12.6% (Male LibriSpeech Dev). Similarly the C_{llr}^{min} improves across all data subsets, with increases from 0.04 to 0.13. We also note that the values of C_{llr}^{min} are

Table 3: WER rates for original and anonymized voices on the datasets specified in the VPC. Results are produced using 1 GMM component and 99% variance retained. In parenthesis we report the difference with the baseline system.

Dataset	Audio Type	Dev. WER (%)	Test WER (%)
LibriSpeech	Original	3.83	4.14
	Anonymized	10.02 (+3.63)	7.09 (+0.36)
VCTK	Original	10.79	12.81
	Anonymized	16.3 (+0.91)	16.98 (+1.75)

Table 4: Results for the ZEBRA and Linkability additional metrics evaluated in the Voice Privacy Challenge for our system using one GMM component and 99% of variance retained. Presented results are for the test split of both datasets. Difference from the x-vector baseline is given in brackets.

Dataset	Gen.	Scenario	ZEBRA		Linkability
			$\mathbf{D_{ECE}}$	$\mathbf{\log_{10}(l)}$	$D_{\leftrightarrow}^{\mathbf{sys}}$
LibriSpeech	F	O-O	0.58	3.98(C)	0.90
		O-A	0.03(+0.03)	0.82(+0.51)(A)	0.15(+0.08)
		A-A	0.02(-0.09)	0.72(-1.77)(A)	0.09(-0.20)
	M	O-O	0.69	3.92(C)	0.96
		O-A	0.00(+0.00)	0.16(-0.12)(A)	0.06(-0.02)
		A-A	0.02(-0.05)	0.50(-1.90)(A)	0.11(-0.09)
VCTK (diff)	F	O-O	0.59	3.65(C)	0.88
		O-A	0.01(+0.01)	0.74(+0.61)(A)	0.07(+0.02)
		A-A	0.08(-0.03)	1.43(-0.44)(B)	0.24(-0.04)
	M	O-O	0.67	3.92(C)	0.95
		O-A	0.01(+0.01)	1.17(+1.17)(B)	0.06(-0.00)
		A-A	0.05(-0.06)	1.86(+0.62)(B)	0.14(-0.16)

close to a perfect score of 1 in many cases, indicating very strong anonymization performance, and that two versions of the same voice anonymized are rarely confused with one another.

We observe a small increase in Word Error Rate (WER) across all of the datasets and dataset splits, when comparing the results from the large language model used in the VPC evaluation, as shown in Table 3. Overall the increases in WER are fairly small and the results slightly worse than those of the VPC x-vector baseline.

4.1.3. Additional Metrics

In this section we calculate the values of the additional metrics proposed during the VPC: (i) ZEBRA Framework (Nautsch et al., 2020) (ii) Linkability (Maouche et al., 2020) and (iii) Voice Similarity Matrices (Noé et al., 2020) Metrics. All of these metrics are computed using the existing scores output by the speaker verifiability model.

We focus on the results for the test split of both datasets.

ZEBRA framework. Table 4 shows the results for $\mathrm{D_{ECE}}$ and $\log_{10}(\mathrm{l})$ computed with the ZEBRA framework. The expected privacy disclosure, $\mathrm{D_{ECE}}$, gives a score for the average level of protection afforded to a population, with a score of 0 corresponding to *perfect privacy* and is thus optimal. The worst case

Dataset	Split	De_{ID}		G_{VD}	
		M	F	M	F
LibriSpeech	Dev	0.99 (-0.01)	0.97 (-0.03)	-3.69 (+5.07)	-1.96 (+7.21)
	Test	0.99 (-0.01)	0.96 (-0.02)	-2.80 (+6.18)	-2.74 (+7.33)
VCTK	Dev	1.00 (-0.00)	0.97 (-0.02)	-2.85 (+9.82)	-2.53 (+6.28)
	Test	1.00 (-0.00)	0.98 (-0.01)	-2.93 (+8.80)	-2.92 (+7.36)

Table 5: De-Identification and Voice Distinctiveness (Gain) performance derived from Voice Similarity Matrices. Results are calculated for our system using 1 GMM component and 99% variance captured by PCA. Difference from x-vector baseline shown in brackets

privacy disclosure, $\log_{10}(\mathrm{l})$, gives a score for the protection afforded to the worst individual, again with 0 being optimal. We observe that our system performs slightly worse than the baseline in the O-A scenario for both metrics, although still retains an A grade for the worst case in all but one case. For the A-A scenario we improve over the x-vector baseline, with values for $\mathrm{D_{ECE}}$ becoming lower across all data subsets. The worst-case privacy disclosure, $\log_{10}(\mathrm{l})$, also improves in the A-A scenario, with the exception of the male VCTK (diff) dataset. Furthermore the VCTK (diff) dataset letter grades are B, implying an adversary would be incorrect once in every 10 to 100 attempts (Nautsch et al., 2020). These results for $\log_{10}(\mathrm{l})$ suggest that our system does not perform equally well for all individuals, causing a high value for VCTK in particular due to poor anonymization of a specific (or several) individual(s).

Linkability. Table 4 shows the linkability, $D_{\leftrightarrow}^{\mathrm{sys}}$, results. The linkability measures the difference between the target and non-target score distributions, and can capture anonymization problems not detected by metrics such as $\mathrm{C_{llr}^{min}}$. For linkability the optimal score is 0. These results display a similar pattern to the other population wide metrics. For the O-A scenario we have variable results, with increases in $D_{\leftrightarrow}^{\mathrm{sys}}$ for both female datasets and no change and a decrease for the male datasets. For the A-A scenario we see large decreases in $D_{\leftrightarrow}^{\mathrm{sys}}$, showing a clear reduction in linkability. These numbers still remain higher than those for O-A, but the performance disparity between the two scenarios is reduced.

Voice Similarity Matrix Metrics. Table 5 shows the De_{ID} and G_{VD} results calculated from voice similarity matrices for our datasets. Deidentification, De_{ID}, measures the ease with which speaker can be linked between original and anonymised, with scores of 1 being optimal. The results for De_{ID} show a slight degradation in performance from the baseline, which had close to perfect scores of 1 (> 0.99) for all combinations of dataset and split. Voice distinctiveness, G_{VD}, measures how distinctive individual voices are, with greater than 0 dB indicating more distinctive voices and less than 0 being less distinctive voices in the anonymized space. For G_{VD} we see an improvement for all metrics, with the absolute values being better for female, but the magnitude of changes overall being slightly larger for males. This increase in voice distinctiveness compared to the baseline clearly demonstrates that our x-vector generation technique produces more diverse anonymous voices than the original technique.

4.2. Alternate ASV Evaluation

Our proposed system utilizes x-vectors for replacing the identity component of the audio. The VPC evaluation framework and associated metrics also use a state-of-the-art x-vector extractor, which could cause masking of potential issues, if the anonymization properties were not maintained into other spaces. As such we validate that the anonymization results hold when also computed with an i-vector (Dehak et al., 2011)-based speaker verification system.

Setup. We repeat the speaker verification experiments conducted in the VPC analysis, using the pre-trained i-vector model found in the Kaldi examples[1]. The features are 24 MFCCs with a frame length of 25ms, and an energy-based voice activity detection (VAD) system determines which frames contain speech. The UBM is a 2048 component full-covariance GMM. The i-vector model extracts a 400 dimensional i-vector, and is trained using the 100,000 longest utterances from the VoxCeleb 1 and 2 training datasets. A PLDA backend, which uses an LDA dimension of 200, is used for scoring the utterances, with the system achieving an EER of 5.3% on the VoxCeleb test datasets.

We extract i-vectors using this trained system for the original and anonymous audio, and for each of the datasets, that we evaluated across the scenarios in section 4.1. This allows us to compare the x-vector and i-vector results directly.

We focus on the EER and C_{llr}^{min}metrics and compare the results on the test splits of the LibriSpeech test and VCTK (diff) test datasets.

Results. Figure 6 shows the EER and C_{llr}^{min}results for the x-vector and i-vector systems on the test splits for all of the anonymization scenarios. In most cases performance is similar for the x-vector and i-vector systems, with only small differences observed between the two systems. In general the difference in performance is more pronounced on female voices than male voices. The largest degradations in performance come on the female split of the VCTK database, in both the O-A and A-A scenarios.

These large drops in EER and C_{llr}^{min}point to the potential for some overspecialization in the x-vector space for female performance. Alternatively, it could be related to dataset imbalances between males and females, the effects of which were observed in examining the x-vector extractor in Section 3.2 and in the VPC results in Section 4.1.2. Overall, these results suggest that the anonymization is not limited only to the x-vector space.

4.3. Forced Dissimilarity

As mentioned previously the forced dissimilarity (FD) θ_{FD} parameter repeats the GMM sampling of a pseudo x-vector if the x-vector is too near to the original user's voice. However the specific value of θ_{FD} used may impact the anonymization performance of the system, and as such we evaluate the system with varying θ_{FD}.

[1] https://kaldi-asr.org/models/m7

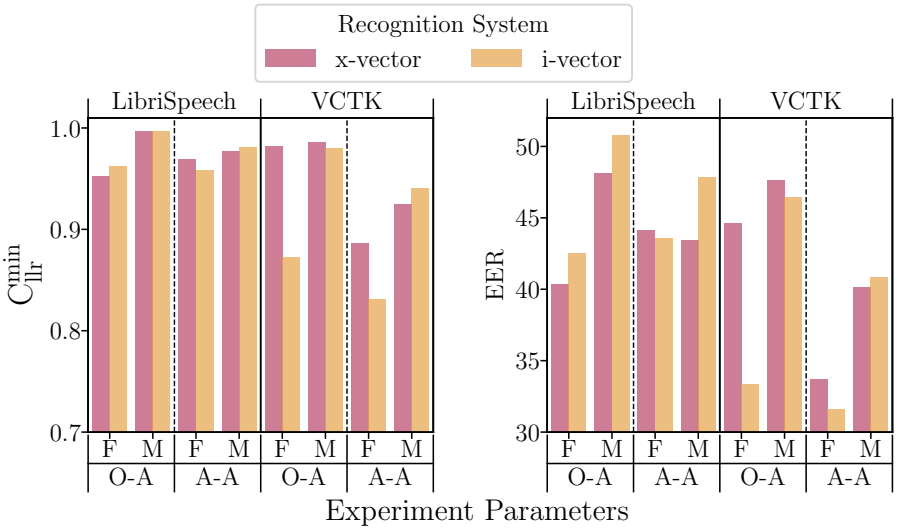

Figure 6: Alternate ASV System results for the test splits of both datasets on all scenarios, comparing the C_{llr}^{min} and EER when calculated using the VPC x-vector system and an alternate i-vector system. Note the y-axes does not start at 0.

Setup. We evaluate the effect of θ_{FD} using the VPC framework, again focusing on the results for C_{llr}^{min}. We run the evaluation framework with $\theta_{FD} \in [0.2, 1]$ (1 corresponds to no similarity constraint).

Results. Figure 7 presents the C_{llr}^{min} for both the O-A and A-A configurations, for the test splits of both datasets and for both genders. We observe that for the O-A scenario the values for C_{llr}^{min} remain high, and slightly increase with a smaller θ_{FD}. This result is as expected, as by forcing less similarity, and thus a bigger distance between the original voice (and thus the enrollment voice) and the target x-vector, the resultant voice also becomes increasingly distant from the original and thus the enrollment voice.

For the A-A scenario we observe a different effect, in that with smaller θ_{FD}, i.e., forcing more distance from the original voice, the C_{llr}^{min} decreases. This is likely because the available area of the x-vector hyperspace is restricted by θ_{FD}, meaning that when anonymizing the same voice twice, both the target x-vectors come from a region that becomes increasing small as θ_{FD} decreases, resulting in voices that are more similar. This is potentially problematic, and shows that small values of θ_{FD} should not be used.

The effects on the overall WER for the test split of all of the datasets are seen in Figure 8. We observe that for both datasets, as θ_{FD} decreases we experience an increase in WER. This may be due to links remaining between the non-speaker identity components of the anonymization and the original speaker identity, resulting in a conflict between features when transforming to a very different x-vector, and thus worse performance.

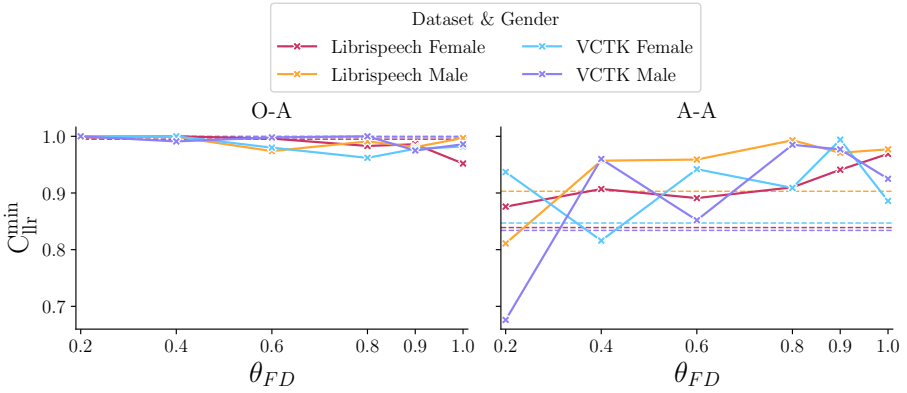

Figure 7: Plots showing C_{llr}^{min}values for the test split of the datasets for varied values of the forced dissimilarity parameter θ_{FD} for the model, for both scenarios in the VPC challenge. The values of the Baseline system are shown with the dashed lines.

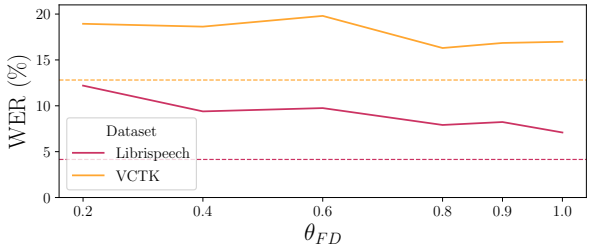

Figure 8: Plots showing WER as a function of the forced dissimilarity parameter θ_{FD}. WER values for the original datasets are marked with the dashed lines.

Overall our results suggest that values of θ_{FD} need to be chosen carefully, and small values should be avoided due to their negative impact on WER and the A-A scenario. If using FD we would recommend a value of 0.9, as this yields strong performance in O-A, A-A and WER, but gives the guarantee of avoiding a transformation to a very similar voice.

4.4. Examining Resultant X-Vectors

The overall architecture for the system relies on the assumption that the voice can be decoupled into the identifying components (described by the x-vector) and the speech content (described by the bottleneck features and F_0) (Fang et al., 2019). In order to examine this assumption, we evaluate the resultant x-vectors produced by the system (i.e., the x-vectors extracted from the final audio). If the assumption holds, we expect to see that the resultant x-vectors are as distant from the original voice as any other voice, and that they closely resemble the target x-vector supplied to the synthesis models.

Setup. We perform this analysis on the VCTK test dataset. Anonymized audio

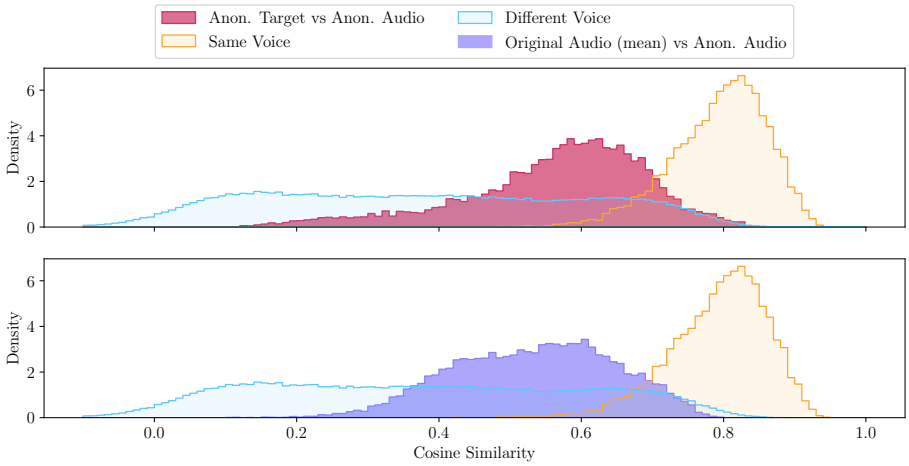

Figure 9: Similarity (cosine) comparison of resultant x-vectors with the target generated by the GMM. We observe that voices are more dissimilar than two copies of the same (original) voice, but not as dissimilar as if they were a different voice entirely.

is generated for all of the enrollment utterances for each user, with the same pseudo target being used for each of these utterances. Resultant x-vectors are then extracted from this anonymized audio, using the same x-vector extractor as used for generating the dataset to train the GMM. We analyze the distances between the pseudo outcome and the pseudo target using the cosine distance between them, as well as comparing this to other sets of distances between original voices and the newly anonymized voices. The x-vectors used for computing these distances are extracted with the extractor used by our system, and not the separate VPC evaluation x-vector extractor. We also compute two reference distributions, composed of the original voices compared with themselves, and of the original voices compared with different voices.

Results. Figure 9 shows two sets of histogram distributions of the cosine similarity scores, for the anonymized audio. In the upper plot of Figure 9 we examine the x-vectors of the final audio post synthesis, with the target x-vector supplied in synthesis. We see that the distribution of these does not mirror that of a normal voice compared with itself, and instead is shifted left-ward (i.e., less similar) and has a further spread distribution. This suggests that the synthesis algorithms do not result in audio that can be considered (in general) to be spoken by the same voice as the target vector, degrading anonymization performance in the process.

The bottom plot, comparing the original audio's mean x-vector with the anonymized version of that voice, shows a distribution that does not match either of our reference distributions. In an optimal system, the similarity distrbution should match the different voice distribution, however we observe that the distribution is shifted to the right and more condensed. This implies that

the anonymized voices are more similar to their original voice than would be ideal, and thus that the synthesis process must be retaining some bias toward the original voice.

Taken together, these results highlight deficiencies in the audio synthesis process, showing that it comes up short at recreating the target accurately and suggesting that the assumption that the non-x-vector features (BN and F_0) do not contain identifying information does not hold.

4.5. Subjective Naturalness

As part of the VPC the organizers computed subjective results for each of the submissions. As we have improved the parameters within our system since the VPC, we do not report the results provided by the challenge organizers here, but they can be found in Tomashenko (2020). Instead, we conduct Mean Opinion Score (MOS) tests to assess the subjective naturalness of the voices produced by our system.

Setup. We use the Amazon Mechanical Turk platform to conduct our MOS tests. We follow the ITU-T Recommendation P.808 (ITU-T, 2018) for conducting these listening tests, with each audio file being assessed on scale from 1 (poor) to 5 (excellent). As directed by the recommendation, we implement a qualification phase, in which we verify participants are native English speakers, are using headphones, are in a quiet environment and of normal hearing. Participant training is conducted with 5 samples that are selected to cover the range of sample qualities. Training samples are identical for all workers. Throughout the training and rating phases gold standard and trapping questions are used, as per the recommendation, to ensure participants are attentive to the task. Workers who fail to answer these correctly have their scores discarded.

We perform our MOS for 100 randomly selected anonymised audio files from the VCTK test dataset. We use the original audio, the baseline system, our proposed system, and our proposed system with $\theta_F D = 0.9$, giving a total of 400 audio files to be scored. We use the same random audio files for each set i.e. the original utterance, and 3 anonymised versions of it.

Workers are presented with sets of 12 samples at a time and paid for each set of 12 they complete. Each of these 12 samples contains at least one audio sample from each of the four conditions (Original + 3 anonymisation methods), to prevent bias introduced by workers who do not rate all samples. Following the p.808 recommendation we incentivize workers financially to rate at least half of the total samples. Each sample is rated by at least 6 distinct workers, with an average of 9.57 ratings per sample. We received ethical approval from our institution for this study, reference CS_C1A_21_010.

Results. The computed MOS can be seen in Table 6. We conduct statistical significance testing, after removing bias, pairwise between each of the audio sets using a Mann-Whitney U test, following the method given in Rosenberg and Ramabhadran (2017). We find that the differences between the MOS are statistically significant at the 1% level for all pairs of audio sets, except between

Type	Total Ratings	MOS	Std. Dev.
Original	1039	4.432146	0.648822
Baseline	926	2.816415	0.800534
Proposed System	918	2.501089	0.860815
Proposed System w/ $\theta_{FD} = 0.9$	945	2.503704	0.826108

Table 6: Mean Opinion Scores for original audio and varied audio creation system. The differences between pairs of audio sets are statistically significant ($p < 0.01$) for all pairs except the two versions of the proposed system.

the two sets based on the Proposed System. Full test statistics and p-values are reported in Appendix B

We observe that all of the audio anonymization systems have much lower opinion scores that the original audio. This highlights the need for improvement in the overall anonymization method.

We also observe that our proposed technique has a slightly worse MOS than the Baseline system, scoring 2.50 and 2.82 respectively. Whilst this difference is fairly small, it is statistically significant, and suggests that the baseline method produces x-vectors that are more natural sounding than our method. This difference could be because the Gaussian space captured by the GMM is not guaranteed to contain only natural sounding voices, and thus can contain x-vectors that produce poor audio. The larger standard deviation also suggests this could be the case.

5. Discussion

In this section, we discuss the overall anonymization capabilities of our system and possible concerns that need further investigation, as well as signposting directions for future work to investigate in order to improve this technique, as well as others based on NSF models.

5.1. Overall Anonymization Assessment

Our experiments focused on two key scenarios throughout, one where the original voice is compared to the anonymized voice, and one where two instances of the same voice anonymized are compared with one another. The x-vector selection technique we developed is intended to maximize performance on the second one of these scenarios, however the two are interlinked, and poor performance on one impacts the other, as well as the overall usefulness of the system.

Our experiments demonstrate that our x-vector selection technique improves the anonymization performance of the system, particularly in improving the diversity of the anonymized voices. We observe that values of C_{llr}^{min} approaches a perfect score in many scenarios, indicating strong anonymization, although worse performance was generally observed in the female voices analyzed. This could be caused by the x-vector extractor's weaker performance on female voices,

meaning the female voices occupy a smaller space. Potential causes of this lopsided x-vector extractor performance could be imbalanced datasets, or due to female voices containing less spectral information than male voices to begin with, particularly when a single extractor must handle all voices, covering a wider spectrum than a gender specific extractor. Further investigation will be required in order to better understand and mitigate this issue.

The results from the additional metrics developed for the voice privacy challenge also achieved strong results, although we observe some worse performance in similar scenarios as to the original metrics. In particular the results for the $\log_{10}(\mathrm{l})$ metric show that for some individuals performance may be poor. This could be an effect of the random pseudo x-vector selection from the GMM, or could be explained with our analysis of resultant x-vectors in section 4.4.

The resultant x-vectors analysis in Section 4.4 also showed that the outcome utterances are not close enough to the x-vector target to be considered the same voice as it. Improvement in this aspect of the system is also likely to improve anonymization results further. These results also show that separation between identity and speech content is not perfect within the system. Whilst this finding does not appear to impact the metrics assessed with the VPC, it does imply that an adversary attempting de-anonymization gains some information about the original voice from measuring its distance to other voices, and thus improvements in the process would result in further privacy gains. One potential avenue of exploration could be to reduce the feature size of the xvectors and bottleneck features, to reduce the information that could be present in both of them.

5.2. Speech Quality Results

The speech quality of the anonymised audio was assessed by both WER of the resulting audio, as well as MOS tests to rate its subjective quality.

The WER results for both the baseline and our proposed technique were worse than for the unanonymized audio. Across all four datasets our proposed system performed worse (0.36% to 3.63%) than the baseline. In terms of MOS, our proposed system scored 2.5, compared to 2.8 for the Baseline and 4.43 for the original audio.

These sets of results highlight two problems that need further attention. Firstly, the x-vector anonymization and re-synthesis system that underpins both the baseline and our proposed system needs further improvement, to increase the naturalness of audio that is generated using it. Secondly, the worse performance of our proposed system in terms of speech quality needs further investigation, to determine why the x-vectors produced by our method lead to worse audio. One potential cause could be that the averaging method used by the baseline can not have extreme values, as the averaging method constrains the xvectors to values more central in the space. Our method does not have such a constraint, and there is no guarantee that all the x-vectors that are modeled by the GMM-PCA space represent naturally sounding speakers, and thus poorly performing speakers could be sampled from this space.

Further work could investigate optimizing the x-vector sampling process to produce more natural speakers, as opposed to our focus of encouraging diversity in the produced x-vectors.

6. Conclusions

In this paper we presented our technique for speaker anonymization, by utilizing GMMs to generate pseudo x-vectors to transform voices to. Our work expands on the initial study of this technique presented as a system description for the Voice Privacy Challenge 2020, and further explores the optimal parameters for our system and conducts further experiments into the anonymization.

We demonstrate that our system performs particularly strongly when comparing two (differently) anonymized versions of the same voice, and in particular outperform the VPC baseline by a large margin. We also investigate the properties of the x-vectors taken from the produced audio. These experiments show that there remains space to further optimize the synthesis models, producing output audio that more closely resembles the target x-vectors, and thus differs more the original voice, improving anonymization and privacy.

Finally we discuss future avenues to be explored in developing this system further, in particular highlighting the need for improved word error rates and better naturalness and intelligibility, if this system (or others based on similar synthesis models) are to be able to provide a useful solution for voice anonymization.

7. Acknowledgments

This work was supported by a grant from Mastercard and by the UK Engineering and Physical Sciences Research Council (EPSRC) [grant numbers EP/N509711/1, EP/P00881X/1]. We would also like to thank the Voice Privacy Challenge organizers for arranging the challenge, as well as the reviewers of this manuscript for all of their helpful comments and feedback.

References

Abou-Zleikha, M., Tan, Z.H., Christensen, M.G., Jensen, S.H., 2015. A discriminative approach for speaker selection in speaker de-identification systems, in: 2015 23rd European Signal Processing Conference (EUSIPCO), pp. 2102–2106. doi:10.1109/EUSIPCO.2015.7362755.

Arik, S., Chen, J., Peng, K., Ping, W., Zhou, Y., 2018. Neural voice cloning with a few samples, in: Advances in Neural Information Processing Systems, pp. 10019–10029.

Bahmaninezhad, F., Zhang, C., Hansen, J., 2018. Convolutional Neural Network Based Speaker De-Identification 2016, 255–260. doi:10.21437/odyssey.2018-36.

Chung, J.S., Nagrani, A., Zisserman, A., 2018. Voxceleb2: Deep speaker recognition, in: Proc. Interspeech 2018, pp. 1086–1090. URL: http://dx.doi.org/10.21437/Interspeech.2018-1929, doi:10.21437/Interspeech.2018-1929.

Cox, R.V., Bock, D.E., Bauer, K.B., Johnston, J.D., Synder, J.H., 1987. Analog Voice Privacy System. AT&T Technical Journal 66, 119–131. doi:10.1002/j.1538-7305.1987.tb00480.x.

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing 19, 788–798. doi:10.1109/TASL.2010.2064307.

Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N., Bonastre, J.F., 2019. Speaker Anonymization Using X-vector and Neural Waveform Models , 3–8URL: http://arxiv.org/abs/1905.13561, arXiv:1905.13561.

ITU-T, 2018. Recommendation P.808: Subjective evaluation of speech quality with a crowdsourcing approach. URL: https://www.itu.int/rec/T-REC-P.808/en.

Jin, Q., Toth, A.R., Schultz, T., Black, A.W., 2009. Speaker de-identification via voice transformation. Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009 , 529–533doi:10.1109/ASRU.2009.5373356.

Kolchinsky, A., Tracey, B.D., 2017. Estimating mixture entropy with pairwise distances. Entropy 19, 361.

Liu, L.J., Ling, Z.H., Yuan-Jiang, Ming-Zhou, Dai, L.R., 2018. Wavenet vocoder with limited training data for voice conversion. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2018-Septe, 1983–1987. doi:10.21437/Interspeech.2018-1190.

Magarinos, C., Lopez-Otero, P., Docio-Fernandez, L., Rodriguez-Banga, E., Erro, D., Garcia-Mateo, C., 2017. Reversible speaker de-identification using pre-trained transformation functions. Computer Speech & Language 46, 36–52.

Maouche, M., Srivastava, B.M.L., Vauquier, N., Bellet, A., Tommasi, M., Vincent, E., 2020. A Comparative Study of Speech Anonymization Metrics, in: Proc. Interspeech 2020, pp. 1708–1712. URL: http://dx.doi.org/10.21437/Interspeech.2020-2248, doi:10.21437/Interspeech.2020-2248.

Nagrani, A., Chung, J.S., Zisserman, A., 2017. Voxceleb: A large-scale speaker identification dataset, in: Proc. Interspeech 2017, pp. 2616–2620. URL: http://dx.doi.org/10.21437/Interspeech.2017-950, doi:10.21437/Interspeech.2017-950.

Nautsch, A., Patino, J., Tomashenko, N., Yamagishi, J., Noé, P.G., Bonastre, J.F., Todisco, M., Evans, N., 2020. The privacy zebra: Zero evidence biometric recognition assessment.

Noé, P.G., Bonastre, J.F., Matrouf, D., Tomashenko, N., Nautsch, A., Evans, N., 2020. Speech pseudonymisation assessment using voice similarity matrices. arXiv preprint arXiv:2008.13144 .

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: an asr corpus based on public domain audio books, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE. pp. 5206–5210.

Pobar, M., Ipšić, I., 2014. Online speaker de-identification using voice transformation, in: 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1264–1267. doi:10.1109/MIPRO.2014.6859761.

Rosenberg, A., Ramabhadran, B., 2017. Bias and statistical significance in evaluating speech synthesis with mean opinion scores, in: Proc. Interspeech 2017, pp. 3976–3980. URL: http://dx.doi.org/10.21437/Interspeech.2017-479, doi:10.21437/Interspeech.2017-a479.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al., 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 4779–4783.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., 2018. X-VECTORS : ROBUST DNN EMBEDDINGS FOR SPEAKER RECOGNITION David Snyder , Daniel Garcia-Romero , Gregory Sell , Daniel Povey , Sanjeev Khudanpur Center for Language and Speech Processing & Human Language Technology Center of Excellence The Johns Hopkins Un. Icassp2018 , 5329–5333.

Srivastava, B.M.L., Tomashenko, N., Wang, X., Vincent, E., Yamagishi, J., Maouche, M., Bellet, A., Tommasi, M., 2020. Design choices for x-vector based speaker anonymization. arXiv preprint arXiv:2005.08601 .

Tomashenko, N., 2020. The voiceprivacy 2020 challenge - challenge setup and results. URL: https://www.voiceprivacychallenge.org/docs/1___VoicePrivacy_challenge_setup_and_results_N_Tomashenko.pdf. odyssey 2020.

Tomashenko, N., Srivastava, B.M.L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.F., Noé, P.G., Todisco, M., 2020a. Introducing the VoicePrivacy initiative .

Tomashenko, N., Srivastava, B.M.L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.F., Noé, P.G., Todisco, M., 2020b. The VoicePrivacy 2020 Challenge evaluation plan URL: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf.

Turner, H., Lovisotto, G., Martinovic, I., 2019. Attacking speaker recognition systems with phoneme morphing, in: European Symposium on Research in Computer Security, Springer. pp. 471–492.

Turner, H., Lovisotto, G., Martinovic, I., 2020. Speaker anonymization with distribution-preserving x-vector generation for the voiceprivacy challenge 2020. arXiv preprint arXiv:2010.13457 .

Yamagishi, J., Veaux, C., MacDonald, K., et al., 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92) .

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R.J., Jia, Y., Chen, Z., Wu, Y., 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech, in: Proc. Interspeech 2019, pp. 1526–1530. URL: http://dx.doi.org/10.21437/Interspeech.2019-2441, doi:10.21437/Interspeech.2019-2441.

Appendix A. GMM entropy estimators

As shown in Section 3.4.2, estimating the GMM entropy is useful in order to best choose the parameters in a system development phase. While in this paper we directly estimate entropy using the log-likelihood of generated samples (Equation 2), in presence of large datasets or in online applications it might be useful to have faster estimators of entropy. Here we report the GMM entropy estimators introduced in Kolchinsky and Tracey (2017), which we show generally provide very tight entropy bounds:

$$\hat{H}(X) = H(X|C) - \sum_i \pi_i \log \sum_j \pi_j \exp(-D(p_i||p_h)),$$

$$H(X|C) = \frac{1}{2} \sum_i \pi_i [\log |\Sigma_i| + d \log 2\pi + d],$$

for the lower bound we replace D with the Chernoff α-divergence distance function $C_\alpha(p_1||p_2)$:

$$C_\alpha(p_1||p_2) = \frac{(1-\alpha)\alpha}{2}(\mu_1 - \mu_2)^T((1-\alpha)\Sigma_1 + \alpha\Sigma_2)^{-1}(\mu_1 - \mu_2) + \frac{1}{2} \ln \Big(\frac{|(1-\alpha)\Sigma_1 + \alpha\Sigma_2)|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha} \Big),$$

System One	System Two	U Statistic	p-value
Original	Baseline	75224.5	< 0.0001
Original	Proposed System	50757.5	< 0.0001
Original	Proposed System w/ $\theta_{FD} = 0.9$	46916.5	< 0.0001
Baseline	Proposed System	315373.0	< 0.0001
Baseline	Proposed System w/ $\theta_{FD} = 0.9$	320107.5	< 0.0001
Proposed System	Proposed System w/ $\theta_{FD} = 0.9$	431770.0	0.432

Table B.7: Mann-Whitney U Test Statistics and p-values for MOS scores between each pair of systems evaluated.

for the upper bound, we replace D with the Kullback-Leibler divergence $\mathrm{KL}(p_1||p_2)$:

$$\mathrm{KL}(p_1||p_2) = \Big[\ln \frac{1}{2}|\Sigma_2| - \ln|\Sigma_1| + (\mu_1 - \mu_2)^T \Sigma_1^{-1}(\mu_1 - \mu_2) + \mathrm{tr}\Big(\Sigma_2^{-1}\Sigma_1\Big) - d\Big].$$

In all equations, d indicates the Gaussian's dimensionality, α is set to 0.5, Σ_i is the co-variance matrix of the i-th component, μ_i are the means of the i-th component.

Appendix B. MOS Statistical Significance Tests

We perform statistical significant testing following the method outline in Rosenberg and Ramabhadran (2017). We use normalized-rank normalization to first normalize the scores by participant and then by utterance, to correct for participant bias and utterance bias respectively.

We then use these scores to conduct a Mann-Whitney U test, between each pair of audio sets. In all cases the null hypothesis is that the distributions are equal to one another. Table B.7 shows the U statistic and p-value for each of the pairwise sets. We see that the differences are statistically significant for all pairs, except the Proposed System with and without Forced Distancing.