Speaker Anonymization with Distribution-Preserving X-Vector Generation for the VoicePrivacy Challenge 2020

Henry Turner, Giulio Lovisotto, Ivan Martinovic

University of Oxford, UK

firstname.lastname@cs.ox.ac.uk

Abstract

In this paper, we present a Distribution-Preserving Voice Anonymization technique, as our submission to the VoicePrivacy Challenge 2020. We notice that the challenge baseline system generates fake X-vectors which are very similar to each other, significantly more so than those extracted from organic speakers. This difference arises from averaging many X-vectors from a pool of speakers in the anonymization processs, causing a loss of information. We propose a new method to generate fake X-vectors which overcomes these limitations by preserving the distributional properties of X-vectors and their intra-similarity. We use population data to learn the properties of the X-vector space, before fitting a generative model which we use to sample fake X-vectors. We show how this approach generates X-vectors that more closely follow the expected intra-similarity distribution of organic speaker X-vectors. Our method can be easily integrated with others as the anonymization component of the system and removes the need to distribute a pool of speakers to use during the anonymization. Our approach leads to an increase in EER of up to 16.8% in males and 8.4% in females in scenarios where enrollment and trial utterances are anonymized versus the baseline solution, demonstrating the diversity of our generated voices.

Index Terms: voice anonymization, voice privacy, X-vector

1. Introduction

Recent advances in voice cloning have lead to extremely realistic synthetic voices [1, 2] and have shown how few voice samples are actually required to bypass voice authentication systems [3, 4]. As voice is personally identifiable, protecting voice data privacy from leaks and adversaries is necessary.

Speech anonymization is a novel technique which aims to anonymize voice data while retaining both the words spoken in the audio and the way they are spoken, such as tone and delivery. Ideally, anonymized voices need to be different from the voice of the original speaker, in a way that guarantees that an anonymized voice is not linkable to the original speaker. The VoicePrivacy Challenge 2020 [5], which this paper is part of, aims to drive forward the creation of voice anonymization systems. This is done by providing a common set of datasets, a baseline anonymization system and a framework for assessing the performance of voice anonymization.

In this paper we note that the X-vector anonymization proposed in the VoicePrivacy Challenge 2020 baseline system leads to fake X-vectors which underutilize their multi-dimensional vector space. Consequently, they tend to be very similar to each other, leading to anonymized voices that are similar as a result. We show how this abnormal similarity is evident by looking at the distribution of cross-similarity between pairs of X-vectors, comparing the distribution of fake and original similarities. We then propose a method that better leverages the vector space by learning the properties of this space from population data and fitting a generative model on a reduced-dimensionality space. The generative model is then used to sample fake X-vectors for anonymization. We show the performance of our method comparing it to the baseline, showing how our generation method significantly improves the diversity between anonymous voices while retaining the baseline performance across other metrics. Our main contributions are as follows:

1. We analyze the shortcomings of the baseline anonymous X-vector generation method, showing that generated X-vectors tend to be much more similar to each other than original X-vectors are.
2. We present a general method that improves on the anonymous X-vector generation by learning the distributional properties of the X-vector space and by fitting a generative model on this space, where X-vectors can be sampled from. Our method also removes the requirement of having a pool of speakers' X-vectors during the anonymization process (as in the baseline), which may lead to privacy leakages.
3. We evaluate our method within the VoicePrivacy Challenge 2020, showing its improved performance in creating differing anonymous voices.

2. Related Work

2.1. Speaker Anonymization

Speaker privacy is not a new concept, with works on securing and encrypting voices existing for several decades, dating back to the analog processing era [6]. This physical layer anonymization has its uses, but approaches that operate at this level either do not mask the voice itself, such as by adding a signal to existing audio [7], or render the audio unintelligble without a decryption key, preventing use for other legitimate purposes.

In this work we focus specifically on anonymization, in which personally identifiable attributes of the speech signal are supressed, but leaving intact all other aspects. Past work in this area includes using voice transformation to convert voices to a specific special speaker identity [8] or using a Convolutional Neural Network (CNN) to convert each speaker to a new anonmyous voice, created as a function of a set of transformation features between the source voice and a database of voices [9]. The level of anonymization offered by previous work is not immediately clear, and hence the VoicePrivacy Challenge 2020 has been created to evaluate systems with common datasets, protcols and metrics [10].

2.2. The VoicePrivacy Challenge 2020

The VoicePrivacy Challenge 2020 [5] provides the setting for this paper, and defines a specific goal, set of datasets, and set of metrics for the evaluation and comparison of voice anonymization systems. The challenge seeks solutions for a scenario

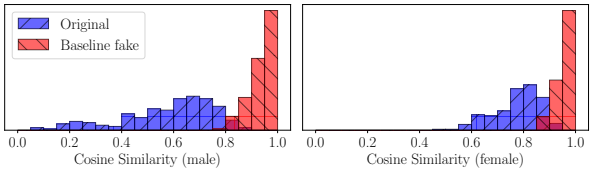

Figure 1: *Distribution of cross-cosine similarities between pairs of X-vectors from original voices and from the baseline fakes. The baseline fake X-vectors do not follow the same distribution of cosine similarities as the original X-vectors: these fake X-vectors are much more similar to one another than X-vectors extracted from organic speakers.*

where 'Speakers want to hide their identity whilst still allowing all other downstream goals to be achieved' [10]. This is done by converting a speaker to a *pseudo-speaker*, the new identity of the original speaker.

In order to meet the task of achieving downstream goals the following system requirements are given: (a) output a speech waveform, (b) hide speaker identity as much as possible, (c) distort other speech characteristics as little as possible, (d) ensure that all trial utterances from a given speaker appear to be uttered by the same pseudo-speaker, while trial utterances from different speakers appear to be uttered by different pseudo-speakers.

The challenge provides a common set of permitted datasets, to facilitate an even playing field. Likewise it provides the evaluation framework, consisting of a set of objective metrics presented in this work, as well as subjective metrics calculated by the challenge organisers in the future.

2.3. Threat Model

The challenge assumes that the attackers have access to one or more anonymized trial utterances, and possibly also to original or anonymized enrollment utterances for each speaker. The threat model states that the attacker does not have access to the anonymization system applied by the user. Whilst our submission operates under this threat model, we do not believe this assumption is necessarily the most reasonable for a speaker anonymization system. In fact, within the security field it is typical to assume that an attacker knows the details of the system (Kerckhoffs's principle).

3. System Overview

3.1. Rationale

Our system design follows the same approach as that of [11]. Fang et al. [11] proposed three techniques for generating fake X-vectors: (i) nearest speakers, (ii) random selection and (iii) range selection. The VoicePrivacy Challenge 2020 baseline system uses a variant of the last of these techniques, selecting the 200 furthest away X-vectors, and then averaging a random selection of 100 within these.

The rationale behind this work is that the fake X-vectors generation techniques mentioned above introduce a bias which leads to their distribution being different from that of the original X-vector's. In particular, we notice that mantaining the X-vector cross-similarity properties of the distribution is desirable, i.e., the similarity between fakes should have the same behavior as the similarity between originals. We show in Figure 1 the cross-similarity between each pair of X-vectors in the original pool and the fake pool. From Figure 1, it is evident that the

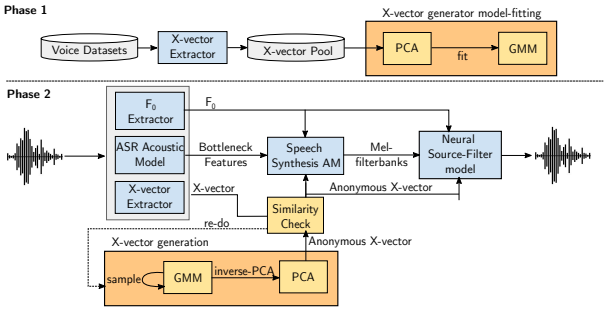

Figure 2: *Voice Anonymization system diagram. We replace the baseline X-vector anonymization module with a new generation method (shown in orange in the diagram).*

similarities in the fake pool are higher compared to the original pool[1]. This behavior comes as a consequence of averaging many X-vectors in the fake generation phase and leads to the global X-vector space being underutilized, which brings the following disadvantages:

- less entropy in fake X-vector space: anonymized voices are more similar to each other than pairs of original voices,
- reduced privacy, as it's easier to tell anonymized and original voices apart.
- the system requires a pool of X-vectors to sample from in the anonymization phase, which may lead to privacy leaks as this information needs to be shipped with the system.

In the following we explain how our system improves on the X-vector generation to maintain the desired similarity properties.

3.2. Method

In our method, we focus on improving the fake X-vector generation of the baseline system. In the baseline, three types of features are extracted for a speaker, the fundamental frequency, bottleneck features and the X-vector. The X-vector describes the speaker identity while the other features only encode the speech content [11].

Following the weakness outlined in Section 3.1, 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, we use a Gaussian Mixture Model (GMM). By using a generative model we avoid the bias introduced by the baseline fake X-vector generation, which generates them by averaging subgroups of population vectors. Whenever a voice needs to be anonymized, a reduced-dimensionality vector is randomly sampled from the GMM and then brought back into the 512-dimensional X-vector space by applying the PCA inverse transform. We note that the X-vector generation could be addressed by training a generative adversarial network, however GMMs have been shown to better generalize the captured distributions [12] and do not suffer from membership inference attacks (which could harm the system privacy guarantees) [13]. In Section 4.1 we show how we choose parameters for our method by monitoring the cross-cosine similarity between pairs of fake and original X-vectors.

As in the baseline, in the later stages of the anonymization

[1]We also note that female X-vectors follow a different distribution than male ones, probably an effect of the unbalanced training data

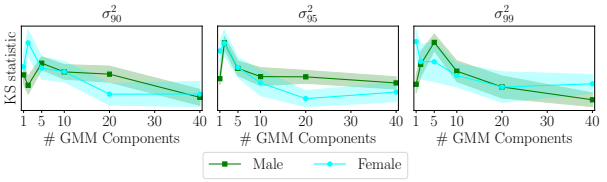

Figure 3: *KS statistic between cross-similarity distributions of our fake X-vectors and VoxCeleb X-vectors, for varied PCA retained variance and GMM no. of components, per gender.*

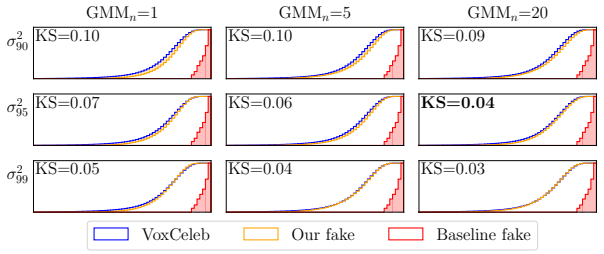

Figure 4: *eCDF comparison between our fake, Voxceleb test split and baseline X-vectors, for female speakers (male distributions are similar). The combination of parameters chosen for the evaluation is in bold. KS values are between VoxCeleb and our fakes, and are only directly comparable across rows.*

a Speech Synthesis acoustic model is used to generate Mel-filterbanks, which are fed with the F0 and new X-vector to a Neural source-filter model to generate audio. We train and reuse the models in the exact same way as the baseline for this, with the exception that we use the VoxCeleb1 [14] and VoxCeleb2 [15] datasets in our pool of speaker X-vectors, in addition to the LibriTTS [16] train-other-500 dataset. Figure 2 gives a full overview of how these system components fit together.

3.3. Forced Dissimilarity

Occasionally, our method might generate an X-vector from the GMM which is relatively close to the original user's voice, which may comprimise their anonymity. To mitigate this risk we propose an optional similarity check, termed *forced dissimilarity*, between the speaker's X-vector and the newly generated fake X-vector. If the cosine distance between the two X-vectors is above some threshold, then a new fake X-vector is generated. This operation adds almost no overhead to the system as it only requires a repeated sample from the GMM.

4. Experiments

4.1. Determining Optimal Parameters

4.1.1. Setup

To evaluate the performance of our fake X-vector generation, we perform an analysis on the resulting cross-similarity distribution of the generated vectors while varying the number of PCA and GMM components. We monitor how close the cross-similarity distribution between the fake and original X-vectors are (as in Figure 1) using the Kolmogorov-Smirnov (KS) test between the two distributions. The KS test quantifies the distance between two empirical cumulative distribution functions (eCDF), with lower scores implying that two distributions are more similar. For PCA, we focus on three values for the total amount of variance captured, namely 90%, 95% and 99%.

We setup the evaluation as follows, we extract all the X-vectors from VoxCeleb1, VoxCeleb2 (4,451 and 2,912 for male and female), for each gender we perform a 50% train-test split and we train our PCA+GMM on the training part. We then sample from the GMM and apply the PCA inverse transform to obtain 512-dimensional fake X-vectors. Then we compute the distribution of cross-cosine similarity on fake X-vectors and on the remaining 50% testing split, and we compute the KS statistic between the distributions. For the GMM we learn a diagonal covariance matrix, set the maximum number of EM iterations to 1,000 and the convergence tolerance to 10^{-16}.

4.1.2. Results

We report in Figure 3 the results for the three PCA models with increasing number of components used to fit the GMM, and we report some examples of the resulting eCDFs in Figure 4. We found that using one or two GMM component(s) lead to a relatively good fit for males but not for females, where there is a clearer decreasing trend as the no. of GMM components increases. Figure 4 shows how much closer our fake X-vectors approximate the cross-similarity distributions found in the VoxCeleb data compared to the baseline fake X-vectors. While increasing the number of components generally leads to a greater similarity between the distributions, in order to not overfit to the VoxCelebs data we choose to settle on using 95% of PCA-retained variance and 20 GMM components. This allows us to have a good approximation of the 512-dimensional X-vector space without requiring an overly complicated model.

4.2. VoicePrivacy Challenge 2020 Results

We focus mainly on Equal Error Rates (EER) and C_{llr}^{min} in our analysis, as C_{llr} is more ambiguous due to non-calibration [10]. Table 1 and Table 2 show the evaluation results computed our method, with and without forced dissimilarity (FD), respectively. Both tables also show the comparison to the baseline[2]. For FD, we set the similarity threshold to 0.8.

We find that our results for both versions of our method perform similarly. For scenarios where original enrollment and anonymized trial data are anonymized we achieve EERs of 42.0-46.6% for females and 45.2-50.2% for males on our standard system. These EER's and corresponding C_{llr}^{min} are slightly lower than the baseline solution, but are still close to 50%, which would indicate perfect anonymization.

In scenarios where both enrollment and trial data are anonymized we achieve EERs of 31.0-38.1% for females and 39.1-42.6% for males, for our standard system. For males this represent an improvement over the baseline of 8.2-16.8%, and -1.2-8.4% for females[3]. C_{llr}^{min} values follow a similar pattern of improvement. We believe the disparity in improvement for males and females may be due to the improved performance of the X-vector extractor on males, as seen by the wider distribution of similarities in Figure 1.

The improvements in EER and C_{llr}^{min} for scenarios with enrollment and trials anonymized show that our technique improves the heterogeneity of the anonymized voices. It is likely that some of the remaining difference between our results and a 50% EER is due to the X-vector system not perfectly decomposing the voice into speaker identity and non-speaker identity components. However, our method is general enough that it

[2]We omit results for VCTK (common) due to space constraints
[3]Performance is worse on VCTK test for women

Table 1: *Speaker verifiability results for the pretrained ASV_{eval} model. Results for our anonymization method with 20 GMM components and σ^2_{95} PCA, without forced distancing. In parenthesis we report the difference with the baseline system.*

Dataset	Gender	Anonymization		Development			Test		
		Enroll	Trial	EER (%)	C^{min}_{llr}	C_{llr}	EER (%)	C^{min}_{llr}	C_{llr}
LibriSpeech	Female	Original	Original	8.7	0.30	42.9	7.7	0.18	26.8
			Anonymized	43.3(-7.0)	0.97(-0.03)	134.7(-11.3)	42.0(-6.6)	0.97(-0.03)	145.5(-5.9)
		Anonymized		36.8(+1.7)	0.90(+0.02)	11.1(-4.1)	38.1(+8.4)	0.92(+0.12)	13.4(-0.6)
	Male	Original	Original	1.2	0.03	14.2	1.1	0.04	15.3
			Anonymized	50.2(-8.2)	0.98(-0.02)	147.7(-20.8)	49.4(-3.8)	0.98(-0.02)	174.1(+6.9)
		Anonymized		41.5(+11.8)	0.95(+0.14)	12.1(-8.0)	40.8(+8.2)	0.96(+0.12)	17.6(-8.9)
VCTK (diff)	Female	Original	Original	2.9	0.10	1.1	4.9	0.17	1.5
			Anonymized	46.6(-3.4)	0.94(-0.04)	168.0(+5.1)	43.4(-5.5)	0.98(-0.02)	148.3(+5.9)
		Anonymized		31.0(+1.5)	0.84(+0.02)	13.5(+3.3)	33.0(-1.2)	0.86(-0.02)	17.8(+5.5)
	Male	Original	Original	1.4	0.05	1.2	2.1	0.07	1.8
			Anonymized	45.2(-10.2)	0.99(-0.01)	154.8(-11.7)	46.7(-7.1)	0.99(-0.01)	162.5(-3.1)
		Anonymized		39.1(+13.0)	0.94(+0.18)	10.4(-8.4)	42.6(+16.8)	0.97(+0.23)	14.5(-1.8)

Table 2: *Speaker verifiability results for the pretrained ASV_{eval} model. The table shows the results for our anonymization method with 20 GMM components and σ^2_{95} PCA, with FD at a threshold of 0.8. In parenthesis we report the difference with the baseline system.*

Dataset	Gender	Anonymization		Development			Test		
		Enroll	Trial	EER (%)	C^{min}_{llr}	C_{llr}	EER (%)	C^{min}_{llr}	C_{llr}
LibriSpeech	Female	Original	Original	8.7	0.30	42.9	7.7	0.18	26.8
			Anonymized	43.6(-6.7)	0.97(-0.03)	134.3(-11.7)	42.3(-6.2)	0.97(-0.02)	146.8(-4.6)
		Anonymized		36.9(+1.8)	0.90(+0.03)	11.3(-3.9)	39.0(+9.3)	0.93(+0.13)	13.9(-0.1)
	Male	Original	Original	1.2	0.03	14.2	1.1	0.04	15.3
			Anonymized	49.5(-8.9)	0.98(-0.02)	147.8(-20.7)	49.4(-3.8)	0.98(-0.02)	175.1(+7.9)
		Anonymized		41.6(+11.9)	0.94(+0.14)	11.9(-8.2)	42.5(+10.0)	0.97(+0.13)	19.2(-7.3)
VCTK (diff)	Female	Original	Original	2.9	0.10	1.1	4.9	0.17	1.5
			Anonymized	46.4(-3.6)	0.94(-0.04)	167.4(+4.5)	43.2(-5.7)	0.98(-0.02)	147.8(+5.4)
		Anonymized		30.5(+1.1)	0.82(+0.01)	12.9(+2.7)	32.8(-1.4)	0.86(-0.03)	18.1(+5.8)
	Male	Original	Original	1.4	0.05	1.2	2.1	0.07	1.8
			Anonymized	45.5(-9.9)	0.99(-0.01)	155.3(-11.2)	47.5(-6.2)	0.99(-0.01)	162.6(-3.0)
		Anonymized		38.8(+12.7)	0.94(+0.18)	10.6(-8.2)	42.5(+16.6)	0.96(+0.22)	14.7(-1.6)

Table 3: *WER rates for original and anonymized voices, with and without forced distancing at a threshold of 0.8. In parenthesis we report the difference with the baseline system.*

		WER (%)	
Dataset	Anonymization	Dev.	Test
LibriSpeech	Original	3.83	4.14
	Anonymized	6.75 (+0.25)	7.26 (+0.49)
	Anonymized FD	6.63 (+0.13)	7.16 (+0.39)
VCTK	Original	10.79	12.81
	Anonymized	15.35 (-0.15)	15.56 (+0.03)
	Anonymized FD	15.22 (-0.28)	15.63 (+0.10)

could be applied to other systems which improve the voice identity extraction part.

Comparing our standard and FD methods, we notice that the results are similar, with the FD method having slightly better results with only trials anonymized slightly worse in trial and enroll anonymized scenarios. We do not expect the FD method to significantly impact our results, as it is intend as a measure to prevent transformations to anonymized voices that are very similar to the original voice, as opposed to improving results.

We observe that the word error rates (WER) for both our standard and FD systems are in the same regions as the baseline, with a slight degredation in quality on the LibriSpeech dataset, and a slight improvement overall for the VCTK scenarios, shown in Table 3.

5. Conclusions

In this work, we propose a scheme to anonymize voices in a way that better maintains the natural diversity of voices, as compared to previous approaches. We maintain this diversity by learning the properties of the X-vector space and using a generative model to sample from it, showing that such method better captures the distribution of similarities between fake vectors. This increase in the diversity of anonymized voices makes them more distinguishable from one another, as evidenced by improved results in scenarios where both enrollment and trial are anonymized. In our work we also propose to use forced dissimilarity, which allows a speaker to ensure that the anonymized voice they produce is not too similar to their own voice.

We experimentally validate that our proposed system produces voices that are more diverse, and evaluate our system using the VoicePrivacy Challenge 2020 baseline system. Our results in the challenge show a slight degredation in performance of anonymized voices against the original enrolled voices, but show a strong improvement when comparing two versions of the same persons voice anonymized. Our results also reveal worse performance for females than males, which we believe to be a result of the unbalanced dataset used for training, and highlight the opportunity to improve such bias.

6. Acknowledgements

This work was generously supported by a grant from Mastercard and by the Engineering and Physical Sciences Research Council [grant numbers EP/N509711/1, EP/P00881X/1]

7. References

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

[2] L. J. Liu, Z. H. Ling, Yuan-Jiang, Ming-Zhou, and L. R. Dai, "Wavenet vocoder with limited training data for voice conversion," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-Septe, no. September, pp. 1983–1987, 2018.

[3] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 019–10 029.

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

[5] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "The VoicePrivacy 2020 Challenge evaluation plan," 2020. [Online]. Available: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf

[6] R. V. Cox, D. E. Bock, K. B. Bauer, J. D. Johnston, and J. H. Synder, "Analog Voice Privacy System." *AT&T Technical Journal*, vol. 66, no. 1, pp. 119–131, 1987.

[7] K. Hashimoto, J. Yamagishi, and I. Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, pp. 5500–5504, 2016.

[8] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification via voice transformation," *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009*, pp. 529–533, 2009.

[9] F. Bahmaninezhad, C. Zhang, and J. Hansen, "Convolutional Neural Network Based Speaker De-Identification," vol. 2016, no. June, pp. 255–260, 2018.

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

[11] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker Anonymization Using X-vector and Neural Waveform Models," pp. 3–8, 2019. [Online]. Available: http://arxiv.org/abs/1905.13561

[12] E. Richardson and Y. Weiss, "On GANs and GMMs," in *Advances in Neural Information Processing Systems*, no. NeurIPS, 2018, pp. 5847–5858.

[13] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[14] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[15] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1929

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