The Risks and Detection of Overestimated Privacy Protection in Voice Anonymisation

Michele Panariello[1,*], *Sarina Meyer*[2,*], *Pierre Champion*[3], *Xiaoxiao Miao*[4], *Massimiliano Todisco*[1], *Ngoc Thang Vu*[2], *Nicholas Evans*[1]

[1]EURECOM, France; [2]Institute for Natural Language Processing, University of Stuttgart, Germany; [3]Inria, France; [4]Singapore Institute of Technology, Singapore

michele.panariello@eurecom.fr, sarina.meyer@ims.uni-stuttgart.de

Abstract

Voice anonymisation aims to conceal the voice identity of speakers in speech recordings. Privacy protection is usually estimated from the difficulty of using a speaker verification system to re-identify the speaker post-anonymisation. Performance assessments are therefore dependent on the verification model as well as the anonymisation system. There is hence potential for privacy protection to be overestimated when the verification system is poorly trained, perhaps with mismatched data. In this paper, we demonstrate the insidious risk of overestimating anonymisation performance and show examples of exaggerated performance reported in the literature. For the worst case we identified, performance is overestimated by 74% relative. We then introduce a means to detect when performance assessment might be untrustworthy and show that it can identify all overestimation scenarios presented in the paper. Our solution is openly available as a fork of the 2024 VoicePrivacy Challenge evaluation toolkit.

Index Terms: speaker anonymisation, voice privacy, evaluation

1. Introduction

Voice anonymisation is the task of processing a speech recording to conceal the voice identity of the speaker while retaining utility — with utility being defined according to some downstream task [1]. State-of-the-art solutions operate upon multiple features extracted from the input speech signal. Typically, they encode at least the desired information pertaining to the downstream task (e.g. spoken and emotional content). Voice conversion is then applied to synthesise an output speech signal which is devoid of the original voice identity but which contains the same desired information as the input. The anonymised output then contains the voice belonging to a so-called *target speaker*. Anonymisation is successful if an attacker cannot establish the voice identity of the original speaker.

Privacy protection is typically measured by simulating attacks with an automatic speaker verification (ASV) system which has the aim of linking anonymised data to the original speaker. Performance is inversely proportional to ASV reliability. Simulated attacks should be as strong as possible so that anonymisation performance is measured in a worst-case sense. This is usually achieved by training the ASV model using utterances anonymised by the same anonymisation system under evaluation. Within the scope of the VoicePrivacy Challenge (VPC) [2, 3], this is referred to as the *semi-informed* attack model. The VoicePrivacy Attacker Challenge [4] was launched recently to foster the development of stronger attacks to improve the reliability of performance estimates. At the time of writing, the semi-informed attack model and evaluation recipe described in [3] remains the de-facto standard.

Estimates of privacy protection reflect both that of the anonymisation system as well as the ASV system used for evaluation. Performance can be overestimated if the ASV system performs sub-optimally for reasons other than the anonymisation itself. For example, previous work [5] showed that some system design choices might prevent the ASV system from learning discriminative cues from anonymised data, resulting in exaggerated estimates of anonymisation performance. They are due not to strong anonymisation, but rather to the use of a suboptimal ASV attack model. The inadvertent overestimation of performance can be challenging to avoid; contrary to the design of reliable ASV systems, strong anonymisation calls for *poor* ASV performance. Poor performance can be achieved relatively easily, perhaps due to inadequate training choices or even design oversight. As a result, and other than to protect scientific integrity, there is comparatively less incentive to optimise an ASV model used for the evaluation of anonymisation systems. This raises challenges in benchmarking; results depend on the degree to which each ASV attack model is optimised with respect to the anonymisation system under evaluation.

In this paper, we report an investigation of potential evaluation pitfalls and propose what is, to the best of our knowledge, the first solution for their detection. We show that exaggerated privacy protection estimates can result from a mismatch between the distributions of anonymised test data and that used to train the ASV attack model. We demonstrate the pitfalls using artificial scenarios and then show similar trends in the evaluation of practical anonymisation systems reported in the literature. Once the distribution mismatch is addressed, we show that levels of anonymisation performance can, in some cases, be much lower than those reported. Finally, we show that the proposed approach to detection is effective in protecting against untrustworthy performance estimates.

2. Anonymisation methods

We report an evaluation of anonymisation performance for a set of scenarios involving some form of mismatch between the anonymisation system under test and that used to generate ASV training data. In this section we describe the set of systems used in our experiments. They include 3 VPC 2024 baseline systems [3]. For reasons which will become apparent later, we also used one more system based on self-supervised learning (SSL) and orthogonal householder neural networks (OHNN) [6].

Baseline **B3** [7] uses a transformer-based system [8] to produce a phonetic transcript of the input audio. An acoustic model is used with a prosodic representation of the input and a GAN-generated [9] speaker embedding to produce an anonymised

[*]These authors contributed equally.

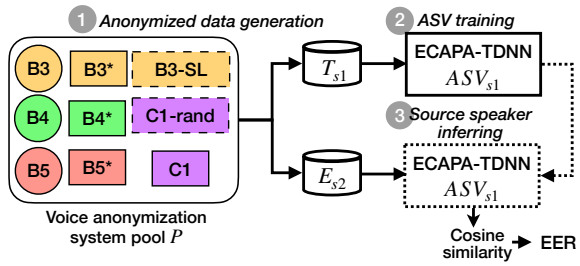

Figure 1: *Illustration of the mismatched evaluation procedure (best viewed in colour).* s_1 *and* s_2 *are voice anonymisation systems from* P*. Matched:* $s_1 = s_2$*. Full mismatch (Section 3.1):* $s_1 \neq s_2$*, different colours. Partial mismatch (Section 3.2):* $s_1 \neq s_2$*, same colour, different shapes. Hidden mismatch (Section 3.3):* $s_1 \neq s_2$*, same colour, different borders.*

Mel spectrogram which is converted to a waveform using a HiFi-GAN vocoder [10]. Baseline **B4** [11] extracts a representation of the spoken content from the input utterance in the form of semantic tokens. Using the EnCodec [12] encoder, a speaker representation in the form of acoustic tokens is extracted from a second, distinct utterance which contains the voice of the target speaker. The two sets of tokens are modeled in auto-regressive fashion to generate a new set of anonymised acoustic tokens which are decoded to a waveform using the EnCodec decoder. Baseline **B5** [5] is based upon the quantisation of wav2vec 2.0 [13] bottleneck features, which are concatenated with the F0 curve and a one-hot representation of the target speaker identity. Voice conversion is performed using a HiFi-GAN vocoder. The SSL-OHNN system, henceforth referred to as **C1**, uses a HuBERT-based soft content encoder [14] to extract the spoken content and an ECAPA-TDNN model [15] to extract a speaker embedding from the original utterance. Anonymisation of the speaker embedding is performed using an OHNN while voice conversion is performed directly using a HiFi-GAN vocoder.

We evaluate anonymisation performance according to the usual VPC 2024 policy [16, 3] and as shown in Figure 1. The evaluation of system S is performed using two sets of anonymised LibriSpeech data [17]: (a) the LibriSpeech test data (E_{S}) [3]; (b) the LibriSpeech *train-clean-360* set of data [17] which is used by the attacker to train an ASV system using similarly anonymised data. The datasets are anonymised in one of two different ways: (a) at the *utterance level*, in which each input utterance is anonymised towards a different target speaker; (b) at the *speaker level* in which, for all utterances produced by the same original speaker, anonymisation is performed towards the same target speaker.

For each anonymisation system S, the attacker trains a new ASV model $\mathrm{ASV}_{\mathrm{S}}$ using anonymised training data T_{S}. A data partition corresponding to 10% of T_{S} is set aside and used for validation during training. For all experiments reported in this paper, $\mathrm{ASV}_{\mathrm{S}}$ is an ECAPA-TDNN model [15]. Its use for privacy evaluation is mandated by the VPC evaluation plan [3] and makes our results comparable to others reported in the anonymisation literature. Speaker embeddings are extracted from the final hidden layer for anonymised enrollment and test utterances. The speaker of the former is either the same (a target trial) or different (a non-target trial) to that of the original test utterance (pre-anonymisation). Embeddings are scored using a cosine distance backend. As per [3], privacy protection is expressed in terms of the equal error rate (EER) for $\mathrm{ASV}_{\mathrm{S}}$ when evaluated using E_{S}. The higher the EER, the better the anonymisation: the attacker has more difficulty to reidentify the original speaker. An EER of 50% would indicate full anonymisation: the attacker cannot do better than to guess at random whether the speaker of the enrollment and test utterances is the same. This attack is referred to as *semi-informed*: the attacker knows the system S used to generate E_{S} and uses S to produce T_{S} with which to train $\mathrm{ASV}_{\mathrm{S}}$, as well as to anonymise the enrollment utterances. Readers are referred to [3, 18] for details.

Table 1: *Results for full mismatch scenarios (EER, %). The diagonal corresponds to the matched condition, the remaining values to the full mismatch.*

Attacker	Evaluation Data		
	E_{B3}	E_{B4}	E_{B5}
$\mathrm{ASV}_{\mathrm{B3}}$	**27.05**	44.23	44.40
$\mathrm{ASV}_{\mathrm{B4}}$	36.99	**29.82**	41.67
$\mathrm{ASV}_{\mathrm{B5}}$	37.28	38.72	**34.34**

3. Experiments

We first show that differences between the anonymisation systems used to generate anonymised evaluation data E_{S} and that used by the adversary for ASV training T_{S} can lead to privacy protection overestimation. We then show similar findings when the two anonymisation systems differ only in terms of a single module. Next, we show that estimates of performance can be lower still if the attacker instead trains the ASV system in a scenario subtly different to the usual *semi-informed* attack. In all of these cases, there is a *mismatch* between the two anonymisation systems. Last, we propose a technique to detect such mismatches with a view to mitigating the overestimation of anonymisation performance.

3.1. Full mismatch

Under *full mismatch* conditions, the two anonymisation systems are completely different. As illustrated in Figure 1, system s_1 is used by the attacker to generate ASV training data T_{s_1}. System s_2 is used to produce anonymised evaluation data E_{s_2}. We performed experiments using all possible pairings among systems B3, B4 and B5. Results are presented in Table 1. Each row reflects performance for $\mathrm{ASV}_{\mathrm{s}_1}$ trained using T_{s_1}, while columns correspond to evaluation data E_{s_2}. Shown in bold face along the diagonal are results for matched conditions where $\mathrm{s}_1 = \mathrm{s}_2$. This is the usual *semi-informed* attack scenario. Off-diagonal results correspond to the conditions of interest where $\mathrm{s}_1 \neq \mathrm{s}_2$. While this setup does not represent a realistic evaluation scenario, the implications of mismatched anonymisation systems remain unclear and have not been previously explored.

Our results show that the impact of such a mismatch can be substantial. An EER of 27% is produced when anonymised evaluation data and ASV training data are both generated using the same B3 system. However, the EER increases to 44% EER (near random guess) when the ASV system is trained using data generated using B4 or B5. Results are similar for B4 for which the EER increases from 30% (matched) to 42% (B5). For B5 the effect is less pronounced, yet not insignificant. While the above is an extreme, unrealistic example, we show next the potential for similar performance overestimates even when the differences between each anonymisation system is more subtle, including real examples from the literature.

Table 2: *Results for the partial mismatch scenario.*

Evaluation data	Attacker	EER (%)
E_{B3}	ASV_{B3}	27.05
	$\text{ASV}_{\text{B3*}}$	28.11
E_{B4}	ASV_{B4}	29.82
	$\text{ASV}_{\text{B4*}}$	35.83
E_{B5}	ASV_{B5}	34.34
	$\text{ASV}_{\text{B5*}}$	43.71

3.2. Partial mismatch

We evaluate the performance of an anonymisation system under a more realistic, *partial mismatch* scenario in which the anonymisation systems differ only in terms of a single module. Experiments were performed using the following system variants: **B3*** in which the HiFi-GAN is replaced with a BigVGAN vocoder [19]; **B4*** in which the EnCodec decoder is replaced with the Vocos [20] vocoder trained with character-level conditioning [21]; **B5*** which uses a factorised time delay neural network (TDNN-F) [22] bottleneck feature extractor instead of the wav2vec 2.0 [13] encoder.[1] The experimental setup is unchanged, but this time involves the evaluation of E_{B3} using either ASV_{B3} or $\text{ASV}_{\text{B3*}}$, E_{B4} using either ASV_{B4} or $\text{ASV}_{\text{B4*}}$, and E_{B5} using either ASV_{B5} or $\text{ASV}_{\text{B5*}}$.

Results are presented in Table 2. For evaluation data E_{B3}, the difference in performance is negligible, likely because the use of only a different vocoder has no consequential impact upon anonymisation behaviour. For E_{B4} the impact is more substantial, with a difference of 6% in EER for ASV_{B4} and $\text{ASV}_{\text{B4*}}$. The character conditioning mechanism of B4* likely has a greater impact on anonymisation. For E_{B5} the difference in EER is even more pronounced, with a difference of almost 10%. The use of a different feature extractor early in the pipeline results in almost completely different systems (cf. results in Table 1).

Even a partial mismatch between anonymisation systems can still induce substantially exaggerated performance estimates. While even this scenario might easily be avoided, there are practical examples in the literature. In [6] (C1) and the system proposed in [23], the attacker is assumed to retrain one of the system modules, thereby inducing a partial mismatch. In [24], E_{S} is formed by mixing data generated using two different anonymisation systems, and the attacker is assumed not to know with which system each test utterance is anonymised; since the ASV model cannot be matched to both systems, there is again a partial mismatch. In all of these works, evaluation is reportedly performed under the VPC-defined *semi-informed* attack model. This observation shows that its definition is inadequate and that differences in interpretation can lead to questionable performance comparisons. This finding calls for the design of an automatic method to estimate the reliability of results derived using an ASV system ASV_{S} regardless of the attack definition. We propose one such method in Section 3.4.

3.3. Hidden mismatch

As we now show, some mismatches can be more insidious. We start with the case of system **C1** [6] for which results are shown to the left in Table 3. When evaluation is performed using ASV_{C1} and E_{C1} we obtain an EER of 41%. However, we found

[1] System B5* is identical to B6 in [3]. Denoted here as B5* for consistency as a *variant* of system B5.

Table 3: *Results for the hidden mismatch scenario (EER, %).*

E_{C1}		$E_{\text{B3-SL}}$	
ASV_{C1}	$\text{ASV}_{\text{C1-rand}}$	$\text{ASV}_{\text{B3-SL}}$	ASV_{B3}
40.80	10.38	44.78	27.31

that the EER can be reduced very substantially to 10% (a 74% relative decrease) through a trivial retraining of the ASV system. The retrained system, denoted $\text{ASV}_{\text{C1-rand}}$, stems from the anonymisation of each utterance within $T_{\text{C1-rand}}$ using an embedding extracted for a speaker selected at random from the *LibriTTS train-other-500* dataset [25] rather than that extracted for speakers selected using the OHNN-based target speaker selection strategy. We now illustrate why such a training approach results in a more effective attack.

The OHNN-based target speaker selection strategy is a completely deterministic function $\text{OHNN}(\mathbf{x})$ applied to a speaker embedding $\mathbf{x}$. It is well-known [26, 27] that, if the target speaker selection strategy is not sufficiently random, then utterance-level anonymisation behaves similarly to speaker-level anonymisation. This is because, if two embeddings $\mathbf{x}_1$ and $\mathbf{x}_2$ are extracted from utterances produced by the same speaker, then $\mathbf{x}_1 \approx \mathbf{x}_2$. It follows then that $\text{OHNN}(\mathbf{x}_1) \approx \text{OHNN}(\mathbf{x}_2)$, hence speaker-level, rather than utterance-level anonymisation.

It has been shown previously that speaker-level anonymisation of T_{S} results in a weaker attack [28], yet a thorough explanation for why is lacking. We now offer such an explanation. If speaker A in T_{S} is consistently anonymised towards target speaker B, ASV_{S} may simply learn to associate speaker B's voice characteristics with label A. As a result, it may rely on the $B \Rightarrow A$ mapping instead of learning speaker-discriminative features that persist despite anonymisation. Consequently, ASV_{S} is likely to fail on E_{S}, which includes different speakers and different anonymisation mappings. This is the case for C1 since the attacker retrains the OHNN, thereby resulting in a different transformation $\text{OHNN}(\mathbf{x})$, the learning of unreliable cues, and hence overestimated anonymisation performance. By using a random target speaker selection strategy which results in $T_{\text{C1-rand}}$ being anonymised at the utterance level, $\text{ASV}_{\text{C1-rand}}$ will no longer learn unreliable cues. This results in the learning of other, more reliable cues, hence the lower EER.

To provide empirical evidence in support of these arguments, we conducted a set of experiments using a variant of B3, denoted B3-SL, which performs speaker-level rather than utterance-level anonymisation. We used B3-SL to create $E_{\text{B3-SL}}$ and $T_{\text{B3-SL}}$, both anonymised at the speaker level but with different target speaker mappings. As shown to the right in Table 3, the resulting $\text{ASV}_{\text{B3-SL}}$ system produces an EER of almost 45% when assessed using $E_{\text{B3-SL}}$. We then repeated the evaluation using $E_{\text{B3-SL}}$ and ASV_{B3} trained using data anonymised at the utterance level and obtained a greatly-reduced EER of 27%. $\text{ASV}_{\text{B3-SL}}$ provides an overestimation of privacy protection as a result of being trained using data anonymised at the speaker level. These findings show that the evaluation of anonymisation systems performed using the same system as that used in generating E_{S} may not always provide reliable performance estimates. Such hidden mismatches can be challenging to mitigate.

3.4. Detecting mismatches

The overestimation of privacy protection can be attributed to ASV_{S} *overfitting* to the data distribution of T_{S} due either to the use of mismatched anonymisation systems (Sections 3.1

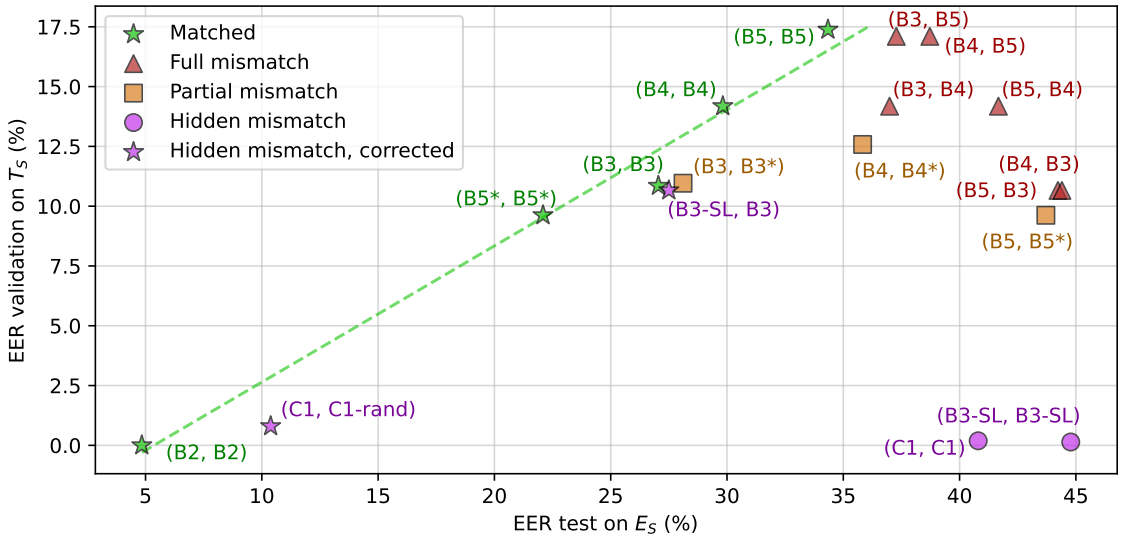

Figure 2: *Values of EER_{val} plotted against EER_{test} for all considered systems, denoted as (E_S, T_S) pairs. The green dashed line is the regression line of the "Matched" systems. Systems falling below this line have a potentially mismatched evaluation.*

and 3.2) or the use of ineffective training strategies (Section 3.3). This causes ASV_S to underperform on E_S. In the following we propose one solution to detect overfitting using validation data [29]. Our hypothesis is that differences in EER for validation and test data derived using an attacking ASV system can serve as an indication of overfitting. The validation data is the same 10% of T_s used for ASV training (see Section 2). In contrast to E_S, the validation data contains the same speakers as T_s. We design a new ASV protocol for the validation data. For each speaker, 5 utterances are used for enrollment. Those remaining are reserved for ASV trials. On average, there are 8 trials per speaker, 3 of which are target trials. ASV_S is trained in the usual way and then applied to the validation data in addition to the usual evaluation data E_S resulting in a pair of performance estimates, EER_{val} and EER_{test}. We performed this experiment using systems B3, B4 and B5 under full, partial and hidden mismatch conditions. To provide more data points, we also used the B5 variant B5* in addition to the less competitive B2 baseline [30].

Results are illustrated in Figure 2, where the x and y axes correspond to EER_{test} and EER_{val} respectively. Green stars correspond to matched scenarios for 2024 VPC baseline systems. Red triangles correspond to full mismatch (Section 3.1). Orange squares correspond to partial mismatch (Section 3.2). Purple circles indicate hidden mismatch, while purple stars correspond to the same scenarios with corrected mismatch (Section 3.3). Annotations signify the systems involved in each evaluation, e.g. (B4,B5) corresponds to E_{B4} being used for evaluation of ASV_{B5}. For all cases, EER_{val} is substantially lower than EER_{test}: this is on the account of the overlap between speakers in the validation set and the training set T_S. The gap between EER_{val} and EER_{test} is more pronounced for mismatched scenarios. For instance, in the case of matched (B3,B3), the EER_{test} of 27% falls to an EER_{val} of 11%, a relative drop of 60%. However, for the full mismatched case of (B4,B3) an EER_{test} of 44% falls to an EER_{val} of 10%, a relative drop of 76%. The trend is consistent; for mismatched scenarios the gap between EER_{test} and EER_{val} is higher than that for matched scenarios.

Also plotted in Figure 2 is a regression line computed over the EER_{test}, EER_{val} point pairs corresponding to matched scenarios, all of which we assume to be well-evaluated reference cases. Nearly all other points, all corresponding to mismatched scenarios, fall below this line suggesting that the difference between EER_{val} and EER_{test} serves as an indicator of how well-trained or well-matched is ASV_S to the evaluation of anonymisation system S: the greater the gap, the higher the chance of there being a mismatch between training and evaluation data, and therefore a less reliable estimation of anonymisation performance. We also note that cases of hidden mismatch are the farthest from the regression line. However, corrected counterparts are closer to the regression line, indicating more reliable performance estimation.

We argue that the community should consider the adoption of such detection techniques for the evaluation of anonymisation systems so as to reduce the risk of overestimating performance. To encourage work in this direction, we have integrated our approach into a publicly available fork of the 2024 VPC evaluation toolkit[2] which includes the automatic computation of EER_{val} within the validation step.

4. Conclusions

In this paper we demonstrate the risk of overestimating anonymisation performance. Using several state-of-the-art anonymisation approaches, we show that the risk stems from mismatches between the data used to train the speaker verification system employed for evaluation and the anonymised data under test. We demonstrate the risk with artificially introduced mismatch and that similar, hidden mismatches also afflict the evaluation of practical systems reported in the literature leading to exaggerated reports of performance. Based upon the comparison of results derived for test and validation sets, our method to detect overestimated performance is effective in identifying all examples reported in the paper. Consequently, we advocate for the adoption of such detection techniques within the community and for future editions of the VoicePrivacy Challenge to protect the trust in performance estimates. Future work should extend our analysis to other sources of mismatch and their detection.

[2]https://github.com/DigitalPhonetics/voice-privacy-evaluation

5. Acknowledgements

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project: Multilingual Controllable Voice Privacy (VoiPy) - Project number 533241795.

6. References

[1] 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," in *Interspeech 2020*, 2020, pp. 1693–1697.

[2] M. Panariello, N. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. Evans, E. Vincent, and J. Yamagishi, "The VoicePrivacy 2022 Challenge: Progress and Perspectives in Voice Anonymisation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3477–3491, 2024.

[3] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, "The VoicePrivacy 2024 Challenge Evaluation Plan," 2024. [Online]. Available: https://arxiv.org/abs/2404.02677

[4] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The First VoicePrivacy Attacker Challenge Evaluation Plan," 2024. [Online]. Available: https://arxiv.org/abs/2410.07428

[5] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," Ph.D. dissertation, Université de Lorraine, 2023.

[6] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. Tomashenko, "Speaker Anonymization Using Orthogonal Householder Neural Network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3681–3695, 2023.

[7] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody Is Not Identity: A Speaker Anonymization Approach Using Prosody Cloning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[8] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *International Conference on Machine Learning (ICML)*, 2022, pp. 17 627–17 643.

[9] H. Liu, X. Gu, and D. Samaras, "Wasserstein GAN with quadratic transport cost," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4832–4841.

[10] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033.

[11] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker Anonymization Using Neural Audio Codec Language Models," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 4725–4729.

[12] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *Transactions on Machine Learning Research*, 2023, featured Certification, Reproducibility Certification.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[14] B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in *Proc. ICASSP*. IEEE, 2022, pp. 6562–6566.

[15] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech 2020*, 2020, pp. 3830–3834.

[16] S. Meyer, X. Miao, and N. T. Vu, "VoicePAT: An Efficient Open-Source Evaluation Toolkit for Voice Privacy Research," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 257–265, 2024.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[18] B. M. Lal Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2802–2806.

[19] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A Universal Neural Vocoder with Large-Scale Training," in *The Eleventh International Conference on Learning Representations*, 2023.

[20] H. Siuzdak, "Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis," in *The Twelfth International Conference on Learning Representations*, 2024.

[21] M. Panariello, M. Todisco, and N. Evans, "Preserving spoken content in voice anonymisation with character-level vocoder conditioning," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 12–16.

[22] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Interspeech 2018*, 2018, pp. 3743–3747.

[23] L. Chen, K. A. Lee, W. Guo, and Z.-H. Ling, "Modeling Pseudo-Speaker Uncertainty in Voice Anonymization," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 601–11 605.

[24] H. L. Xinyuan, Z. Cai, A. Garg, K. Duh, L. P. García-Perera, S. Khudanpur, N. Andrews, and M. Wiesner, "HLTCOE JHU Submission to the Voice Privacy Challenge 2024," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 61–66.

[25] 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 *Interspeech 2019*, 2019, pp. 1526–1530.

[26] M. Panariello, M. Todisco, and N. Evans, "Vocoder drift in x-vector–based speaker anonymization," in *Interspeech 2023*, 2023, pp. 2863–2867.

[27] ——, "Vocoder drift compensation by x-vector alignment in speaker anonymisation," in *3rd Symposium on Security and Privacy in Speech Communication*, 2023, pp. 16–20.

[28] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, "Differentially Private Speaker Anonymization," *Proceedings on Privacy Enhancing Technologies*, 2023.

[29] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer Publishing Company, Incorporated, 2014.

[30] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker Anonymisation Using the McAdams Coefficient," in *Interspeech 2021*, 2021, pp. 1099–1103.