Introducing the VoicePrivacy initiative

Natalia Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, et al.

Introducing the VoicePrivacy Initiative

N. Tomashenko[1], B. M. L. Srivastava[2], X. Wang[3], E. Vincent[4], A. Nautsch[5], J. Yamagishi[3,6], N. Evans[5], J. Patino[5], J.-F. Bonastre[1], P.-G. Noé[1], M. Todisco[5]

[1]LIA, University of Avignon, France [2]Inria, France [3]NII, Tokyo, Japan [4]Université de Lorraine, CNRS, Inria, LORIA, France [5]EURECOM, France [6]University of Edinburgh, UK

organisers@lists.voiceprivacychallenge.org

Abstract

The VoicePrivacy initiative aims to promote the development of privacy preservation tools for speech technology by gathering a new community to define the tasks of interest and the evaluation methodology, and benchmarking solutions through a series of challenges. In this paper, we formulate the voice anonymization task selected for the VoicePrivacy 2020 Challenge and describe the datasets used for system development and evaluation. We also present the attack models and the associated objective and subjective evaluation metrics. We introduce two anonymization baselines and report objective evaluation results.

Index Terms: privacy, anonymization, speech synthesis, voice conversion, speaker verification, automatic speech recognition

1. Introduction

Recent years have seen mounting calls for the preservation of privacy when treating or storing personal data. This is not least the result of the European general data protection regulation (GDPR). While there is no legal definition of privacy [1], speech data encapsulates a wealth of personal information that can be revealed by listening or by automated systems [2]. This includes, e.g., age, gender, ethnic origin, geographical background, health or emotional state, political orientations, and religious beliefs, among others [3, p. 62]. In addition, speaker recognition systems can reveal the speaker's identity. It is thus of no surprise that efforts to develop privacy preservation solutions for speech technology are starting to emerge. The VoicePrivacy initiative aims to gather a new community to define the tasks of interest and the evaluation methodology, and to benchmark these solutions through a series of challenges.

Current methods fall into four categories: deletion, encryption, distributed learning, and anonymization. Deletion methods [4, 5] are meant for ambient sound analysis. They delete or obfuscate any overlapping speech to the point where no information about it can be recovered. Encryption methods [6, 7] such as fully homomorphic encryption [8] and secure multiparty computation [9], support computation upon data in the encrypted domain. They incur significant increases in computational complexity, which require special hardware. Decentralized or federated learning methods aim to learn models from distributed data without accessing it directly [10]. The derived data used for learning (e.g., model gradients) may still leak information about the original data, however [11].

Anonymization refers to the goal of suppressing personally identifiable attributes of the speech signal, leaving all other attributes intact[1]. Past and recent attempts have focused on noise addition [12], speech transformation [13], voice conversion [14–17], speech synthesis [18, 19], or adversarial learning [20]. In contrast to the above categories of methods, anonymization appears to be more flexible since it can selectively suppress or retain certain attributes and it can easily be integrated within existing systems. Despite the appeal of anonymization and the urgency to address privacy concerns, a formal definition of anonymization and attacks against it is missing. Furthermore, the level of anonymization offered by existing solutions is unclear and not meaningful because there are no common datasets, protocols and metrics.

For these reasons, the VoicePrivacy 2020 Challenge focuses on the task of speech anonymization. This paper is intended as a general reference about the Challenge for researchers, engineers and privacy professionals. Details for participants are provided in the evaluation plan [21] and on the challenge website[2].

The paper is structured as follows. The anonymization task and the attack models, the datasets, and the metrics are described in Sections 2, 3, and 4, respectively. The two baseline systems and the corresponding objective evaluation results are presented in Section 5. We conclude in Section 6.

2. Anonymization task and attack models

Privacy preservation is formulated as a game between *users* who publish some data and *attackers* who access this data or data derived from it and wish to infer information about the users [22, 23]. To protect their privacy, the users publish data that contain as little personal information as possible while allowing one or more downstream goals to be achieved. To infer personal information, the attackers may use additional prior knowledge.

Focusing on speech data, a given privacy preservation scenario is specified by: (i) the nature of the data: waveform, features, etc., (ii) the information seen as personal: speaker identity, traits, spoken contents, etc., (iii) the downstream goal(s): human communication, automated processing, model training, etc., (iv) the data accessed by the attackers: one or more utterances, derived data or model, etc., (v) the attackers' prior knowledge: previously published data, privacy preservation method applied, etc. Different specifications lead to different privacy preservation methods from the users' point of view and different attacks from the attackers' point of view.

2.1. Privacy preservation scenario

VoicePrivacy 2020 considers the following scenario, where the terms *"user"* and *"speaker"* are used interchangeably. Speakers want to hide their identity while still allowing all other downstream goals to be achieved. Attackers have access to one or more utterances and want to identify the speakers.

[1]In the legal community, the term "anonymization" means that this goal has been achieved. Here, it refers to the task to be addressed, even when the method being evaluated has failed. We expect the VoicePrivacy initiative to lead to the definition of new, unambiguous terms.

[2]https://www.voiceprivacychallenge.org/

2.2. Anonymization task

To hide his/her identity, each speaker passes his/her utterances through an anonymization system. The resulting anonymized utterances are referred to as *trial* data. They sound as if they had been uttered by another speaker called *pseudo-speaker*, which may be an artificial voice not corresponding to any real speaker.

The task of challenge participants is to design this anonymization system. In order to allow all downstream goals to be achieved, this system should: (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[3].

Requirement (c) is assessed via *utility* metrics: automatic speech recognition (ASR) decoding error rate using a model trained on *original*, i.e., unprocessed data and subjective speech intelligibility and naturalness (see Section 4). Requirement (d) and additional downstream goals including ASR training will be assessed in a post-evaluation phase (see Section 6).

2.3. Attack models

The attackers have access to: (a) one or more anonymized trial utterances, (b) possibly, original or anonymized *enrollment* utterances for each speaker. They do not have access to the anonymization system applied by the user. The protection of personal information is assessed via *privacy* metrics, including objective speaker verifiability and subjective speaker verifiability and linkability. These metrics assume different attack models.

The objective speaker verifiability metrics assume that the attackers have access to a single anonymized trial utterance and several enrollment utterances. Two sets of metrics are used for original vs. anonymized enrollment data (see Section 4.1). In the latter case, we assume that the trial and enrollment utterances of a given speaker have been anonymized using the same system, but the corresponding pseudo-speakers are different.

The subjective speaker verifiability metric (Section 4.2) assumes that the attackers have access to a single anonymized trial utterance and a single original enrollment utterance. Finally, the subjective speaker linkability metric (Section 4.2) assumes that the attackers have access to several anonymized trial utterances.

3. Datasets

Several publicly available corpora are used for the training, development and evaluation of speaker anonymization systems.

3.1. Training set

The training set comprises the 2,800 h *VoxCeleb-1,2* speaker verification corpus [24,25] and 600 h subsets of the *LibriSpeech* [26] and *LibriTTS* [27] corpora, which were initially designed for ASR and speech synthesis, respectively. The selected subsets are detailed in Table 1 (top).

3.2. Development set

The development set comprises *LibriSpeech dev-clean* and a subset of the VCTK corpus [28] denoted as *VCTK-dev* (see Table 1, middle). With the above attack models in mind, we split them into trial and enrollment subsets. For *LibriSpeech dev-clean*, the speakers in the enrollment set are a subset of those in the trial set. For *VCTK-dev*, we use the same speakers for enrollment and trial and we consider two trial subsets, denoted as *common* and *different*. The *common* trial subset is composed of utterances $\#1-24$ in the VCTK corpus that are identical for all speakers. This is meant for subjective evaluation of speaker verifiability/linkability in a text-dependent manner. The enrollment and *different* trial subsets are composed of distinct utterances for all speakers.

Table 1: *Number of speakers and utterances in the VoicePrivacy 2020 training, development, and evaluation sets.*

Subset			Female	Male	Total	#Utter.
Training	VoxCeleb-1,2		2,912	4,451	7,363	1,281,762
	LibriSpeech train-clean-100		125	126	251	28,539
	LibriSpeech train-other-500		564	602	1,166	148,688
	LibriTTS train-clean-100		123	124	247	33,236
	LibriTTS train-other-500		560	600	1,160	205,044
Development	LibriSpeech dev-clean	Enrollment	15	14	29	343
		Trial	20	20	40	1,978
	VCTK-dev	Enrollment	15	15	30	600
		Trial (common)				695
		Trial (different)				10,677
Evaluation	LibriSpeech test-clean	Enrollment	16	13	29	438
		Trial	20	20	40	1,496
	VCTK-test	Enrollment	15	15	30	600
		Trial (common)				70
		Trial (different)				10,748

3.3. Evaluation set

Similarly, the evaluation set comprises *LibriSpeech test-clean* and a subset of VCTK called *VCTK-test* (see Table 1, bottom).

4. Utility and privacy metrics

Following the attack models in Section 2.3, we consider objective and subjective privacy metrics to assess anonymization performance in terms of speaker verifiability and linkability. We also propose objective and subjective utility metrics to assess whether the requirements in Section 2.2 are fulfilled.

4.1. Objective metrics

For objective evaluation, we train two systems to assess speaker verifiability and ASR decoding error. The first system denoted ASV_{eval} is an automatic speaker verification (ASV) system, which produces log-likelihood ratio (LLR) scores. The second system denoted ASR_{eval} is an ASR system which outputs a word error rate (WER). Both are trained on *LibriSpeech train-clean-360* using Kaldi [29].

4.1.1. Objective speaker verifiability

The ASV_{eval} system for speaker verifiability evaluation relies on x-vector speaker embeddings and probabilistic linear discriminant analysis (PLDA) [30]. Three metrics are computed: the equal error rate (EER) and the LLR-based costs C_{llr} and $C_{\text{llr}}^{\text{min}}$. Denoting by $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ the false alarm and miss rates at threshold θ, the EER corresponds to the threshold θ_{EER} at which the two detection error rates are equal, i.e., $\text{EER} = P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$. C_{llr} is computed from PLDA

[3] This is akin to "pseudonymization", which replaces each user's identifiers by a unique key. We do not use this term here, since it often refers to the distinct case when the identifiers are tabular data and the data controller stores the correspondence table linking users and keys.

scores as defined in [31, 32]. It can be decomposed into a discrimination loss ($C_{\text{llr}}^{\text{min}}$) and a calibration loss ($C_{\text{llr}} - C_{\text{llr}}^{\text{min}}$). $C_{\text{llr}}^{\text{min}}$ is estimated by optimal calibration using monotonic transformation of the scores to their empirical LLR values.

As shown in Fig. 1, these metrics are computed and compared for: (1) original trial and enrollment data, (2) anomymized trial data and original enrollment data, (3) anomymized trial and enrollment data. The number of target and impostor trials is given in Table 2.

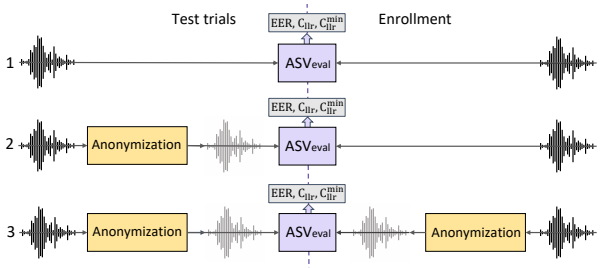

Figure 1: *ASV evaluation.*

Table 2: *Number of speaker verification trials.*

Subset		Trials	Female	Male	Total
Development	LibriSpeech dev-clean	Target	704	644	1,348
		Impostor	14,566	12,796	27,362
	VCTK-dev	Target (common)	344	351	695
		Target (different)	1,781	2,015	3,796
		Impostor (common)	4,810	4,911	9,721
		Impostor (different)	13,219	12,985	26,204
Evaluation	LibriSpeech test-clean	Target	548	449	997
		Impostor	11,196	9,457	20,653
	VCTK-test	Target (common)	346	354	700
		Target (different)	1,944	1,742	3,686
		Impostor (common)	4,838	4,952	9,790
		Impostor (different)	13,056	13,258	26,314

4.1.2. *ASR decoding error*

ASR_{eval} is based on the state-of-the-art Kaldi recipe for LibriSpeech involving a factorized time delay neural network (TDNN-F) acoustic model (AM) [33, 34] and a trigram language model. As shown in Fig. 2, the (1) original and (2) anonymized trial data is decoded using the provided pretrained ASR_{eval} model and the corresponding WERs are calculated.

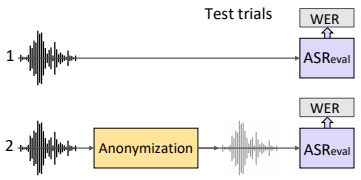

Figure 2: *ASR decoding evaluation.*

4.2. Subjective metrics

Subjective metrics include speaker verifiability, speaker linkability, speech intelligibility, and speech naturalness. They will be evaluated using listening tests carried out by the organizers.

4.2.1. *Subjective speaker verifiability*

To evaluate subjective speaker verifiability, listeners are given pairs of one anonymized trial utterance and one distinct original enrollment utterance of the same speaker. Following [35], they are instructed to imagine a scenario in which the anonymized sample is from an incoming telephone call, and to rate the similarity between the voice and the original voice using a scale of 1 to 10, where 1 denotes 'different speakers' and 10 denotes 'the same speaker' with highest confidence. The performance of each anonymization system will be visualized through detection error tradeoff (DET) curves.

4.2.2. *Subjective speaker linkability*

The second subjective metric assesses speaker linkability, i.e., the ability to cluster several utterances into speakers. Listeners are asked to place a set of anonymized trial utterances from different speakers in a 1- or 2-dimensional space according to speaker similarity. This relies on a graphical interface, where each utterance is represented as a point in space and the distance between two points expresses subjective speaker dissimilarity.

4.2.3. *Subjective speech intelligibility*

Listeners are also asked to rate the intelligibility of individual samples (anonymized trial utterances or original enrollment utterances) on a scale from 1 (totally unintelligible) to 10 (totally intelligible). The results can be visualized through DET curves.

4.2.4. *Subjective speech naturalness*

Finally, the naturalness of the anonymized speech will be evaluated on a scale from 1 (totally unnatural) to 10 (totally natural).

5. Baseline software and results

Two anonymization baselines are provided.[4] We briefly introduce them and report the corresponding objective results below.

5.1. Anonymization baselines

The primary baseline shown in Fig. 3 is inspired from [18] and comprises three steps: (1) extraction of x-vector [30], pitch (F0) and bottleneck (BN) features; (2) x-vector anonymization; (3) speech synthesis (SS) from the anonymized x-vector and the original F0+BN features. In *Step 1*, 256-dimensional BN features encoding spoken content are extracted using a TDNN-F ASR AM trained on *LibriSpeech train-clean-100* and *train-other-500* using Kaldi. A 512-dimensional x-vector encoding the speaker is extracted using a TDNN trained on *VoxCeleb-1,2* with Kaldi. In *Step 2*, for every source x-vector, an

[4] https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

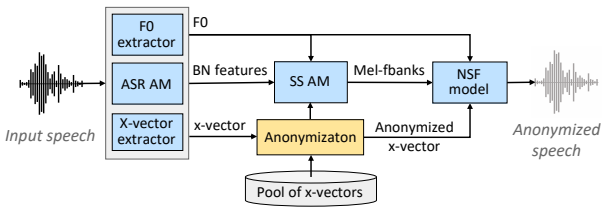

Figure 3: *Primary baseline anonymization system.*

Table 3: *Speaker verifiability achieved by the pretrained ASV_{eval} model. The primary baseline is used for anonymization.*

Dataset	Gender	Anonymization		Development			Test		
		Enroll	Trial	EER (%)	$C_{\text{llr}}^{\text{min}}$	C_{llr}	EER (%)	$C_{\text{llr}}^{\text{min}}$	C_{llr}
LibriSpeech	Female	original	original	8.67	0.304	42.86	7.66	0.183	26.79
			anonymized	50.14	0.996	144.11	47.26	0.995	151.82
		anonymized	anonymized	36.79	0.894	16.35	32.12	0.839	16.27
	Male	original	original	1.24	0.034	14.25	1.11	0.041	15.30
			anonymized	57.76	0.999	168.99	52.12	0.999	166.66
		anonymized	anonymized	34.16	0.867	24.72	36.75	0.903	33.93
VCTK (different)	Female	original	original	2.86	0.100	1.13	4.89	0.169	1.50
			anonymized	49.97	0.989	166.03	48.05	0.998	146.93
		anonymized	anonymized	26.11	0.760	8.41	31.74	0.847	11.53
	Male	original	original	1.44	0.052	1.16	2.07	0.072	1.82
			anonymized	53.95	1.000	167.51	53.85	1.000	167.82
		anonymized	anonymized	30.92	0.839	23.80	30.94	0.834	23.84

Table 4: *ASR decoding error achieved by the pretrained ASR_{eval} model. The primary baseline is used.*

Dataset	Anonymization	Dev. WER (%)	Test WER (%)
LibriSpeech	original	3.83	4.15
	anonymized	6.39	6.73
VCTK (comm.+diff.)	original	10.79	12.82
	anonymized	15.38	15.23

anonymized x-vector is computed by finding the N farthest x-vectors in an external pool (*LibriTTS train-other-500*) according to the PLDA distance, and by averaging N^* randomly selected vectors among them[5]. In *Step 3*, an SS AM generates Mel-filterbank features given the anonymized x-vector and the F0+BN features, and a neural source-filter (NSF) waveform model [36] outputs a speech signal given the anonymized x-vector, the F0, and the generated Mel-filterbank features. The SS AM and NSF models are both trained on *LibriTTS train-clean-100*. See [21, 37] for further details.

The secondary baseline is a simpler, formant-shifting approach provided as additional inspiration [38].

5.2. Objective evaluation results

Table 3 reports the values of objective speaker verifiability metrics obtained before/after anonymization with the primary baseline.[6] The EER and $C_{\text{llr}}^{\text{min}}$ metrics behave similarly, while interpretation of C_{llr} is more challenging due to non-calibration[7]. We hence focus on the EER below. On all datasets, anonymization of the trial data greatly increases the EER. This shows that the anonymization baseline effectively increases the users' privacy. The EER estimated with original enrollment data (47 to 58%), which is comparable to or above the chance value (50%), suggests that full anonymization has been achieved. However, anonymized enrollment data result in a much lower EER (26 to 37%), which suggests that F0+BN features retain some information about the original speaker. If the attackers have access to such enrollment data, they will be able to re-identify users almost half of the time. Note also that the EER is larger for females than males on average. This further demonstrates that failing to define the attack model or assuming a naive attack model leads to a greatly overestimated sense of privacy [23].

[5]In the baseline, we use $N = 200$ and $N^* = 100$.

[6]Results on VCTK (common) are omitted due to space constraints.

[7]In particular, $C_{\text{llr}} > 1$ is not a problem, since we care more about discrimination metrics than score calibration metrics in the first edition.

Table 4 reports the WER achieved before/after anonymization with the primary baseline. While the absolute WER stays below 7% on LibriSpeech and 16% on VCTK, anonymization incurs a large WER increase of 19 to 67% relative.

The results achieved by the secondary baseline are inferior and detailed in [21]. Overall, there is substantial potential for challenge participants to improve over the two baselines.

6. Conclusions

The VoicePrivacy initiative aims to promote the development of private-by-design speech technology. Our initial event, the VoicePrivacy 2020 Challenge, provides a complete evaluation protocol for voice anonymization systems. We formulated the voice anonymization task as a game between users and attackers, and highlighted three possible attack models. We also designed suitable datasets and evaluation metrics, and we released two open-source baseline voice anonymization systems. Future work includes evaluating and comparing the participants' systems using objective and subjective metrics, computing alternative objective metrics relating to, e.g., requirement (d) in Section 2.2, and drawing initial conclusions regarding the best anonymization strategies for a given attack model. A revised, stronger evaluation protocol is also expected as an outcome.

In this regard, it is essential to realize that the users' downstream goals and the attack models listed above are not exhaustive. For instance, beyond ASR decoding, anonymization is extremely useful in the context of anonymized data collection for ASR training [20]. It is also known that the EER becomes lower when the attackers generate anonymized training data and retrains ASV_{eval} on this data [23]. In order to assess these aspects, we will ask volunteer participants to share additional data with us and run additional experiments in a post-evaluation phase.

7. Acknowledgment

VoicePrivacy was born at the crossroads of projects VoicePersonae, COMPRISE (`https://www.compriseh2020.eu/`), and DEEP-PRIVACY. Project HARPOCRATES was designed specifically to support it. The authors acknowledge support by ANR, JST, and the European Union's Horizon 2020 Research and Innovation Program, and they would like to thank Md Sahidullah and Fuming Fang. Experiments presented in this paper were partially carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

8. References

[1] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding," in *Interspeech*, 2019, pp. 3695–3699.

[2] A. Nautsch, A. Jimenez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado *et al.*, "Preserving privacy in speaker and speech characterisation," *Computer Speech and Language*, vol. 58, pp. 441–480, 2019.

[3] COMPRISE, "Deliverable N°5.1: Data protection and GDPR requirements." [Online]. Available: https://www.compriseh2020.eu/files/2019/06/D5.1.pdf

[4] A. Cohen-Hadria, M. Cartwright, B. McFee, and J. P. Bello, "Voice anonymization in urban sound recordings," in *2019 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019, pp. 1–6.

[5] F. Gontier, M. Lagrange, C. Lavandier, and J.-F. Petiot, "Privacy aware acoustic scene synthesis using deep spectral feature inversion," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 886–890.

[6] M. A. Pathak, B. Raj, S. D. Rane, and P. Smaragdis, "Privacy-preserving speech processing: cryptographic and string-matching frameworks show promise," *IEEE Signal Processing Magazine*, vol. 30, no. 2, pp. 62–74, 2013.

[7] P. Smaragdis and M. Shashanka, "A framework for secure speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1404–1413, 2007.

[8] S.-X. Zhang, Y. Gong, and D. Yu, "Encrypted speech recognition using deep polynomial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5691–5695.

[9] F. Brasser, T. Frassetto, K. Riedhammer, A.-R. Sadeghi, T. Schneider, and C. Weinert, "VoiceGuard: Secure and private speech processing," in *Interspeech*, 2018, pp. 1303–1307.

[10] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6341–6345.

[11] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients — how easy is it to break privacy in federated learning?" *arXiv preprint arXiv:2003.14053*, 2020.

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

[13] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X.-Y. Li, Y. Wang, and Y. Deng, "Voicemask: Anonymize and sanitize voice input on mobile devices," *arXiv preprint arXiv:1711.11460*, 2017.

[14] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification via voice transformation," in *ASRU*, 2009.

[15] M. Pobar and I. Ipšić, "Online speaker de-identification using voice transformation," in *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2014, pp. 1264–1267.

[16] F. Bahmaninezhad, C. Zhang, and J. H. Hansen, "Convolutional neural network based speaker de-identification." in *Odyssey*, 2018, pp. 255–260.

[17] C. Magariños, P. Lopez-Otero, L. Docio-Fernandez, E. Rodriguez-Banga, D. Erro *et al.*, "Reversible speaker de-identification using pre-trained transformation functions," *Computer Speech & Language*, vol. 46, pp. 36–52, 2017.

[18] 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," in *Speech Synthesis Workshop*, 2019, pp. 155–160.

[19] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa, "Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release," *arXiv preprint arXiv:2004.07442*, 2020.

[20] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-preserving adversarial representation learning in ASR: Reality or illusion?" in *Interspeech*, 2019, pp. 3700–3704.

[21] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans *et al.*, "The VoicePrivacy 2020 Challenge evaluation plan," 2020. [Online]. Available: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf

[22] J. Qian, F. Han, J. Hou, C. Zhang, Y. Wang, and X.-Y. Li, "Towards privacy-preserving speech data publishing," in *2018 IEEE Conference on Computer Communications (INFOCOM)*, 2018, pp. 1079–1087.

[23] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2802–2806.

[24] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Interspeech*, 2017, pp. 2616–2620.

[25] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, 2018, pp. 1086–1090.

[26] 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.

[27] 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, pp. 1526–1530.

[28] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019. [Online]. Available: https://datashare.is.ed.ac.uk/handle/10283/3443

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel *et al.*, "The Kaldi speech recognition toolkit," 2011.

[30] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[31] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[32] D. Ramos and J. Gonzalez-Rodriguez, "Cross-entropy analysis of the information in forensic speaker recognition," in *Odyssey*, 2008.

[33] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi *et al.*, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.

[34] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Interspeech*, 2015, pp. 3214–3218.

[35] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods," in *Odyssey*, 2018, pp. 195–202.

[36] X. Wang and J. Yamagishi, "Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis," in *Speech Synthesis Workshop*, 2019, pp. 1–6.

[37] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design choices for x-vector based speaker anonymization," in *Interspeech*, 2020.

[38] J. Patino, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," Eurecom, Tech. Rep. EURECOM+6190, 2020. [Online]. Available: http://www.eurecom.fr/publication/6190