The following publication K. A. Lee, Z. Liu, L. Chen and Z. -H. Ling, "Pinhole Effect on Linkability and Dispersion in Speaker Anonymization," in IEEE Signal Processing Letters, vol. 32, pp. 4144-4148, 2025 is available at https://doi.org/10.1109/lsp.2025.3624588.

Pinhole Effect on Linkability and Dispersion in Speaker Anonymization

Kong Aik Lee, Zeyan Liu, Liping Chen, and Zhen-Hua Ling

***Abstract*—Speaker anonymization aims to conceal speaker-specific attributes in speech signals, making the anonymized speech unlinkable to the original speaker identity. Recent approaches achieve this by disentangling speech into content and speaker components, replacing the latter with pseudo- speakers. The anonymized speech can be mapped either to a common pseudo-speaker shared across instances or to distinct pseudo-speakers unique to each instance. This paper investigates the impact of these mapping strategies on three key dimensions: speaker linkability, dispersion in the anonymized speaker space, and de-identification from the original identity. Our findings show that using distinct pseudo-speakers increases speaker dispersion and reduces linkability compared to common pseudo-speaker mapping, while maintaining de-identification, thereby enhancing overall privacy preservation. These observations are interpreted through the proposed *pinhole effect*, a conceptual framework introduced to explain the relationship between mapping strategies and anonymization performance. The hypothesis is validated through empirical evaluation.**

***Index Terms*—Privacy-preserving speech processing, voice privacy preservation, speaker anonymization.**

I. Introduction

SPEAKER anonymization is the task of altering the speaker's voice to hide their identity to the greatest possible extent while leaving all other speech attributes intact [1]–[3]. For instance, speech signals are anonymized to conceal the identity of the interviewee on a television program while keeping the spoken contents. In a wider context, speaker anonymization is posed as a privacy-preservation solution, alongside homomorphic encryption [4], [5] and federated learning [6]. Different from the latter, speaker anonymization transforms speech signals into a privacy-preserving format that aligns with current pipelines. This practicality has led to its widespread adoption, as anonymized speech can be used in downstream speech processing tasks (e.g., speech and emotion recognition) with minimal or no modifications to existing systems [7], while preserving speaker's privacy.

Mainstream speaker anonymization approach is based on replacing the speaker's voice attributes with those of a pseudo speaker [3], [8], [9]. In this approach, the input speech could be anonymized to a common or distinct pseudo-speaker. Both of these have their pros and cons. For instance, distinct pseudo-speakers are useful for a multi-party conversational setting when multiple speakers are involved. In contrast, a common pseudo-speaker is often used in passer-by interviews. It has been stipulated that mapping to a common pseudo-speaker would lead to more effective privacy preservation since all anonymized speech would sound alike. This paper shows that this first intuition does not hold. More specifically, we hypothesize that mapping to distinct pseudo-speakers reduces linkability (i.e., likelihood of re-identification) and thereby enhances privacy preservation. We validate this hypothesis through experiments using two different speaker anonymization systems. Furthermore, we demonstrate that this phenomenon can be explained by the *pinhole effect*, a conceptual framework proposed for the first time in this paper.

This work was supported in part by the Innovation and Technology Fund, Hong Kong SAR (MHP/048/24), the National Key R&D Program of China (2024YFE0217200), and the FRF-CU (WK2100000043).

Kong Aik Lee is with The Hong Kong Polytechnic University and the Research Centre for Data Science & Artificial Intelligence, Hong Kong (e-mail: kong-aik.lee@polyu.edu.hk). Zeyan Liu, Liping Chen, and Zhen-Hua Ling are with the University of Science and Technology of China, Hefei, China (e-mail: xy671231@mail.ustc.edu.cn; lipchen@ustc.edu.cn; zhling@ustc.edu.cn).

II. Speaker Anonymization

Speaker anonymization belongs to a subclass of privacy-preserving technology (PPT). PPT aims to protect data privacy while the data is being processed, at rest on a system, or in transit between systems [10]. Put it formally. Let X_p be the private data in X, where X_p leads to the inference of Y (e.g., identity of the person, gender, ethnicity):

$$X_p \rightarrow Y \tag{1}$$

Privacy preservation aims to keep private data X_p opaque to an insider or an outsider, when the data X is being processed, in transit, or stored.

Speaker anonymization realizes the goal of privacy preservation by removing or concealing the voice attributes $X_v \subset X_p$ that leads to the inference of speaker identity $Y_{\text{ID}} \in Y$, i.e., $X_v \rightarrow Y_{\text{ID}}$, while keeping other private and non-private information (linguistic and para-linguistic) untouched. Formally, let X be a speech signal, speaker anonymization encompasses a mapping between the input and the anonymized speech, represented by the function f, as follows:

$$f : X \longmapsto X \backslash X_v \tag{2}$$

where the shorthand $X \backslash X_v$ denotes the set X excluding the subset X_v. The resulting speech signal $X \backslash X_v$ is referred to as the anonymized speech. In practice, since a speech signal cannot exist without a speaker's voice as a carrier of the spoken words, the mapping is realized as

$$f' : X \longmapsto (X \backslash X_v) \cup X_{\text{pseu}} \tag{3}$$

where X_{pseu} represents the pseudo-speaker voice introduced to replace X_v. Pseudo-speakers are artificial speakers created by algorithms that are not linked to any real person, thereby

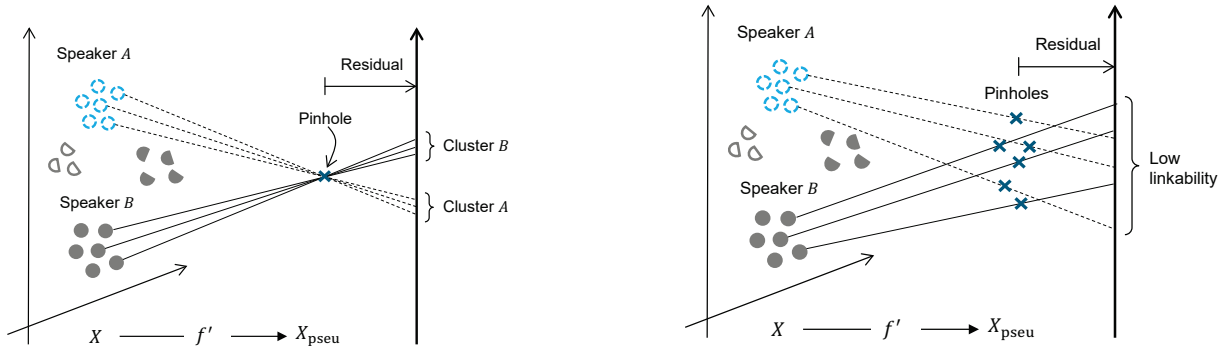


Fig. 1: Pinhole effect in speaker anonymization: (a) a common pseudo-speaker is used across all anonymization instances, resulting in any-to-one (a2o) mappings, and (b) distinct pseudo-speakers are used for each anonymized utterance, resulting in any-to-any (a2a) mappings.

avoiding infringing the privacy of real individuals. Many approaches have been proposed for the generation and selection of pseudo speakers [3], [9], [11]–[15], for example, by taking the average of a subset of speakers selected from a cohort set with the further distance away from the source speaker [16].

III. Pinhole Effect in Speaker Anonymization

In the anonymization process, an important consideration is whether to use the same pseudo-speaker across different anonymization instances. One approach is to apply a common pseudo-speaker to all anonymized utterances, while another is to assign a distinct pseudo-speaker to each session. The former leads to an any-to-one mapping, while the latter leads to an any-to-any mapping.

In the case when a common pseudo-speaker $X^i_{\text{pseu}} = X_{\text{pseu}}$ is used for all i utterances, the anonymization function f' in (3) becomes an any-to-one mapping. On the other hand, the function f' implements an any-to-any mapping when $X^i_{\text{pseu}} \neq X^j_{\text{pseu}}$ for $i \neq j$. In Figs. 1 (a) and (b), we illustrate the any-to-one and any-to-any mappings graphically assuming a two-dimensional feature space. Due to the imperfection in the removal of speaker voice attributes, residual attributes cause the anonymized utterances from the same speaker to cluster together. In Fig. 1(a), this is illustrated as beams of light passing through a pinhole, where beams originating from the same source (i.e., in analogy to speech utterances from the same speaker) passing through the pinhole will cluster around in the same area. In Fig. 1(b), by using multiple pinholes (i.e., analogous to multiple pseudo-speakers), the anonymized utterances from the same speakers are scattered apart in the anonymized space. In this case, the pseudo-speaker attributes overwhelm the residual.

The *pinhole effect* and its implications for speaker anonymization can be summarized as follows:

- **Dispersion:** Any-to-any mapping leads to greater dispersion in speaker representations of anonymized speech compared to any-to-one mapping.
- **Linkability:** Any-to-any mapping reduces speaker similarity among anonymized utterances, thereby lowering linkability relative to any-to-one mapping.
- **De-identification:** The speaker similarity between original and anonymized speech does not differ significantly regardless of the number of pinholes (one or multiple). Thus, both any-to-one and any-to-any mappings achieve a comparable level of de-identification.

The implication for *speaker linkability* arises directly from the assertion on *speaker dispersion*, as greater dispersion leads to reduced linkablity between anonymized utterances. Specifically, low linkability is achieved by assigning one distinct pseudo-speaker per session (e.g., meeting), preventing anonymized utterances from being linked across sessions. Both of these relate to the distribution of anonymized speaker representations on the right-hand side of the pinhole(s) in Figs. 1 (a) and (b). In contrast, the assertion concerning speaker *de-identification* involves a cross-side comparison, evaluating the speaker similarity between original speech (left side) and anonymized speech (right side) across the pinhole(s).

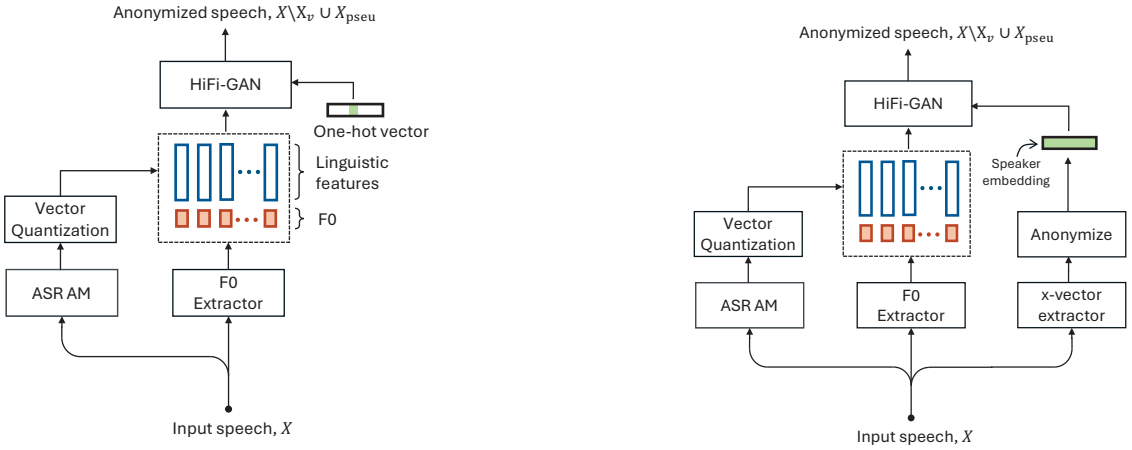
IV. Experiment

We validate the pinhole effect using two speaker anonymization systems developed for the VoicePrivacy 2024 (VPC2024) Challenge [17] under two anonymization settings. In the first, all utterances are anonymized to a common pseudo-speaker, corresponding to the any-to-one mapping strategy. In the second, each utterance is anonymized using a different pseudo-speaker, corresponding to the any-to-any mapping strategy.

A. Models

We used two speaker anonymization systems in the experiments. The first system (`SYS1`) corresponds to the B5 baseline in VPC2024 [1]. As depicted in Fig. 2(a), the system consists of (i) an ASR acoustic model (ASR AM) to extract speech features containing the linguistic content, and (ii) a pitch tracking model to extract $F0$ features. Vector quantization (VQ) is applied to the linguistic features, while per-utterance mean-variance normalization (MVN) is applied to the $F0$ features; both aim to reduce residual voice attributes. The quantized

[1] https://github.com/deep-privacy/SA-toolkit



(a) SYS1: One-hot speaker representation. (b) SYS2: Continuous speaker embedding.

Fig. 2: Two speaker anonymization systems used in the experiments to validate the *pinhole effect*.

features are then fed together with the $F0$ prosodic feature into a HiFi-GAN neural source-filter model [18] to synthesize an anonymized speech. SYS1 can be configured for either any-to-one or any-to-any mapping by selecting an appropriate one-hot speaker vector. For the former, a constant one-hot speaker ID is used for all anonymized utterances to produce a common pseudo-speaker. For the latter, a different one-hot speaker ID is randomly assigned to each anonymized utterance. In the any-to-one setting, speaker ID 103 was arbitrarily chosen from a pool of 251 speakers for the experiments.

The second system (SYS2) is a variant of the B5 baseline, where the one-hot vector input to the HiFi-GAN vocoder is replaced with a speaker embedding. As shown in Fig. 2(b), the pseudo-speaker embedding ϕ_{pseu} is fed to a vocoder with the other two components (i.e., the $F0$ and linguistic features) to synthesize the anonymized speech. SYS2 can also be configured for either any-to-one or any-to-any mapping by manipulating the speaker embedding. In the experiments, utterances are anonymized to a common pseudo-speaker, obtained by averaging the x-vector embeddings from the *LibriSpeech train-clean-100* [19]. For any-to-any mapping, each utterance is anonymized using a distinct pseudo-speaker embedding, computed by averaging the x-vector embeddings of 100 utterances randomly selected from *LibriSpeech train-clean-100*.

B. Datasets

The configurations and datasets used to train each module of the anonymization systems are described as follow. The ASR AM comprises a wav2vec2 [20] front-end with three additional TDNN-F layers [21]. As in the VPC2024 B5 baseline, the wav2vec2 model was pre-trained on *VoxPopuli* dataset [22] and then fine-tuned on *LibriSpeech train-clean-100* [19]. The VQ has a codebook size of 48, with a dimensionality of 256. The x-vector extractor [23] was trained on the *VoxCelb-1* and *VoxCeleb-2* datasets [24]. As in most implementations, the x-vector has a dimensionality of 512. The HiFi-GAN vocoder was trained on the *LibriSpeech train-clean-100* with both ASR-AM and x-vector extractor frozen. A Pytorch implementation of YAAPT pitch tracking algorithm [25] was used for the $F0$ extraction.

TABLE I: Evaluation of anonymization methods on the LibriSpeech dataset using ASV EER (where higher values indicate better anonymization) and ASR WER (where lower values indicate better linguistic preservation).

	Test set	Partition	ORG	SYS1	SYS2
EER (%)	libri-dev	F	10.51	33.37	34.94
		M	0.93	31.94	34.32
	libri-test	F	8.76	31.84	33.73
		M	0.42	32.19	32.74
	avg		5.16	32.23	33.93
WER (%)	libri-dev	-	1.79	3.95	3.88
	libri-test	-	1.84	4.15	4.01

C. Performance metrics

Our experiments were carried out following the evaluation protocol provided in VPC2024 [17]. The privacy-preserving capability of the anonymization systems was evaluated using ASV tests, with performance measured by the *equal error rate* (EER). To minimize inference risk, anonymized speech should not be successfully verified by an ASV system, which corresponds to a higher EER. The ability to preserve linguistic content was assessed using ASR tests, measured by the *word error rate* (WER). Since the goal is to retain speech information other than the speaker's identity, the anonymized speech should have a WER as close as possible to that of the original speech.

Table I shows the baseline performance of the two anonymization systems, SYS1 and SYS2, using the any-to-one mapping to a common voice. The ASV evaluations were conducted in a a gender-dependent manner. The average EERs across the four subsets are included. Comparing to the EERs in the ORG column, when no anonymization was applied, both anonymization systems increase tremendously the EER to over 30%. Both anonymization systems obtained WERs comparable with the original speech (ORG) without anonymization. These results show that the anonymization

TABLE II: Scatter values and ratio J for original and anonymized speech under different systems and mapping strategies.

Method	Mapping	Tr($W^\top S_w W$)	Tr($W^\top S_b W$)	J
ORG	–	206.71	305.39	1.477
SYS1	a2o	674.27	30.14	0.047
SYS1	a2a	1224.04	38.19	0.031
SYS2	a2o	730.91	31.83	0.045
SYS2	a2a	2192.49	48.95	0.023

TABLE III: ASV EER (%) on the VPC2024 LibriSpeech Dev and Test subsets under any-to-one (a2o) and any-to-any (a2a) mapping strategies.

Method	Mapping	Libri-Dev		Libri-Test		Avg
		F	M	F	M	
SYS1	a2o	33.37	31.94	31.84	32.19	32.23
SYS1	a2a	34.88	36.21	33.12	32.43	34.16
SYS2	a2o	34.94	34.32	33.73	32.74	33.93
SYS2	a2a	37.03	35.84	34.37	36.62	35.97

systems achieve good privacy preservation while retaining the linguistic information for downstream tasks.

D. Results

The first experiment examines the speaker dispersion of anonymized speech [2]. Using the x-vector extractor described in Section-IV(A), a total of $28,539$ utterances from 251 speakers in the *LibriSpeech train-clean-100* dataset were represented as x-vector embeddings. The within-class and between-class scatter matrices, denoted as S_w and S_b respectively, were computed using the speaker labels. The trace of the within-class scatter matrix, Tr($W^\top S_w W$), quantifies the speaker class compactness. A smaller value indicates tighter clustering. In contrast, the trace of the between-class scatter matrix, Tr($W^\top S_b W$), measures the separation between speakers, where a higher value indicates better speaker separation. Here, W is the matrix of eigenvectors of $S_w^{-1} S_b$. The within-class and between-class scatter values for the original utterances (ORG) are reported in the first row of Table II. The last column shows the scatter ratio $J = \text{Tr}((W^\top S_w W)^{-1}(W^\top S_b W))$. While the between-class scatter reduces, the within-class scatter increases substantially after anonymization. The resulting reduction in the scatter ratio J indicates increased speaker dispersion and, consequently, lower linkability among anonymized utterances. Comparing the mapping strategies, any-to-any (a2a) mapping yields a lower scatter ratio and thus higher dispersion than any-to-one (a2o) mapping in both systems.

In the second experiment, we examine the identity *linkability* property asserted by the pinhole effect. We assume that an attacker attempts to verify the speaker identity using anonymized speech utterances. In this case, anonymized utterances were used for enrollment and test in ASV. We experimented with two settings described in Section IV-A. In the first setting, utterances were anonymized using a common pseudo-speaker leading to an a2o mapping as in Fig 1(a). In the second setting, utterances were anonymized with distinct pseudo-speakers leading to an a2a mapping as in Fig 1(b). Table III shows the ASV EER on the VPC2024 LibriSpeech Dev and Test subsets, extending Table I with results for the a2a setting. Looking at the second and third row for anonymization system SYS1, a2a mapping leads to higher EER compared to a2o mapping. The EER increment amount to 5.35% on average. Bootstrap resampling was performed to compute the 95% confidence intervals of the EER differences between a2a and a2o, confirming that the improvements achieved by a2a are statistically significant ($p < 0.05$). Similar trend is observed for anonymization system SYS2 in the last two rows. These results support the assertion that a2a mapping reduces linkability between anonymized utterances, due to greater dispersion across anonymized samples.

[2]https://github.com/VoicePrivacy/Pinhole-effect-in-anonymization

TABLE IV: ASV EER (%) when original speech is used for enrollment and anonymized speech for testing, measuring de-identification.

Method	Mapping	Libri-Dev		Libri-Test		Avg
		F	M	F	M	
SYS1	a2o	47.87	49.38	50.34	48.80	49.10
SYS1	a2a	47.58	48.27	48.72	51.00	48.89
SYS2	a2o	48.72	48.27	47.81	49.00	48.45
SYS2	a2a	49.01	47.98	49.26	48.60	48.71

The third experiment examines the *de-identification* property asserted by the pinhole effect. With reference to Fig. 1 (a) and (b), this setting corresponds to the comparison of the original speech on the left to the anonymized speech on the right of the pinhole(s). In this scenario, we assume that an attacker attempts to verify an anonymized speech utterance as if it was spoken by the same speaker given an original speech utterance without anonymization. Table IV shows the ASV EER. The EERs are relatively higher compared to the EERs in Table III, since the enrollment utterances were not anonymized. Comparing the EERs for any-to-one and any-to-any mappings, there are no substantial differences between using a common or distinct pseudo-speaker for anonymization. This observation supports the second assertion of the pinhole effect, indicating that both mapping strategies achieve comparable de-identification performance.

V. Conclusion

We have introduced the *pinhole effect* as a conceptual framework to explain the identity linkability behavior frequently observed in speaker anonymization systems. By modeling the anonymization process as a mapping function from original speaker identities to pseudo-speakers, we examined two key strategies: mapping to a common pseudo speaker (any-to-one) and mapping to distinct pseudo speakers (any-to-any). Our analysis shows that anonymizing each utterance to a distinct pseudo-speaker significantly reduces speaker linkability by increasing the dispersion in the anonymized speaker space. While both mapping strategies achieve comparable de-identification performance (i.e, the anonymized speech cannot be reliably traced back to the original speaker), the use of distinct pseudo-speakers offers a clear advantage in lowering linkability, which is a desirable property in privacy-preserving speech processing.

REFERENCES

[1] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche, "The VoicePrivacy 2020 Challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, Jul. 2022.

[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] L. Chen, W. Gu, K. A. Lee, W. Guo, and Z.-H. Ling, "Pseudo-speaker distribution learning in voice anonymization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 272–285, 2025.

[4] 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, Mar. 2013.

[5] S.-X. Zhang, Y. Gong, and D. Yu, "Encrypted speech recognition using deep polynomial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5691–5695.

[6] D. Leroy, A. Coucke, T. Lavril, T. Gisselbrecht, and J. Dureau, "Federated learning for keyword spotting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6341–6345.

[7] S. T. Arasteh, T. Arias-Vergara, P. A. Pérez-Toro, T. Weise, K. Packhäuser, M. Schuster, E. Nöth, A. Maier, and S. H. Yang, "Addressing challenges in speaker anonymization to maintain utility while ensuring privacy of pathological speech," *Communications Medicine*, vol. 4, no. 1, p. 182, 2024.

[8] F. Fang, X. Wang, J. Yamagishi, M. T. I. Echizen, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Proc. Speech Synthesis Workshop*, 2019, pp. 155–160.

[9] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi, "Privacy and utility of x-vector based speaker anonymization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2383–2395, Jul. 2022.

[10] D. W. Archer, B. de Balle Pigem, D. Bogdanov, M. Craddock, A. Gascon, R. Jansen, M. Jug, K. Laine, R. McLellan, O. Ohrimenko, M. Raykova, A. Trask, and S. Wardley, "UN handbook on privacy-preserving computation techniques," 2023. [Online]. Available: https://arxiv.org/abs/2301.06167

[11] P. Champion, D. Jouvet, and L. Anthony, "Are disentangled representations all you need to build speaker anonymization systems?" in *Proc. Interspeech*, 2022, pp. 2793–2797.

[12] 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*, p. 3681–3695, 2023.

[13] J. Yao, Q. Wang, P. Guo, Z. Ning, and L. Xie, "Distinctive and natural speaker anonymization via singular value transformation-assisted matrix," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2944–2956, 2024.

[14] H. Turner, G. Lovisotto, and I. Martinovic, "Generating identities with mixture models for speaker anonymization," *Computer Speech & Language*, vol. 72, p. 101318, 2022.

[15] S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, "Anonymizing speech with generative adversarial networks to preserve speaker privacy," in *Proc. SLT*, 2023, pp. 912–919.

[16] 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 *Proc. Interspeech*, 2020, pp. 1713–1717.

[17] 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," *arXiv preprint arXiv:2404.02677*, 2024.

[18] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*. IEEE, 2015, pp. 5206–5210.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[21] 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.

[22] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 993–1003.

[23] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2017, pp. 999–1003.

[24] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[25] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2002, pp. 361–364.