A Preliminary Study on Sectional Voice Anonymization and Detection

Shaoqi Tang*, Zeyan Liu*, Liping Chen*, Kong Aik Lee†, Tomoki Toda‡, Zhenhua Ling*
* University of Science and Technology of China, China
Emails: tsqslh97877256tk@gmail.com, xy671231@mail.ustc.edu.cn, {lipchen, zhling} @ustc.edu.cn,
† The Hong Kong Polytechnic University, China
Email: kong-aik.lee@polyu.edu.hk
‡ Nagoya University, Japan
Email: tomoki@icts.nagoya-u.ac.jp

***Abstract*—Facilitated by the speech generation framework that disentangles the speaker attribute from the original speech and represents it as an embedding vector, voice anonymization is accomplished by substituting the original speaker embedding vector with that of a pseudo-speaker. To date, voice anonymization techniques are required to anonymize a speech utterance to a single pseudo-speaker, leaving the methodology for anonymizing an utterance into multiple pseudo-speakers unexplored. To this end, this paper investigates a sectional voice anonymization method, which partitions the original utterance into sections and replaces the original speaker with distinct pseudo-speakers in each section. Subsequently, to align with the single pseudo-speaker requirement defined for anonymized speech, a detection algorithm is investigated to examine the presence of multiple pseudo-speakers in anonymized speech. Experimental results on the LibriSpeech dataset demonstrated that the sectional anonymization method was able to enhance voice privacy protection capability, and the presence of multiple pseudo-speakers in anonymized utterances can be detected. Audio samples can be found in https://voiceprivacy.github.io/sectional-voice-anonymization-and-detection/.**

I. Introduction

Recent advancements in speech technologies [1]–[4] have led to increasing threats to the voice privacy information conveyed by speech signals. For example, personal information of the speaker, such as identity, age, gender, and emotional state, may be compromised. Moreover, high-quality artificial speech can be synthesized in a person's voice for impersonation. Such threats to voice privacy highlight the necessity for voice privacy protection techniques [5]–[7]. Driven by three VoicePrivacy Challenges (VPCs) [8]–[10], voice anonymization techniques have been remarkably prompted and emerged as one of the mainstream approaches for voice privacy protection.

Facilitated by the speech generation technique based on information disentanglement, voice anonymization can be achieved by substituting the original speaker attribute with that of a pseudo-speaker. Particularly, the speaker attribute is disentangled from the original speech and represented with an embedding vector. A pseudo-speaker embedding is created to substitute the original one and used to generate the anonymized speech. Within such a framework, pseudo-speaker embedding creation poses a fundamental challenge, emphasizing the need to distinguish the pseudo-speaker from the original speaker while ensuring the uniqueness of the pseudo-speaker [5]. Research on pseudo-speaker construction has led to the development of methods based on reference pools [11]–[13], transformations of original speaker embeddings [14], [15], and generative models [16], [17], respectively.

The current definition of the voice anonymization task confines that a single pseudo-speaker is present in the anonymized speech, achieved by replacing the speaker attribute of the entire original utterance with that of the pseudo-speaker. As a result, the method for anonymizing a speech utterance into multiple pseudo-speakers remains uninvestigated. To this end, this paper presents a preliminary study on voice anonymization using multiple pseudo-speakers and its detection. Specifically, a sectional voice anonymization method is proposed in which the original utterance is anonymized at the section level, with each section being assigned a distinct pseudo-speaker. Based on that, to align with the single pseudo-speaker requirement defined for anonymized speech, a detection method is investigated to examine the presence of multiple speakers in the anonymized speech. Experiments conducted on the LibriSpeech dataset demonstrated that:

1) The application of multiple pseudo-speakers was able to enhance the efficacy of voice privacy protection compared to a single pseudo-speaker, probably due to the increased complexity of the pseudo-speaker in the anonymized speech.

2) The presence of multiple pseudo-speakers in the anonymized speech can be effectively detected.

II. Background

Fig. 1 presents the voice anonymization framework based on the information disentanglement mechanism [13], which is applied in our study. Given an original speech $\mathcal{O}$, the

Corresponding author: Liping Chen.

This work was supported in part by the National Key Research and Development Program of China (Project No. 2024YFE0217200) and the Innovation and Technology Fund of the Hong Kong SAR (Project No. MHP/048/24)

This work was supported in part by the National Key Research and Development Program Project 2024YFE0217200, the Innovation and Technology Fund of the Hong Kong SAR MHP/048/24, the National Natural Science Foundation of China under Grant U23B2053, and the Fundamental Research Funds for the Central Universities WK2100000043.

979-8-3315-7206-8/25/$31.00

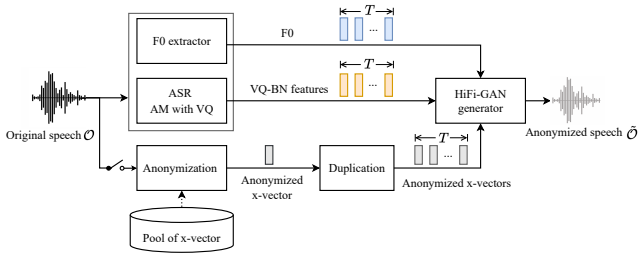


Fig. 1: Illustration of voice anonymization framework. The inputs of the original speech and the reference pool of x-vector to the anonymization module are optional, depending on the anonymization algorithm, as indicated by the switch and dotted line, respectively.

F0 features are extracted with an F0 extractor. The VQ-BN feature vectors are derived as bottleneck (BN) features extracted with an acoustic model (AM) used in automatic speech recognition (ASR), processed with vector quantization (VQ). Assume that the number of frames in $\mathcal{O}$ is T. Both the F0 and VQ-BN features are of length T, and serve as representations of prosodic and linguistic content attributes within $\mathcal{O}$, respectively. Besides, an x-vector is extracted from $\mathcal{O}$ with an x-vector extractor, representing the speaker attribute. Thereafter, an anonymization module is applied to generate the anonymized x-vector, representing the pseudo-speaker. It is then duplicated T times to match the length of the F0 and VQ-BN features. Finally, the F0, VQ-BN, and anonymized x-vector sequences are input into the HiFi-GAN generator to generate the anonymized speech $\tilde{\mathcal{O}}$.

Specifically, as depicted in Fig. 1, the original speech $\mathcal{O}$ and the reference pool of x-vector are optional inputs to the anonymization module, depending on the specific anonymization algorithm [11]–[17]. In our study, the random selection strategy is applied in the anonymization module, which excludes $\mathcal{O}$ as input, wherein the anonymized x-vector is obtained by randomly sampling from the reference x-vector pool.

III. Sectional voice anonymization

In sectional voice anonymization, the original utterance is divided into sections, with a specific pseudo-speaker assigned to each section to generate the anonymized speech. The process to obtain the anonymized x-vector sequence is shown in Fig. 2. The anonymization module, which takes the original speech as input, is illustrated in Fig. 2(a). Given the original speech utterance $\mathcal{O}$, it is first segmented into K sections, represented as $\{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_K\}$. For the k-th section $\mathcal{S}_k$ $(k = 1, ..., K)$, an anonymized x-vector $\boldsymbol{x}_k$ is randomly selected from the reference pool, yielding K anonymized x-vectors $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_2, ..., \boldsymbol{x}_K\}$, which are distinct from each other.

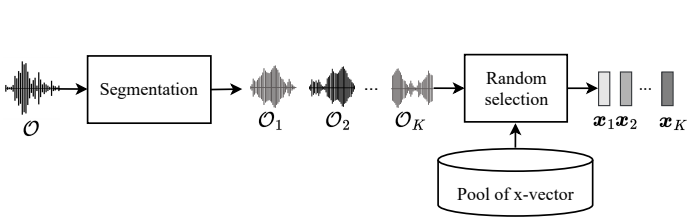
Particularly, the voice activity detection (VAD) algorithm is applied to partition $\mathcal{O}$ into segments, with cutting points occurring in silence intervals. Given N segments obtained via VAD, denoted as $\{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_N\}$, and a minimum duration

Input: VAD segments $\{\boldsymbol{S}_1, \boldsymbol{S}_2, ..., \boldsymbol{S}_N\}$, section duration threshold $D_{\min}$.
Output: Sections $\mathcal{Q} = \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_K\}$.
1. Initialize section set: $\mathcal{Q} = \{\ \}$;
2. Initialize intermediate section: $\boldsymbol{Q} = [\ ]$;
3. Initialize index $n = 1$;
4. **while** *True* **do**
  if $len\,(\boldsymbol{Q}) < D_{\min}$ **then**
    // Append segments to create a section until its length exceeds the threshold $D_{\min}$. Here, the function $len\,(\bullet)$ returns the duration of non-silent speech.
    $\boldsymbol{Q} = [\boldsymbol{Q}, \boldsymbol{S}_n]$;
    $n = n + 1$;
  end
  else
    if *n==N-1* **then**
      // Check the last segment
      if $len\,(\boldsymbol{S}_{n+1}) < D_{\min}$ **then**
        // Append the last segment to the current section
        $\boldsymbol{Q} = [\boldsymbol{Q}, \boldsymbol{S}_{n+1}]$;
        $n = n + 1$;
      end
    end
    // Save the current section
    $\mathcal{Q} = \{\mathcal{Q}, \boldsymbol{Q}\}$;
    $\boldsymbol{Q} = [\ ]$;
  end
  if $n > N$ **then**
    return $\mathcal{Q}$;
  end
end

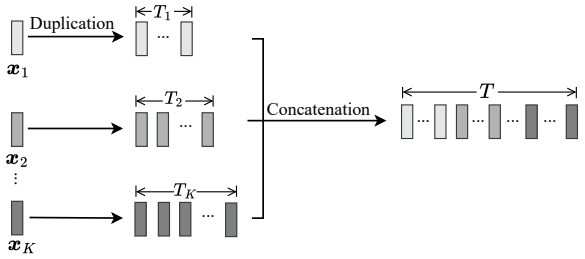
Algorithm 1: Algorithm for deriving sections from VAD segments.

$D_{\min}$ for sections, the sections $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K\}$ are derived to ensure that the duration of the non-silent speech in each section is at least $D_{\min}$. The algorithm for deriving sections from VAD segments is detailed in Alg. 1.

Assume the length of the k-th section to be T_k $(k = 1, \ldots, K)$, such that $\sum_{k=1}^{K} T_k = T$, where T denotes the total number of frames in the original speech. As depicted in Fig. 2(b), $\boldsymbol{x}_k$ is duplicated T_k times, resulting in a subsequence of anonymized x-vectors for the k-th section. Thereafter, the K subsequences are concatenated to form the anonymized x-vector sequence. Finally, combined with the F0 and VQ-BN sequences extracted from $\mathcal{O}$, it is fed into the HiFi-GAN generator to generate the anonymized speech $\tilde{\mathcal{O}}$. Thereby, multiple pseudo-speakers are present in $\tilde{\mathcal{O}}$, with one pseudo-speaker assigned to each section.



(a) Illustration of the anonymization module for sectional voice anonymization.



(b) Duplication of the anonymized x-vector.

Fig. 2: Anonymized x-vector sequence generation in sectional voice anonymization.

IV. Detection

The detection of multiple pseudo-speakers in the anonymized speech $\tilde{\mathcal{O}}$ is performed on the sections. Given $\tilde{\mathcal{O}}$, it is first partitioned into sections, represented as $\{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_L\}$, with L denoting the number of sections. In our work, three segmentation techniques are examined as detailed in the experiments, including voice activity detection, speaker change detection [18], and speaker diarization [19]. Among these three methods, the speaker purity of the sections increased progressively from voice activity detection to speaker diarization. Thereafter, the speaker similarity is scored on the sections, quantifying the likelihood of the utterance being anonymized to multiple pseudo-speakers.

In this study, given the sections $\{\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_L\}$, a speaker embedding vector is extracted from each, resulting in $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_L\}$. Two types of scores are examined, calculated on the speaker embedding vetors of successive sections $\{\boldsymbol{x}_l, \boldsymbol{x}_{l+1}\}$ $(l = 1, \ldots, L-1)$ and exhaustive pairs of sections $\{\boldsymbol{x}_l, \boldsymbol{x}_{j>l}\}$ $(l = 1, \ldots, L-1;\ j = 2, \ldots, L)$, respectively detailed as follows.

$$s_{\text{suc}} = \frac{1}{L-1} \sum_{l=1}^{L-1} \frac{\boldsymbol{x}_l \boldsymbol{x}_{l+1}}{\|\boldsymbol{x}_l\|_2 \, \|\boldsymbol{x}_{l+1}\|_2} \tag{1}$$

$$s_{\text{exh}} = \frac{2}{L\,(L-1)} \sum_{l=1}^{L-1} \sum_{j=l+1}^{L} \frac{\boldsymbol{x}_l \cdot \boldsymbol{x}_j}{\|\boldsymbol{x}_l\|_2 \, \|\boldsymbol{x}_j\|_2} \tag{2}$$

where $\| \bullet \|_2$ is the L2 norm. The subscripts $_{\text{suc}}$ and $_{\text{exh}}$ are short for *successive* and *exhaustive*, respectively. Finally, the score is compared with a threshold to decide whether multiple pseudo-speakers are present in the anonymized speech.

V. Experiments

A. Datasets & configurations

In our experiments, the anonymization framework was built following the open-source toolkit published in [1]. The ASR AM with VQ model applied in the open-source toolkit was used in our experiments, which was built upon a wav2vec2 model pre-trained on VoxPopuli [20], and fine-tuned on the LibriSpeech [21] train-clean-100 dataset. An open-source pre-trained x-vector encoder [22] was employed as the speaker encoder, available at [2]. The HiFi-GAN generator was trained using the train-clean-100 subset from the LibriSpeech dataset. The reference pool adopted the train-clean-100 subset of Librispeech. Evaluations were conducted on the development and test subsets, denoted as libri-dev and libri-test, respectively, following the configurations provided by VoicePrivacy Challenge 2024 (VPC2024)[3]. All recordings applied in our experiments were resampled to 16 kHz.

B. Anonymization evaluations

1) Evaluation metrics: In our experiments, automatic speaker verification (ASV) evaluations were conducted to examine the voice privacy protection capability. In these evaluations, the ASV models were trained with the train-clean-360 subset of LibriSpeech using its anonymized version. The performances were measured with equal error rates (EERs). Automatic speech recognition (ASR) evaluations were conducted to measure the preservation of linguistic content. As provided by VPC2024, an ASR model based on the wav2vec 2.0 architecture was used. The performances were measured with word error rates (WERs).

2) Compared methods: Two anonymization methods were compared including:

- *Baseline*: As depicted in Fig. 1, given an original speech utterance, a randomly selected anonymized x-vector from the reference pool was used to generate the anonymized speech.
- *Sectional voice anonymization*: As depicted in Fig. 2, given an original speech utterance, it was first partitioned into sections. A distinct anonymized x-vector was randomly selected for each section to generate the anonymized speech. The WebRTC[4] toolkit was applied for VAD segmentation, and the minimum duration $D_{\min}$ was set to 1 second. On average, each sentence was segmented into 1.97 sections, with a mean duration of 4.90 seconds per section in our evaluation dataset.

3) Results: The EERs obtained on the libri-dev and libri-test sets are presented in Table I. As seen from the table, the sectional anonymization method achieved substantially

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

[2] http://kaldi-asr.org/models/m7

[3] https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024

[4] https://github.com/wiseman/py-webrtcvad

TABLE I: EERs(%) obtained by the baseline and sectional voice anonymization (SVA) methods on the libri-dev and libri-test datasets. The results are presented for female and male genders, respectively.

	Gender	Baseline	SVA
libri-dev	Female	34.80	42.47
	Male	33.10	38.20
libri-test	Female	33.58	38.36
	Male	32.74	39.66

TABLE II: WERs (%) obtained on the libri-dev and libri-test datasets by the baseline and sectional voice anonymization (SVA) methods.

	Baseline	SVA
libri-dev	4.62	4.69
libri-test	4.38	4.47

higher EERs than the baseline method on the four subsets. This demonstrates that anonymizing a speech utterance to multiple pseudo-speakers is able to enhance the voice privacy protection capability compared to a single pseudo-speaker. The superiority should be attributed to the increased complexity of the pseudo-speaker in the anonymized speech.

The WERs obtained on the libri-dev and libri-test datasets are presented in Table II. Between the two anonymization methods, the results show that utterances anonymized using the sectional anonymization method exhibit slightly higher WERs than those anonymized by the baseline method. However, the relative increase is within 2%, indicating that the degradation in speech quality caused by the sectional anonymization method is marginal.

C. Detection evaluations

1) Segmentation methods: In our experiments, voice activity detection, speaker change detection, and speaker diarization algorithms were examined for segmentation the anonymized speech utterances. The configurations are detailed as follows:

- *Voice activity detection (VAD)*: In our detection experiments, VAD was implemented using the open-source pyannote.audio toolkit [5]. The threshold for activity detection was set to 0.50. On average, each sentence was segmented into 3.10 sections, with a mean duration of 2.54 seconds per section in our evaluation dataset.
- *Speaker change detection (SCD)*: The open-source code[6] was used for speaker change point detection. The sensitivity parameter for peak detection was set to 0.05. A minimum section length of 0.45 seconds was applied. On average, each sentence was segmented into 3.19 sections, with a mean duration of 2.49 seconds per section.

[5]https://github.com/pyannote/pyannote-audio

[6]https://github.com/hbredin/fastpages/blob/master/_notebooks/2022-10-23-One-speaker-segmentation-model-to-rule-them-all.ipynb

TABLE III: EERs (%) obtained by computing the speaker similarity scores between successive sections, based on Eq. (1). Results on the libri-dev and libri-test datasets using the three segmentation methods are shown, including voice activity detection (VAD), speaker change detection (SCD), and speaker diarization (SD). The EERs obtained on the four subsets are averaged and presented in the row labeled *Avg* for each method.

Dataset	Gender	VAD	SCD	SD
libri-dev	Female	11.67	9.46	8.83
	Male	6.80	7.44	6.47
libri-test	Female	11.13	6.68	5.94
	Male	8.15	8.49	7.81
Avg		9.44	8.02	7.26

- *Speaker diarization (SD)*: Speaker diarization was implemented using the open-source pyannote.audio toolkit[5]. A minimum segment length of 0.45 seconds was used. On average, each sentence was segmented into 3.22 sections, with an average duration of 2.46 seconds per section.

In the SCD and SD methods, a section was obtained between two successive speaker change points.

2) Evaluation metrics & configurations: Detection performance was evaluated on utterances anonymized to both single and multiple pseudo-speakers using the libri-dev and libri-test datasets. Evaluations were conducted in a gender-dependent manner, with male and female utterances examined separately. Each subset was equally partitioned into two parts: one anonymized to a single pseudo-speaker and the other anonymized to multiple pseudo-speakers. Given an anonymized speech utterance, it was first partitioned into sections using the three segmentation methods, i.e., VAD, SCD and SD. Subsequently, the speaker similarity score was calculated using Eq. (1) and Eq. (2), respectively, for each utterance. Based on the scores from utterances anonymized to single and multiple pseudo-speakers, a threshold corresponding to the equal error rate was obtained. Utterances with scores exceeding this threshold were decided as anonymized to single pseudo-speakers, while those below the threshold were determined as anonymized to multiple pseudo-speakers. The performance of sectionally anonymized speech detection was measured using EERs, with lower EER values indicating better detection performance.

3) Results: The detection results obtained on the speaker similarity derived from the successive sections, as calculated using Eq. (1), are presented in Table III. The EERs of 7% to 10% achieved by the three segmentation methods demonstrated that utterances anonymized to multiple pseudo-speakers were effectively detected. Particularly, the SD method exhibited the best detection performance, as reflected by the lowest average EER, whereas the VAD method obtained the poorest performance, indicated by the highest average EER. This is

TABLE IV: EERs (%) obtained by computing the speaker similarity scores from exhaustive section pairs, according to Eq. (2). Results on the libri-dev and libri-test datasets using the three segmentation methods are included, including voice activity detection (VAD), speaker change detection (SCD), and speaker diarization (SD). The EERs obtained on the four subsets are averaged and presented in the row labeled *Avg* for each method.

Dataset	Gender	VAD	SCD	SD
libri-dev	Female	10.73	6.94	7.26
	Male	5.18	4.85	4.53
libri-test	Female	9.65	4.82	4.82
	Male	4.75	5.77	5.09
Avg		7.58	5.60	5.43

attributed to the degrees of speaker purity within the sections produced by the three methods. The segmentation outputs from the VAD method lacked information about speaker attributes within the segments, while the SCD method incorporated speaker attributes. Additionally, the SD method further refined the speaker segmentations based on SCD, resulting in the highest speaker purity for the SD method, followed by the SCD method.

Besides, the results obtained on the speaker similarity scores computed on the exhaustive section pairs, computed according to Eq. (2), are presented in Table IV. Similar to the observations obtained on the successive section pairs as shown in Table III, the efficacy of detecting multiple pseudo-speakers within the anonymized speech was validated by EERs ranging from 5% to 8%. Moreover, the exhaustive speaker similarity computation yielded lower EERs for the three segmentation methods compared to the successive similarity computation as given in Table III, thereby further improving detection performance for multiple pseudo-speakers.

Furthermore, in our experiments, the SCD and SD segmentations were realized with open-source toolkits, with models trained on recordings, leading to a mismatch with the evaluation data, which were synthesized speech. The SCD and SD methods exhibited suboptimal performances, as indicated by a low F1 score and high detection error rate, respectively. This necessitates future research to improve multiple pseudo-speaker detection.

VI. Conclusions

This paper investigated anonymizing a speech utterance into multiple pseudo-speakers and its detection. A sectional voice anonymization method was studied which assigned a distinct pseudo-speaker to each section of the anonymized speech, resulting in the presence of multiple pseudo-speakers in the anonymized output. Based on that, a detection algorithm was investigated to determine whether multiple speakers were present in the anonymized speech. Experimental results demonstrated that with the introduction of multiple pseudo-speakers in the anonymized speech, the voice privacy protection capability was enhanced. Moreover, the presence of multiple pseudo-speakers in the anonymized utterance could be effectively detected.

References

[1] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[2] X. Tan, J. Chen, H. Liu, *et al.*, "Naturalspeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4234–4245, 2024.

[3] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," *Proc. Interspeech*, 2020, pp. 3830–3834.

[5] N. Tomashenko, B. M. L. Srivastava, X. Wang, *et al.*, "Introducing the voiceprivacy initiative," *Proc. Interspeech*, 2020, pp. 1693–1697.

[6] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," *Proc. Interspeech*, 2021, pp. 1099–1103.

[7] S. Chen, L. Chen, J. Zhang, K. Lee, Z. Ling, and L. Dai, "Adversarial speech for voice privacy protection from personalized speech generation," *ICASSP*, 2024, pp. 11 411–11 415.

[8] N. Tomashenko, X. Wang, E. Vincent, *et al.*, "The VoicePrivacy 2020 Challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101 362, 2022.

[9] N. Tomashenko, X. Wang, X. Miao, *et al.*, "The VoicePrivacy 2022 Challenge evaluation plan," *arXiv preprint arXiv:2203.12468*, 2022.

[10] N. Tomashenko, X. Miao, P. Champion, *et al.*, "The VoicePrivacy 2024 Challenge evaluation plan," *arXiv preprint arXiv:2404.02677*, 2024.

[11] B. M. L. Srivastava, N. Tomashenko, X. Wang, *et al.*, "Design choices for x-vector based speaker anonymization," *Proc. Interspeech*, 2020, pp. 1713–1717.

[12] 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.

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

[14] 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*, 2023.

[15] 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.

[16] H. Turner, G. Lovisotto, and I. Martinovic, "Generating identities with mixture models for speaker anonymization," *Computer Speech & Language*, vol. 72, p. 101 318, 2022.

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

[18] S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," *Proc. DARPA broadcast news transcription and understanding workshop*, vol. 8, 1998, pp. 127–132.

[19] X. Anguera, S. Bozonnet, *et al.*, "Speaker diarization: A review of recent research," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[20] C. Wang, M. Riviere, A. Lee, *et al.*, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *Proc. ACL-IJCNLP*, ACL, 2021, pp. 993–1003.

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

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