

Voice Privacy Preservation with Multiple Random Orthogonal Secret Keys: Attack Resistance Analysis

Kohei Tanaka*, Hitoshi Kiya* and Sayaka Shiota*

* Tokyo Metropolitan University, Japan

Abstract—Recently, opportunities to transmit speech data to deep learning models executed in the cloud have increased. This has led to growing concerns about speech privacy, including both speaker-specific information and the linguistic content of utterances. As an approach to preserving speech privacy, a speech privacy-preserving method based on encryption using a secret key with a random orthogonal matrix has been proposed. This method enables cloud-based model inference while concealing both the speech content and the speaker identity. However, the method has limited attack resistance and is constrained in terms of the deep learning models to which the encryption can be applied. In this work, we propose a method that enhances the attack resistance of the conventional speech privacy-preserving technique by employing multiple random orthogonal matrices as secret keys. We also introduce approaches to relax the model constraints, enabling the application of our method to a broader range of deep learning models. Furthermore, we investigate the robustness of the proposed method against attacks using extended attack scenarios based on the scenarios employed in the Voice Privacy Challenge. Our experimental results confirmed that the proposed method maintains privacy protection performance for speaker concealment, even under more powerful attack scenarios not considered in prior work.

I. INTRODUCTION

Recently, deep learning-based speech-processing systems have commonly been used with smartphones, home assistants, and other edge devices. Such devices have strict computational resource constraints that make it challenging to run inference locally. To address this limitation, the speech-processing systems on these edge devices are typically implemented by offloading the computations of deep learning models to cloud servers. However, the cloud-based systems require sharing speech data with cloud servers. Speech data contains not only the utterance content but also personal identifying information, including language, age, gender, and speaker identity [1], [2]. The personal information embedded in speech data raises significant global privacy concerns. Therefore, under the General Data Protection Regulation (GDPR), an international data protection law, audio data is subject to privacy regulations [3].

Against this background, speech privacy issues have received significant focus [4]–[8]. For example, the Voice Privacy Challenge (VPC), an international competition aimed at advancing speech anonymization technologies, serves as a key initiative within these ongoing efforts [9]–[11]. The major speaker anonymization approaches involve voice conversion-based techniques, which are employed as the core technology in many methods submitted to VPC [12]–[14]. Such speech

anonymization methods aim to conceal the speaker identity while preserving the linguistic content in an utterance.

An encryption-based speech privacy-preserving method with a random orthogonal matrix as a secret key, distinct from the voice conversion-based methods, has been proposed to protect both the speaker identity information and the linguistic content of an utterance [15]. This method enables deep learning model inference while concealing both speaker identity and speech content. However, this method has limited resistance to inference attacks on speech content and speaker identity. It also cannot be applied to mainstream speech-processing self-supervised learning (SSL) models due to restrictions on compatible model architectures. Furthermore, existing evaluations have only considered attackers without access to encryption systems, indicating that more sophisticated adversaries are needed for comprehensive assessments of speech privacy.

In this work, we propose a method that enhances the attack resistance of the conventional encryption-based speech privacy-preserving technique by employing multiple random orthogonal matrices as secret keys. We also introduce an approach to relax the model constraints of the conventional method. This approach enables the application of our method to a broader range of deep learning models, including mainstream SSL models. Furthermore, we investigate the robustness of the proposed method against attacks by proposing extended attack scenarios. These scenarios include assessments against sophisticated adversaries with access to the encryption system, inspired by the VPC attack scenario. The experimental results confirmed that the proposed method can be applied to automatic speech recognition (ASR) and automatic speaker verification (ASV) models that use a widely used SSL model as their frontend. Furthermore, we found that the proposed method maintains attack resistance, particularly for speaker identity concealment, even under these more sophisticated attack scenarios.

The contributions of our work are as follows:

- 1) **Enhanced attack resistance:** We propose a method that employs multiple random orthogonal matrices as secret keys to enhance the attack resistance of the conventional encryption-based speech privacy-preserving method.
- 2) **Broader model applicability:** The proposed method also includes a function that relaxes the model constraints of the conventional method. The function enables the proposed method to adapt to a broader range of deep learning models, including mainstream SSL models,

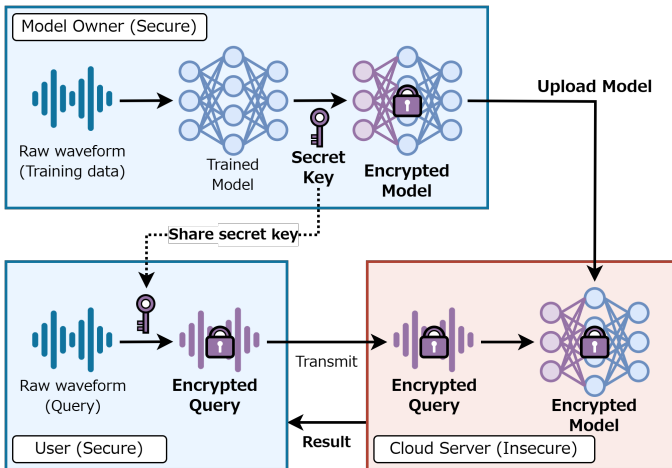


Fig. 1. Privacy-Preserving Scenario

without requiring retraining.

- 3) **Extended evaluation framework:** We advocate comprehensive evaluations using two attack scenarios extended from the scenario employed in VPC, including assessments against sophisticated adversaries with access to the encryption system.

II. PRIVACY-PRESERVING SCENARIO

This section explains the privacy protection scenarios of the encryption-based methods. Figure 1 shows an overview of the scenario, which consists of three domains: model owner, user, and cloud server. The cloud-based model performs inference on the encrypted speech to obtain results on speech processing tasks while protecting user speech privacy. The workflow operates as follows: The model owner trains the model using unencrypted speech datasets, encrypts the model using a secret key, and deploys it to the cloud. The secret key is then shared with an authorized user. The user encrypts query speech using the shared secret key and transmits it to the cloud. The cloud performs inference using the encrypted model without decrypting the query, thereby preventing third parties from accessing the original speech without the secret key. This approach ensures user privacy by making it difficult to infer information from the encrypted queries.

III. PROPOSED METHOD

This section describes the proposed method for enhancing the attack resistance of the conventional encryption-based method and enabling its application to a broader range of deep learning models without retraining.

A. Multiple random orthogonal matrices as secret keys for speech privacy preservation

To enhance the attack resistance of the conventional method, we propose an encryption-based technique that utilizes multiple random orthogonal matrices as secret keys. This approach encompasses the conventional method in which only one matrix is used as a secret key. The following sections explain

the speech encryption and model encryption processes of the proposed method in the scenario described in Section II.

1) *Speech encryption:* Speech encryption is performed by dividing a one-dimensional speech waveform \mathbf{X} into blocks \mathbf{X}_i of size M and multiplying each block \mathbf{X}_i by an M -dimensional random orthogonal matrix. The speech waveform \mathbf{X} , divided into blocks, is represented as follows:

$$\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_i, \dots, \mathbf{X}_t], \quad (1)$$

where t is the total number of blocks. The secret key \mathbf{K}_{mult} contains N random orthogonal matrices of dimension M , as follows:

$$\mathbf{K}_{\text{mult}} = \{\mathbf{K}_1, \dots, \mathbf{K}_n, \dots, \mathbf{K}_N\}. \quad (2)$$

When encrypting speech, each block \mathbf{X}_i is encrypted using a random orthogonal matrix \mathbf{K}_n from \mathbf{K}_{mult} , and the encryption process is as follows:

$$\mathbf{X}_i^{(\mathbf{K}_n)} = \mathbf{X}_i \mathbf{K}_n, \quad n = i \bmod N. \quad (3)$$

The random orthogonal matrix \mathbf{K}_n to be applied to the block \mathbf{X}_i is selected based on the block index i , such that $n = i \bmod N$. Encrypted queries sent to the cloud are formed by concatenating the encrypted blocks, as follows:

$$\mathbf{X}^{(\mathbf{K}_{\text{mult}})} = \left[\mathbf{X}_1^{(\mathbf{K}_1)}, \dots, \mathbf{X}_N^{(\mathbf{K}_N)}, \mathbf{X}_{N+1}^{(\mathbf{K}_1)}, \dots, \mathbf{X}_t^{(\mathbf{K}_{t \bmod N})} \right]. \quad (4)$$

The encrypted speech makes it difficult even for humans to recognize linguistic content and speaker identity.

2) *Model encryption:* Model encryption is a preprocessing step that modifies the model to enable correct prediction from encrypted queries. Similar to the conventional method, this method targets models that directly process one-dimensional speech waveforms as input. Here, we assume that the first layer of the model is a 1D convolutional layer with equal kernel size and stride. The dimension of the random orthogonal matrices in secret key \mathbf{K}_{mult} must be equal to the kernel size. Consider a simplified convolutional layer with kernel size M , stride S , number of output channels 1, and no bias term. The kernel \mathbf{E} of the convolutional layer is expressed as follows:

$$\mathbf{E} = [e_0, \dots, e_k, \dots, e_{M-1}]^\top. \quad (5)$$

Model encryption is performed by multiplying the kernel of the first layer by the transpose \mathbf{K}_n^\top of the matrix applied to the corresponding speech waveform block \mathbf{X}_i . This allows the first layer to cancel out the matrix \mathbf{K}_n applied to block \mathbf{X}_i , making the first layer output equivalent to the unencrypted case. The model encryption process is as follows:

$$\mathbf{E}^{(\mathbf{K}_n)} = \mathbf{K}_n^\top \mathbf{E}. \quad (6)$$

Since N encrypted kernels are prepared, the first layer branches into N encrypted convolutional layers, with each branch processing blocks encrypted with the corresponding \mathbf{K}_n . This approach requires routing encrypted speech blocks to different convolutional layers based on the corresponding matrix \mathbf{K}_n^\top applied to each kernel. While this changes the model structure, it eliminates the need for retraining since the

encrypted kernels are computed from the pre-trained model and the secret key. When using different secret keys, the model can be re-encrypted by applying new keys to the original unencrypted model without retraining.

B. Mitigating Model Restrictions

As explained in Section III-A2, the conventional encryption-based speech privacy preservation method requires that the kernel size and stride of the model's first 1D convolutional layer be equal to each other. This constraint significantly limited the applicability of conventional methods to models. In this section, we explain how the proposed method, through a modification to the waveform block partitioning method, can relax the constraints on its applicability to deep learning models. This approach enables correct computation during the model encryption, even when the stride of the first convolutional layer does not match the kernel size. Similar to Section III-A2, consider a 1D convolutional layer with kernel size M , stride S , and no bias term. The kernel of this convolutional layer is expressed as Equation (5). The input \mathbf{Y} and output \mathbf{Z} of the convolutional layer are expressed as follows:

$$\mathbf{Y} = [y_0, y_1, \dots, y_{T-1}], \quad (7)$$

$$\mathbf{Z} = [z_0, \dots, z_i, \dots, z_{L-1}], \quad (8)$$

where T denotes the length of the input to the convolutional layer, L denotes the length of its output, and $L = \lfloor \frac{T-M}{S} + 1 \rfloor$. Under the condition that the kernel size and stride satisfy $S < M$, we assume that the convolutional layer can be computed as an inner product between the block and the kernel. Each element z_i of the convolutional layer output is expressed as follows:

$$z_i = \sum_{k=0}^{M-1} e_k y_{Si+k} = A_i \mathbf{E}, \quad (9)$$

where $A_i = [y_{Si}, y_{Si+1}, \dots, y_{Si+(M-1)}]$, and A_i denotes a vector consisting of elements from the input \mathbf{Y} that are referenced when computing the output element z_i of the convolutional layer. The array \mathbf{A} is then formed by extracting and concatenating these vectors A_i (A_0 to A_{L-1}) from the input audio \mathbf{Y} . Figure 2 illustrates this construction of \mathbf{A} for the case where $M = 3$ and $S = 2$. As depicted, when the stride S is smaller than the kernel size M , consecutive vectors A_i share $M - S$ overlapping samples from the input \mathbf{Y} . This structured concatenation of overlapping blocks A_i is crucial for enabling the transformation of a convolutional layer with $S < M$ into an equivalent one where the effective stride is $S = M$ when \mathbf{A} is used as input. Here, we consider modifying the convolutional layer to use \mathbf{A} as input to the model. By changing the stride to $S = M$ without changing the kernel size, when \mathbf{A} is provided as input, an output that matches the output of the original convolutional layer can be obtained. The length L_A of the input to the modified convolutional layer becomes as follows:

$$L_A = \left\lfloor \frac{T-M}{S} + 1 \right\rfloor M. \quad (10)$$

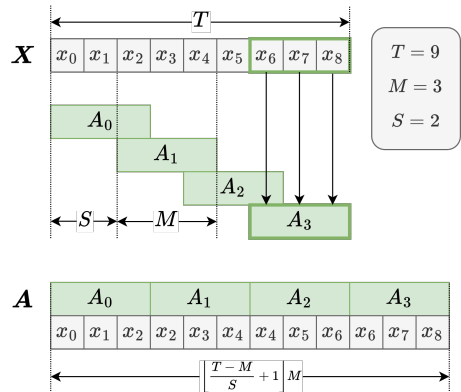


Fig. 2. Block partitioning method

As illustrated in Figure 2, L_A becomes longer than the original input. By utilizing this method, a convolutional layer with $S < M$ can be modified to a convolutional layer with $S = M$ that performs the same computation. As a result, the proposed method can be applied to models with convolutional layers in which the kernel size and stride differ. Speech encryption can be performed on \mathbf{A} using the same procedure described in Section III-A1. For model encryption, we use the same procedure described in Section III-A2 but modify the stride of the model's first convolutional layer to $S = M$.

IV. EVALUATION OF ATTACK RESISTANCE

In this section, we propose two attack scenarios, inspired by VPC 2020 [9] and VPC 2024 [11], to evaluate the attack resistance of the proposed method. Attacks against encrypted queries aim to infer information, specifically linguistic content and speaker identity, from original speech embedded within the encrypted queries. Therefore, we extend the VPC scenarios, which originally focused only on speaker identity concealment, to include linguistic content inference attacks.

A. Adversary model

Adversaries use ASV models and ASR models to infer speaker identity and linguistic content from an utterance, respectively. We assess the attack resistance of the proposed method under scenarios where attackers have access to the following two types of information:

- 1) Encrypted query speech transmitted by users to cloud-based models.
- 2) Encryption algorithm used to encrypt query speech.

In contrast, the secret key used for encryption is inaccessible to adversaries. Adversaries leverage the accessible information, along with ASR and ASV models, to infer linguistic content and speaker identity from encrypted queries. Additionally, adversaries may preprocess encrypted audio or adapt their attack models using the accessible information to improve attack accuracy.

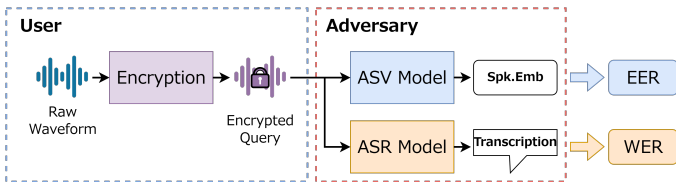


Fig. 3. Attack scenario 1

B. Attack scenario 1

Attack scenario 1 assumes that the adversary performs attacks using only information specified in Section IV-A item (1), i.e., exclusively encrypted queries. This attack scenario is based on the “ignorant attacker” scenario from the VPC2020 [16]. Figure 3 shows an overview of attack scenario 1. The adversary obtains encrypted queries transmitted by users and attempts to infer linguistic content and speaker embeddings using pre-trained ASR and ASV models. We use the following two evaluation metrics to assess attack resistance.

Linguistic content inference: Word Error Rate (WER) between the adversary’s ASR transcription of the encrypted query and the original transcription indicates how accurately the speech content can be inferred.

Speaker identity inference: Equal Error Rate (EER) is calculated based on the similarity between speaker embeddings extracted by the adversary’s ASV model from encrypted queries and speaker embeddings from unencrypted enrollment utterances of the same speaker. Higher WER and EER values indicate greater attack resistance for the encryption method.

C. Attack scenario 2

Attack scenario 2 assumes that the adversary performs attacks using both types of information specified in Section IV-A: (1) encrypted queries and (2) the encryption algorithm. This attack scenario is based on the “semi-informed attacker” scenario from the VPC2024 [11]. Figure 4 shows an overview of attack scenario 2. Unlike in Attack Scenario 1, the adversary in this scenario has knowledge of the encryption algorithm and can generate encrypted datasets using the same encryption system. The adversary fine-tunes ASR and ASV models using these encrypted datasets and uses the adapted models to conduct attacks on encrypted queries. The attack resistance is evaluated using WER and EER metrics. For EER computation, speaker embeddings are extracted from both encrypted query utterances and encrypted enrollment utterances. This approach is used because the ASV model has been adapted to encrypted queries through fine-tuning. Attack Scenario 2 represents a more sophisticated attack than Attack Scenario 1. In this scenario, the adversary leverages models specifically adapted to encrypted queries to improve inference accuracy.

V. EXPERIMENT

This section presents experiments conducted to validate the proposed encryption method and evaluate its attack resistance.

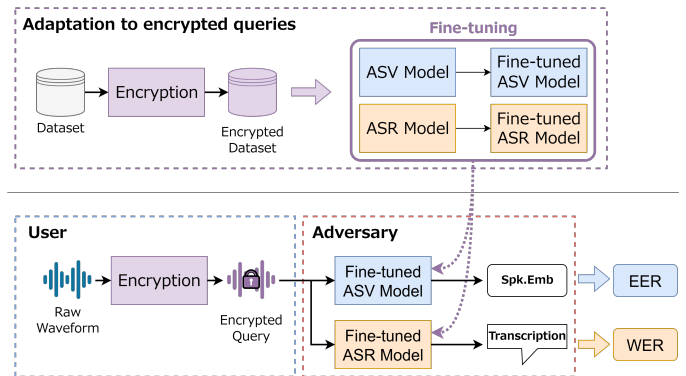


Fig. 4. Attack scenario 2

We conducted experiments to validate that the proposed encryption method can be applied to models with different kernel sizes and strides. This was not possible with conventional methods. Additionally, we evaluated the attack resistance of the encryption method through experiments based on the attack scenarios described in Section IV-A.

A. Pre-trained models

We utilize wav2vec 2.0 [17] Large and wav2vec 2.0 Base as frontends for ASR and ASV models in the experiments. The ASR model backend consists of a 2-layer fully connected layer with 1024 units, while the ASV model uses x-vector [18] for the backend. The first convolutional layer of wav2vec 2.0 has a kernel size of 10 and a stride of 5, representing settings in which previous encryption methods cannot be applied. Both wav2vec 2.0 Large and Base models share the same CNN feature extractor architecture, differing only in the number of Transformer blocks and attention heads. The ASR and ASV models are pre-trained on LibriSpeech [19] and VoxCeleb1 [20] datasets, respectively.

For Attack Scenario 2, we fine-tune the pre-trained models using encrypted datasets to create models adapted to encrypted speech for attacks. The encrypted datasets are constructed by encrypting the same datasets used for pre-training, using randomly generated secret keys for each utterance. For ASR model fine-tuning, we use only the train-clean-100 and train-clean-360 subsets of LibriSpeech. Fine-tuning uses learning rates of $1e-4$ for the frontend and $5e-3$ for the backend.

B. Preprocessing

As mentioned in Section IV-A, adversaries can apply preprocessing to encrypted queries within the scope defined by the adversary model. We apply two preprocessing techniques with dual objectives: improving adversary model performance and stabilizing the fine-tuning process using encrypted datasets in Attack Scenario 2.

The first preprocessing is time-scale adjustment of encrypted queries. As described in Section III-B, the length of encrypted queries given by Equation (10) is stretched because consecutive blocks A_i and A_{i+1} contain overlapping samples of length $M - S$, as shown in Figure 2. To match the original audio

TABLE I
PERFORMANCE OF PRE-TRAINED MODELS UNDER THE PROPOSED
ENCRYPTION METHOD
(NO ENCRYPTION: WER = 1.76[%], EER = 3.66[%])

N	Correct key		Incorrect key	
	WER [%]	EER [%]	WER [%]	EER [%]
1	1.76	3.66	67.1	30.8
3	1.76	3.66	94.2	34.6
5	1.76	3.66	96.6	36.8

length, we remove overlapping samples between consecutive blocks. The processed result \tilde{A} is expressed using blocks A_i from Equation (9) as follows:

$$\tilde{A} = [\tilde{A}_0, \dots, \tilde{A}_i, \dots, \tilde{A}_L], \quad (11)$$

$$\tilde{A}_0 = A_0,$$

$$\tilde{A}_i = (A_i[M - S], A_i[M - S + 1], \dots, A_i[M - 1]).$$

The resulting length is expressed as follows:

$$|\tilde{A}| = |\tilde{A}_0| + \sum_{i=1}^{L-1} |\tilde{A}_i| = M + (L - 1)S = T. \quad (12)$$

This achieves length consistency with the original audio. This preprocessing is used in both Attack Scenarios 1 and 2, and is incorporated into the fine-tuning process for Attack Scenario 2.

The second preprocessing method is low-pass filtering, applied exclusively in Attack Scenario 1 after the first preprocessing step. A low-pass filter with a cutoff frequency of 4 kHz removes high-frequency components from encrypted query speech before being input to the adversary's model.

C. Experimental results

Table I shows the performance of ASR and ASV when encryption is applied to both the pre-trained model and the input speech. N denotes the number of random orthogonal matrices in the secret key. In Table I, when using the correct key, WER and EER are equivalent to the unencrypted case, showing no performance degradation. Here, a correct key means that the speech and model are encrypted with the same secret key. This occurs because the output of the first convolutional layer remains identical to the unencrypted case when both audio and the model share the same secret key, as explained in Section III. In contrast, when audio and the model use different secret keys, WER and EER values increase significantly compared to the correct key scenario. These results demonstrate that the proposed method enables encrypted inference without performance degradation for model configurations incompatible with the conventional method.

Tables II and III show the attack resistance evaluation results for Attack Scenarios 1 and 2. Table II shows that both WER and EER increase with the number of random orthogonal matrices N . Specifically, WER increased from 40.6% ($N = 1$) to 97.5% ($N = 9$), and EER increased from 33.1% to 46.8%. This indicates that increasing N makes it more difficult to infer linguistic content and speaker identity from encrypted speech, thereby improving privacy protection. Similar trends

TABLE II
EVALUATION OF ATTACK RESISTANCE IN ATTACK SCENARIO 1

N	Without preprocessing		LPF	
	WER [%]	EER [%]	WER [%]	EER [%]
1	40.6	33.1	38.3	30.8
3	90.0	44.8	87.1	45.1
5	93.5	46.3	90.0	46.0
7	96.8	46.5	94.4	46.7
9	97.5	46.8	94.9	47.0

TABLE III
EVALUATION OF ATTACK RESISTANCE IN ATTACK SCENARIO 2

N	WER[%]	EER[%]
1	3.66	11.4
3	7.53	22.3
5	9.95	22.8
7	11.4	26.9
9	11.2	27.0

were observed with low-pass filtering, where WER and EER showed slight decreases, but attack resistance was generally maintained.

Attack Scenario 2 results (Table III) show the evaluation under stronger attack conditions in which attackers have access to the encryption system. WER increased from 3.66% ($N = 1$) to 11.2% ($N = 9$), and EER increased from 11.4% to 27.0%, confirming improved attack resistance with increasing N . However, both WER and EER were lower than those in Attack Scenario 1, indicating that attacks using models adapted to encrypted speech are more effective.

These results demonstrate that increasing the number of random orthogonal matrices effectively improves attack resistance in both scenarios. Even against stronger attacks like Attack Scenario 2, a certain level of privacy protection is maintained, particularly for speaker identity concealment.

D. Conclusion

In this work, we proposed an encryption-based method for preserving speech privacy that addresses the limitations of the conventional method. The proposed method enhances attack resistance by utilizing multiple random orthogonal matrices as secret keys and by relaxing model applicability constraints to support a broader range of deep learning models, including mainstream SSL models. To evaluate the attack resistance of the proposed method, we introduced sophisticated attack scenarios that incorporate more challenging attack models and adversaries compared to those typically explored in VPC. The experimental results confirmed that the proposed method enables encryption to be applied to models in which the previous method could not be applied due to architectural constraints. Even under Attack Scenario 2, which assumes more sophisticated adversaries, the results demonstrated that privacy protection performance for speaker identity concealment is maintained to a reasonable extent. Future work should evaluate the attack resistance of the encryption method using models with various architectures and training data beyond that used in this work.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI (Grant Number JP24K14993), SCAT, and ROIS DS-JOINT (026RP2025) to S. Shiota.

REFERENCES

- [1] F. Metze, J. Ajmera, R. Englert, *et al.*, “Comparison of four approaches to age and gender recognition for telephone applications,” in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07*, IEEE, vol. 4, 2007, pp. IV–1089.
- [2] Z. Fan, M. Li, S. Zhou, and B. Xu, *Exploring wav2vec 2.0 on speaker verification and language identification*, 2021. arXiv: 2012.06185 [cs.SD]. [Online]. Available: <https://arxiv.org/abs/2012.06185>.
- [3] 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,” *arXiv preprint arXiv:1907.03458*, 2019.
- [4] H. Kai, S. Takamichi, S. Shiota, and H. Kiya, “Lightweight voice anonymization based on data-driven optimization of cascaded voice modification modules,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 560–566.
- [5] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, *Speaker anonymisation using the mcadams coefficient*, 2021. arXiv: 2011.01130 [eess.AS]. [Online]. Available: <https://arxiv.org/abs/2011.01130>.
- [6] S. Gharib, M. Tran, D. Luong, K. Drossos, and T. Virtanen, “Adversarial representation learning for robust privacy preservation in audio,” *IEEE Open Journal of Signal Processing*, vol. 5, pp. 294–302, 2024, ISSN: 2644-1322. DOI: 10.1109/ojsp.2023.3349113. [Online]. Available: <http://dx.doi.org/10.1109/OJSP.2023.3349113>.
- [7] D. K. Singh, G. P. Prajapati, and H. A. Patil, “Voice privacy using time-scale and pitch modification,” *SN Computer Science*, vol. 5, no. 2, p. 243, 2024.
- [8] S. Tayebi Arasteh, T. Arias-Vergara, P. A. Pérez-Toro, *et al.*, “Addressing challenges in speaker anonymization to maintain utility while ensuring privacy of pathological speech,” *Communications Medicine*, vol. 4, no. 1, p. 182, 2024.
- [9] N. Tomashenko, B. M. L. Srivastava, X. Wang, *et al.*, *The voiceprivacy 2020 challenge evaluation plan*, 2022. arXiv: 2205.07123 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2205.07123>.
- [10] N. Tomashenko, X. Wang, X. Miao, *et al.*, *The voiceprivacy 2022 challenge evaluation plan*, 2022. arXiv: 2203.12468 [eess.AS]. [Online]. Available: <https://arxiv.org/abs/2203.12468>.
- [11] N. Tomashenko, X. Miao, P. Champion, *et al.*, *The voiceprivacy 2024 challenge evaluation plan*, 2024. arXiv: 2404.02677 [eess.AS]. [Online]. Available: <https://arxiv.org/abs/2404.02677>.
- [12] J. Yao, N. Kuzmin, Q. Wang, *et al.*, *Npu-ntu system for voice privacy 2024 challenge*, 2025. arXiv: 2409.04173 [eess.AS]. [Online]. Available: <https://arxiv.org/abs/2409.04173>.
- [13] H. L. Xinyuan, Z. Cai, A. Garg, *et al.*, *Hltcoe jhu submission to the voice privacy challenge 2024*, 2024. arXiv: 2409.08913 [eess.AS]. [Online]. Available: <https://arxiv.org/abs/2409.08913>.
- [14] S. Akti, T. N. Nguyen, Y. Liu, and A. Waibel, “Voice privacy - investigating voice conversion architecture with different bottleneck features,” in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 44–49. DOI: 10.21437/SPSC.2024-8.
- [15] N. Shoko, S. Shiota, H. Kiya, *et al.*, “Speech privacy-preserving methods using secret key for convolutional neural network models and their robustness evaluation,” *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 1, 2024.
- [16] N. Tomashenko, X. Wang, E. Vincent, *et al.*, “The voiceprivacy 2020 challenge: Results and findings,” *Computer Speech & Language*, vol. 74, p. 101362, Jul. 2022, ISSN: 0885-2308. DOI: 10.1016/j.csl.2022.101362.
- [17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.
- [18] 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)*, IEEE, 2018, pp. 5329–5333.
- [19] 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)*, IEEE, 2015, pp. 5206–5210.
- [20] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: A large-scale speaker identification dataset,” *arXiv preprint arXiv:1706.08612*, 2017.