

Investigation of perception inconsistency in speaker embedding for asynchronous voice anonymization

Rui Wang*, Liping Chen*, Kong Aik Lee†, Zhengpeng Zha*, Zhenhua Ling*

* University of Science and Technology of China

E-mail: wangrui256@mail.ustc.edu.cn {lipchen, zhazp, zhling}@ustc.edu.cn

† The Hong Kong Polytechnic University

E-mail: kong-aik.lee@polyu.edu.hk

Abstract—Given the speech generation framework that represents the speaker attribute with an embedding vector, asynchronous voice anonymization can be achieved by modifying the speaker embedding derived from the original speech. However, the inconsistency between machine and human perceptions of the speaker attribute within the speaker embedding remains unexplored, limiting its performance in asynchronous voice anonymization. To this end, this study investigates this inconsistency via modifications to speaker embedding in the speech generation process. Experiments conducted on the FACodec and Diff-HierVC speech generation models discover a subspace whose removal alters machine perception while preserving its human perception of the speaker attribute in the generated speech. With these findings, an asynchronous voice anonymization is developed, achieving 100% human perception preservation rate while obscuring the machine perception. Audio samples can be found in <https://voiceprivacy.github.io/speaker-embedding-eigen-decomposition/>.

I. INTRODUCTION

Advancements in speech technologies have intensified security risks related to the misuse of speaker attributes, necessitating the development of voice privacy protection techniques. In this context, voice anonymization, originating from the 1980s [1], has regained the interest of the community as it provides a viable solution to protecting speaker attributes from being extracted by speaker models. To date, voice anonymization can be realized in both synchronous[2], [3], [4] and asynchronous[5], [6], [7] manners. Synchronous voice anonymization alters both machine-discriminable and human-perceivable speaker attributes. Asynchronous voice anonymization modifies the machine-discriminable attribute while preserving the human-perceivable attribute.

Facilitated by the speech generation framework wherein the speaker attribute is disentangled and represented with an embedding, voice anonymization can be realized by replacing the original speaker with a pseudo-speaker [2], [7]. In asynchronous voice anonymization, the construction of the pseudo-speaker forms a critical challenge. An asynchronous voice anonymization method was proposed in [7], where

Corresponding author: Liping Chen.

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.

the pseudo-speaker embedding was generated by introducing adversarial perturbation to the speaker embedding. In this approach, stronger perturbations provided better protection against machine perception with larger alteration of the machine-discriminable speaker attribute, whereas weaker perturbation better preserves the human-perceivable attribute. Due to the lack of differentiating between machine and human perceptions in the speaker embedding, the adversarial method exhibits limitations in protecting machine perception of the speaker attribute while preserving its human perception.

This paper investigates the differences between machine and human perceptions of speaker attributes. Given a speech generation model that incorporates speaker attribute disentanglement and its representation via an embedding vector, speaker-modified speech utterances are generated through modifications to the speaker embedding. Specifically, the modification is performed in a subspace of the speaker embedding by eliminating its contribution from the embedding. Machine and human perceptions of the speaker attribute within the speaker-modified utterances were examined individually in the experiments. To our knowledge, this is the first investigation of the differences between the machine and human perceptions within the speaker embedding in the context of speech generation. Our contributions are as follows:

- 1) Experimental findings in two speech generation models, FACodec[8] and Diff-HierVC[9], demonstrate that a speaker variability subspace exists, whose removal exclusively influences machine perception of the speaker attribute without affecting its human perception. This reveals the inconsistency between machine and human perceptions of speaker attributes within the speaker embedding.
- 2) An asynchronous voice anonymization method is developed, in which the pseudo-speaker vector is obtained by removing the subspace contribution from the original speaker vector. It achieves a 100% human perception preservation rate while obscuring the machine-discriminable speaker attributes in the anonymized speech.

II. BACKGROUND

Fig. 1 illustrates the framework for information disentanglement, facilitating the generation of speaker-modified speech. The disentanglement-based speech generation framework and

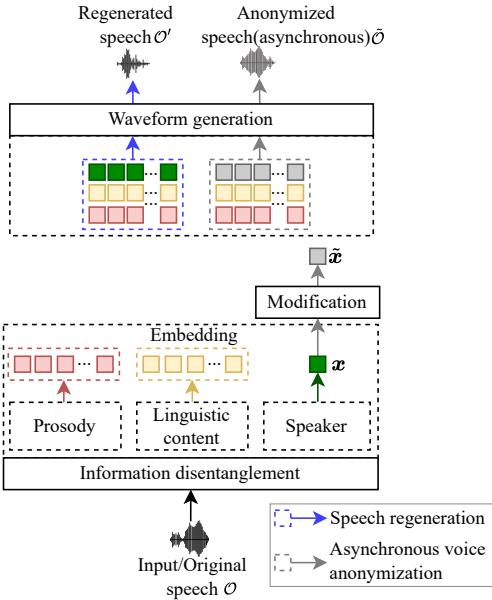


Fig. 1: Speech generation framework based on information disentanglement. Colored boxes represent embeddings of disentangled attributes: red (prosody), yellow (linguistic content), green (original speaker \mathbf{x}). The gray box represents the modified speaker embedding $\tilde{\mathbf{x}}$. The blue and gray rectangular boxes with dotted lines, along with the corresponding arrowed lines, indicate speech regeneration and speaker-modified speech generation processes, respectively.

its application in generating speaker-modified speech are detailed in the following.

A. Disentanglement-based speech generation

Fig. 1 presents the speech generation framework based on information disentanglement and waveform generation. In the information disentanglement phase, given an input speech utterance \mathcal{O} , three distinct attributes, including prosody, linguistic content, and speaker characteristics, are disentangled and represented by separate embedding vectors. A specific configuration is illustrated where prosody and linguistic content attributes are extracted from speech frames and represented with sequences of embedding vectors. Besides, the speaker attribute is encoded for the entire utterance using a single vector \mathbf{x} . In the waveform generation phase, \mathbf{x} is replicated to match the length of the prosody and content embedding sequences. The embedding vectors of the three attributes are input into the waveform generation module, producing a speech waveform \mathcal{O}' , which is a regenerated version of \mathcal{O} . The disentanglement-based speech generation mechanism enables control over generated speech by facilitating the manipulation of speech attributes, especially prosody [10], [11] and speaker characteristics[8], [9], [12], [13].

B. Anonymized speech (asynchronous) generation

In the asynchronous voice anonymization method [7] built upon the speech generation framework depicted in Fig. 1,

given an original utterance, the prosody, linguistic content, and speaker attributes were disentangled and represented with separate embedding vectors. The speaker embedding vector \mathbf{x} , denoted with the green square box, was subsequently modified to $\tilde{\mathbf{x}}$ as the pseudo-speaker vector (the gray square box in Fig. 1). Thereby, the anonymized speech $\tilde{\mathcal{O}}$ was generated by the waveform generation module and used as the anonymized speech, utilizing the original prosody and linguistic content embedding vectors and the modified speaker vector $\tilde{\mathbf{x}}$.

III. SPEAKER MODIFICATION

Given the speech generation framework depicted in Fig. 1, our investigation into the differences between machine and human perceptions of speaker attributes in speaker embedding is conducted through modifications to the original speaker embedding \mathbf{x} , followed by experimental evaluations of both perceptions within the speaker-modified utterances $\tilde{\mathcal{O}}$. The modification is performed within the variability subspaces of speaker embedding as detailed in the following.

A. Variability space

A speaker embedding vector \mathbf{x} is hypothesized to be decomposable with a basis matrix \mathbf{V} as follows:

$$\mathbf{x} = \mathbf{V}\mathbf{c}. \quad (1)$$

Assume \mathbf{x} to be a D -dimensional vector, \mathbf{V} is composed of D orthogonal unit vectors as $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_D]$, with the vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_D\}$ serving as the *basis vectors*. The vector $\mathbf{c} = [c_1, \dots, c_D]^T$ is the coefficient vector, with each element quantifying the contribution of the corresponding basis vector in constructing \mathbf{x} . The space span by \mathbf{V} captures the variability of \mathbf{x} , and is referred to as *variability space*.

Given N speech utterances, the speaker embedding vectors are extracted and denoted as $\mathcal{X} = \cup_{n=1}^N \mathbf{x}_n$. Firstly, the covariance matrix Σ is calculated from \mathcal{X} .

Thereafter, \mathbf{V} is derived via eigen-decomposition of Σ as follows:

$$\Sigma = \mathbf{V}\Lambda\mathbf{V}^T, \quad (2)$$

where $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_D]$ is obtained as the matrix of eigenvectors, and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$ is a diagonal matrix with eigenvalues on the diagonal. Particularly, the eigenvalues are sorted in descending order as $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D$. An identical mechanism for variability space derivation is utilized in principal component analysis (PCA) [14]. Readers are referred to it for mathematical details.

B. Speaker embedding modification

Given the speaker vector \mathbf{x}_n extracted from the n -th utterance, the coefficient vector in space \mathbf{V} is obtained as follows:

$$\mathbf{c}_n = \mathbf{V}^T \mathbf{x}_n, \quad (3)$$

where \mathbf{c}_n consists of D elements as $\{c_{n,1}, \dots, c_{n,D}\}$. The modification of \mathbf{x}_n is achieved by altering \mathbf{c} to $\tilde{\mathbf{c}}_n$ which is conducted on the subspaces of \mathbf{V} . Specifically, a subspace is

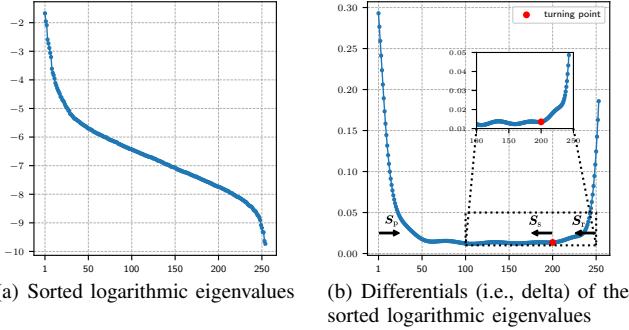


Fig. 2: Sorted logarithmic eigenvalues and their differentials of the speaker variability space within the open-source FACodec model. The speaker embedding is of dimension $D=256$. The arrows in Fig. 2(b) indicate the three modification subspaces S_p , S_s and S_r , starting from their initial dimensions and directing along their spans, respectively.

characterized by three parameters: initial dimension i , subspace size K , and the span direction (forward or backward). It is represented as $S = \{i, K, \text{direction}\}$, indicating that the subspace spans K dimensions from the i -th dimension in the designated direction. The contribution of S is eliminated from \mathbf{x} by setting the coefficients associated with its basis vectors to 0 as follows:

$$\begin{aligned} c_{n,i}, \dots, c_{n,i+K-1} &= 0, & \text{if direction} &= + \\ c_{n,i-K+1}, \dots, c_{n,i} &= 0, & \text{if direction} &= - \end{aligned} \quad (4)$$

where “+” (forward) and “-” (backward) are the span direction of the subspace. Combined with the remaining coefficients, the modified coefficient vector $\tilde{\mathbf{c}}_n$ is obtained. Finally, the modified speaker embedding vector $\tilde{\mathbf{x}}_n$ is obtained as follows:

$$\tilde{\mathbf{x}}_n = \mathbf{V}\tilde{\mathbf{c}}_n. \quad (5)$$

C. Modification subspaces

For presentational clarity, the variability space of the speaker embedding in the open-source FACodec model is adopted for description. Derived from the speaker embedding vector set extracted from the speech utterances in the LibriSpeech train-clean-360[15] dataset, the logarithmic eigenvalues, i.e., $\{\log\lambda_1, \dots, \log\lambda_D\}$, are plotted in Fig. 2(a). Fig. 2(b) illustrates the delta of the logarithmic eigenvalues, computed as follows:

$$a_i = \log\lambda_{i+1} - \log\lambda_i, \quad (6)$$

for $i = 1, \dots, D - 1$. Particularly, the FACodec model employs speaker vectors of dimension $D=256$.

As shown by the arrows in Fig. 2(b), given the delta log-eigenvalue, three subspace region are investigated for the modification of the speaker embedding: primary, secondary, and residual subspaces. The *primary subspace* starts from the 1st dimension and spans in the forward direction of size K_p , denoted as $S_p = \{1, K_p, +\}$. It is composed of the dominant

basis vectors of the variability space. The *secondary subspace* is defined at the turning point where the differential oscillates slightly before and increases monotonically after. The point is marked by the red dot in Fig. 2(b) with the dimension i_s . The secondary subspace starts from i_s and spans in the backward direction with size K_s , denoted as $S_s = \{i_s, K_s, -\}$. The *residual subspace* spans from the last dimension D in the backward direction with size K_r , denoted as $S_r = \{D, K_r, -\}$. Between them, S_r represents the least important subspace in the variability space. The importance of S_s is higher than S_r while lower than S_p . In these subspaces, the subscripts p , s , and r are short for primary, secondary, and residual, respectively.

IV. EXPERIMENTS

A. Dataset & speech generation models

Our evaluations were conducted on the dev-clean subset of the LibriSpeech [15] dataset, including 2,703 utterances from 20 female and 20 male speakers. All recordings were resampled to 16 kHz. The speaker embedding variability space was obtained from the LibriSpeech train-clean-360 dataset. The open-source FACodec and Diff-HierVC models were examined as the speech generation model. The dimensions of the speaker vectors in both models are 256. The turning dimensions i_s are 200 and 218 in FACodec and Diff-HierVC, respectively.

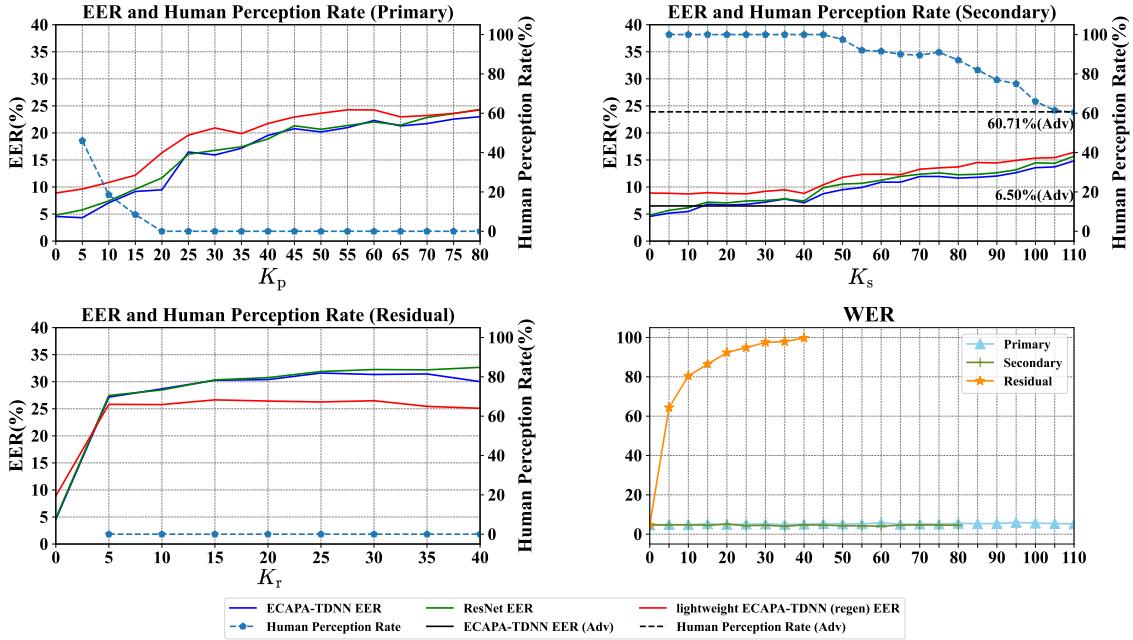
B. Evaluation Metrics

Evaluations were conducted to assess the machine and human perceptions of the speaker attributes in the utterances generated with the modified speaker embedding vectors. Given the degradation introduced by the speech generation model, the original utterance \mathcal{O} was regenerated using its extracted speaker vector \mathbf{x} , giving \mathcal{O}' . It served as the reference for a fair comparison with the speaker-modified speech $\tilde{\mathcal{O}}$. Additionally, in line with the requirement of the voice anonymization task [3], [4], [16], the linguistic content preservation capability was measured.

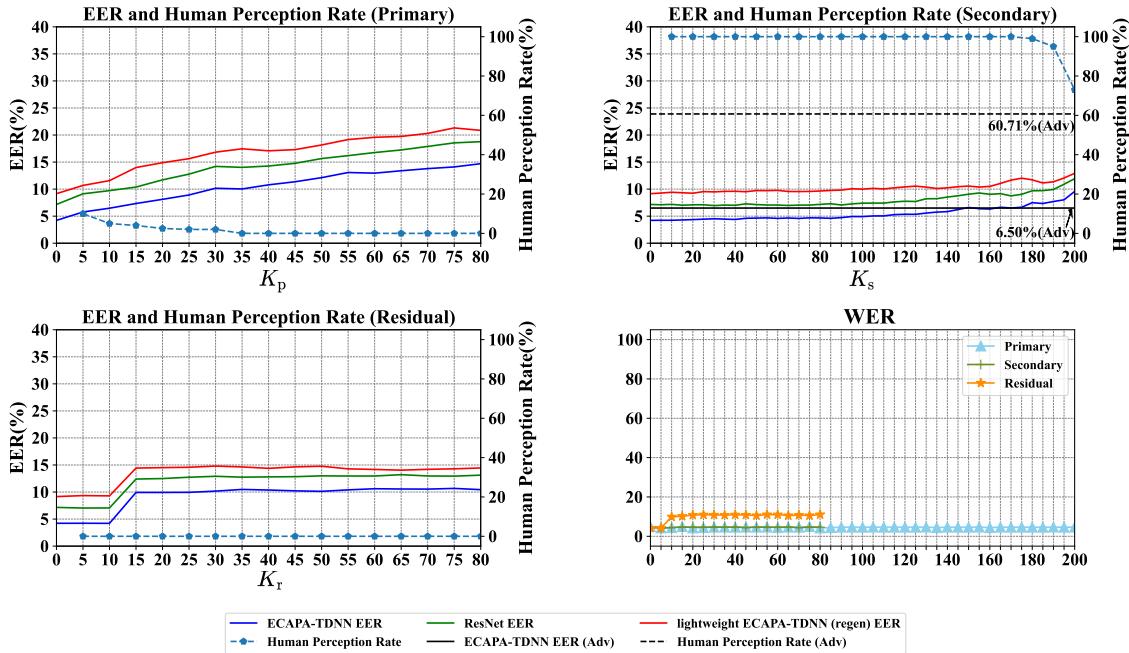
- *Machine perception modification*: Automatic speaker verification (ASV) evaluation was conducted to assess the modification of machine-perceivable speaker attributes. Three speaker embedding extractors were employed: ECAPA-TDNN[17], ResNet[18], and a lightweight ECAPA-TDNN. Specifically, the ECAPA-TDNN and ResNet extractors were trained with the VoxCeleb1 & 2 datasets[19], [20] using the ASVSubtools open-source toolkit[21]. The lightweight ECAPA-TDNN extractor was trained on the regenerated speech of the LibriSpeech train-other-500 subset. The modified speech $\tilde{\mathcal{O}}$ was used for both enrollment and testing. Cosine similarities between speaker embeddings extracted by the three models were used for scoring. ASV performance was measured in terms of equal error rate (EER), where a higher

https://huggingface.co/spaces/amphion/naturalspeech3_facodec

<https://github.com/hayeong0/Diff-HierVC>
<https://github.com/Snowdar/asv-subtools>



(a) FACodec



(b) Diff-HierVC

Fig. 3: Machine perception, human perception, and linguistic content evaluation results across the primary, secondary, and residual subspaces under varying sizes. The results of FACodec and Diff-HierVC are shown in Fig. 3(a) and Fig. 3(b), respectively. The machine perception (measured by EER(%)) and human perception (measured by human perception rate(%)) of the primary and secondary subspace configurations are shown in the first row, in the first and second columns, respectively, while the results for the residual subspace are presented in the left column of the second row. The EERs obtained with the ECAPA-TDNN, ResNet, and lightweight ECAPA-TDNN trained with regenerated (regen) speech are included. The WERs(%) obtained in the three configurations are presented in the right column of the second row. The EER and human perception preservation rate obtained by the adversarial method (Adv) [7] are included in the plots of secondary subspace with solid and dotted black lines, respectively.

EER indicates a stronger alteration of machine-discriminable speaker characteristics. The ASV evaluations were conducted on the trials provided by VPC 2024[16], with scores from male and female trials pooled together for EER calculation.

- *Human perception preservation:* Subjective listening tests were conducted to assess the preservation of human perception. In each test, 200 utterances were randomly selected from the evaluation dataset. For each test utterance, given the pair of its regenerated version \mathcal{O}' and the speaker-modified version \mathcal{O} , five listeners were asked to decide whether the speakers were indistinguishable. Listeners gave a *yes* (indistinguishable) or *no* (distinguishable) for each utterance pair. A pair was decided to be perceived as the same speaker if it got a minimum of three *yes*.
- *Linguistic content preservation:* The preservation of the linguistic content of the original speech was measured with automatic speech recognition (ASR) evaluations. The Whisper model[22] provided by OpenAI was called. The performances were measured with word error rates (WERs).

C. Experimental configurations

Our study investigated various subspaces by varying the sizes of the three subspace types. In the experiments on the primary subspace, the size K_p was examined from 0 to 80 with step 5 for both the FCodec and Diff-HierVC models. In the secondary subspace experiments, the size K_s was examined from 0 to 80 with step 5 for the FCodec model. For the Diff-HierVC model, K_s was examined from 0 to 200 with step 5. In the residual subspace experiments, size K_r was examined from 0 to 40 with step 5 for the FCodec model. For the Diff-HierVC model, K_s was examined from 0 to 80 with step 5. Notably, in these experiments, the sizes of 0 indicate no modifications applied to the speaker embedding, resulting in the speech being the regenerated speech \mathcal{O}' . For comparison, the EER obtained in the ASV evaluation and the human perception preservation rate achieved by the adversarial method proposed in [7] are presented together with the secondary subspaces, represented with *Adv*. The ECAPA-TDNN speaker extractor was utilized in its ASV evaluation.

D. Results

In the primary subspace experiments, as K_p increased from 0, the ASV EERs rose rapidly to approximately 25% on the three speaker extractors, i.e., ECAPA-TDNN, ResNet, lightweight ECAPA-TDNN. This indicates an obvious modification in the machine-discriminable speaker attributes. However, the human perception rates decreased sharply from 46.00% to 0% at $K_s = 20$, indicating its incapability of preserving the human perception of speaker attributes. These observations indicate that the removal of the contribution of the primary

subspace alters both of speaker attributes in speaker-modified speech. This suggests that it is associated with both machine-discriminable and human-perceptible speaker attributes.

The secondary subspace results show that the EER increased as the subspace size increased. Besides, in the FCodec model, the human perception rate remained at 100% up to $K_s = 45$. Similar results are observed in the Diff-HierVC model with K_s ranging from 5 to 170. These observations demonstrate the removal of these secondary subspaces did not change the human perception while obscuring the machine perception of speaker attributes, indicating the inconsistency between the two perceptions. Moreover, the speaker-modified speech attained comparable WERs with the regenerated speech ($K_s = 0$), indicating that such alterations to the speaker embedding did not compromise the linguistic content.

In the residual subspace experiments, the WER significantly increased in the FCodec model. Besides, a notable rise is found in the Diff-HierVC model, increasing from 4.17% to 10.96%. These observations suggest the influence of the residual subspace on linguistic content, demonstrating that the removal of this subspace is incapable of voice anonymization, which requires the preservation of linguistic content.

Seeing from the evaluations with the ECAPA-TDNN speaker extractor, comparing with the regenerated speech ($K_s = 0$), the removal of the secondary subspace $S_s = \{200, 45, -\}$ in FCodec achieved an increase in EER from 4.79% to 8.76%. Similarly, the subspace $S_s = \{218, 170, -\}$ in Diff-HierVC yielded an EER increase from 4.22% to 6.65%. Both modifications preserved human perception at the 100% rate and did not cause degradation to the linguistic content. These results demonstrate that an asynchronous voice anonymization method can be developed by removing the contribution from the subspace in the speaker embedding. Particularly, in the FCodec model, it outperformed the adversarial approach [7] at $K_s = 45$, achieving a higher EER (8.76% vs. 6.50%) and a higher human perception preservation rate (100% vs. 60.71%).

V. CONCLUSIONS AND DISCUSSIONS

This paper investigated the inconsistency between the machine and human perceptions on speaker attributes in the speech generation framework. It was conducted within the speaker variability subspaces of the speech generation models FCodec and Diff-HierVC. Experimental findings reveal that in both models, a subspace within the speaker embedding variability space exists, whereby the removal of its contribution from the speaker embedding alters machine-detectable speaker attributes while preserving human perception. Based on the investigation, an asynchronous voice anonymization method is developed through the removal of the subspace from the speaker embedding.

In addition to enhancing voice privacy protection, future research will focus on comprehensive evaluations of techniques for preserving speaker-independent attributes of the original speech, including speech quality and prosody, etc.

<https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024>

<https://platform.openai.com/docs/api-reference/introduction>

REFERENCES

- [1] R. V. Cox and J. M. Tribble, “Analog voice privacy systems using TFSP scrambling: Full duplex and half duplex,” *The Bell System Technical Journal*, vol. 62, no. 1, pp. 47–61, 1983.
- [2] F. Fang et al., “Speaker anonymization using x-vector and neural waveform models,” in *Speech Synthesis Workshop*, 2019, pp. 155–160.
- [3] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, and et al, “Introducing the VoicePrivacy initiative,” in *Proc. Interspeech*, 2020, pp. 1693–1697.
- [4] N. Tomashenko et al., “The VoicePrivacy 2022 Challenge evaluation plan,” *arXiv preprint arXiv:2203.12468*, 2022.
- [5] M. Chen et al., “VoiceCloak: Adversarial example enabled voice de-identification with balanced privacy and utility,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 2, pp. 1–21, 2023.
- [6] J. Deng, F. Teng, Y. Chen, X. Chen, Z. Wang, and W. Xu, “V-Cloak: Intelligibility-, naturalness-& timbre-preserving real-time voice anonymization,” in *32nd USENIX Security Symposium*, 2023, pp. 5181–5198.
- [7] R. Wang, L. Chen, K. A. Lee, and Z.-H. Ling, “Asynchronous voice anonymization using adversarial perturbation on speaker embedding,” in *Proc. Interspeech*, 2024, pp. 4443–4447.
- [8] Z. Ju et al., “NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models,” in *International Conference on Machine Learning*, 2024, pp. 22 605–22 623.
- [9] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, “Diff-HierVC: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation,” in *Proc. Interspeech*, 2023, pp. 2283–2287.
- [10] Y. Wang et al., “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *International conference on machine learning*, 2018, pp. 5180–5189.
- [11] X. Zhao et al., “Disentangling content and fine-grained prosody information via hybrid ASR bottleneck features for voice conversion,” in *Proc. ICASSP*, 2022, pp. 7022–7026.
- [12] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, “YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone,” in *International Conference on Machine Learning*, 2022, pp. 2709–2720.
- [13] J. Li, W. Tu, and L. Xiao, “FreeVC: Towards high-quality text-free one-shot voice conversion,” in *Proc. ICASSP*, 2023, pp. 1–5.
- [14] C. Bishop, *Pattern recognition and machine learning*. Springer, New York, 2006.
- [15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: an ASR corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.
- [16] N. Tomashenko et al., “The VoicePrivacy 2024 Challenge evaluation plan,” *arXiv preprint arXiv:2404.02677*, 2024.
- [17] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Proc. Interspeech*, 2020, pp. 3830–3834.
- [18] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, “BUT system description to VoxCeleb speaker recognition challenge 2019,” *arXiv preprint arXiv:1910.12592*, 2019.
- [19] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A large-scale speaker identification dataset,” in *Proc. Interspeech*, 2017, pp. 2616–2620.
- [20] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” in *Proc. Interspeech*, 2018, pp. 1086–1090.
- [21] F. Tong et al., “ASV-Subtools: Open source toolkit for automatic speaker verification,” in *Proc. ICASSP*, 2021, pp. 6184–6188.
- [22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*, 2023, pp. 28 492–28 518.