USTC-PolyU system for the VoicePrivacy 2024 Challenge

*Wenju Gu[1], *Zeyan Liu[1], Liping Chen[1], Rui Wang[1], Chenyang Guo[1], Wu Guo[1], Kong Aik Lee[2], Zhen-Hua Ling[1]

[1]NERC-SLIP, University of Science and Technology of China, China
[2]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong

{wjgu,xy671231}@mail.ustc.edu.cn, lipchen@ustc.edu.cn

Abstract

This paper presents the USTC-PolyU voice anonymization system for the VoicePrivacy 2024 Challenge (VPC2024). Our submission is developed based on the disentanglement of the content and non-content attributes present in speech utterances. Particularly, the content encoder adopts the ASR-BN extractor as used in baseline B5 of VPC2024. Meanwhile, a non-content learner which employs the global style token (GST) strategy, is applied to model the non-content attributes, resulting in non-content embedding. Finally, the waveform is generated using a HiFi-GAN generator by utilizing both the content and non-content representations. The anonymization process is completed by the transfer of non-content attributes from the original speech to an utterance with the same emotion. Two systems are submitted with separate trade-offs between privacy protection and utility.

Index Terms: voice anonymization, information disentanglement, content attributes, non-content attributes

1. Introduction

In recent years, driven by the increase in computing power and data amount, neural networks have gone through rapid development. As a result, remarkable progress has been achieved in speech technologies, including automatic speech recognition (ASR) [1], text-to-speech (TTS) [2], voice conversion (VC) [3], speaker recognition [4, 5], and so on. However, the proliferation of personal speech data online also brought the security risks caused by malicious applications of the voice attributes. Thus, the demand for voice privacy protection of such speech is increasing. For example, using speaker recognition technology, one can now recognize the speaker's identity within a speech utterance with high accuracy. As a result, other personal information conveyed in the identified utterance will be exposed, such as interests, opinions, ages, etc. By utilizing personalized speech generation technologies, e.g., TTS [6] or VC [7], it is possible to produce high-quality synthetic speech of an individual. Malicious uses of such deepfake speech utterances include fraud, damaging the reputation of the victim, etc.

Also empowered by the development of speech generation technology, the voice anonymization technique has regained the interest of the community. Specifically, a voice anonymization method based on speaker replacement was introduced in [8]. Within the framework, the speaker, prosody and content information within the original speech are first disentangled and represented separately. The anonymized speech is generated by replacing the representation of the original speaker with a pseudo-speaker. Based on the framework, efforts have been devoted in pseudo-speaker construction, to name a few [9–12]. In parallel, focuses have been devoted into the information disentanglement within the original speech, where the efficacy of the vector quantization strategy has been proven [13]. In [14], the large speech generation model *Suno*[1] was applied wherein the content and speaker attributes are encoded to one-hot vectors, separately. The method is applied as the *baseline **B4*** in VoicePrivacy 2024 Challenge (VPC2024). In [15], the wav2vec 2.0 model, combined with vector quantization, was applied for content encoding, extracting the ASR-BN features. Furthermore, to mitigate the transfer of speaker attributes from the original utterance to the anonymized version, the pitch values obtained from the original speech were transformed to align with those of the pseudo-speaker. The framework is adopted as the *baseline **B5*** in VPC2024.

Our voice anonymization method is based on the information disentanglement, separating and encoding the content and non-content information of speech. Non-content attributes involve characteristics such as speaker identity, prosody, and emotion, among others. In our method, a waveform construction framework is built upon a content encoder, a non-content learner and a HiFi-GAN generator. Particularly, the content encoder adopts the ASR-BN extractor as used in baseline B5. The non-content attributes modeling is fulfilled with a non-content learner, which incorporates the global style token (GST) strategy. Given an input speech, the non-content learner generates a non-content embedding that captures the specific characteristics of the non-content attributes. The HiFi-GAN generator reconstructs the speech by utilizing both the content and non-content embeddings. Given a well-trained waveform construction model, the anonymization of an original speech is carried out through a four-step process. Firstly, the content encoder is employed to extract the content encoding. Secondly, the emotion is predicted, and an utterance with the same emotion from a pre-defined speech pool is selected as the reference speech for the pseudo-speaker. Thirdly, the non-content embedding is estimated from the reference speech. Lastly, the anonymized speech is generated by utilizing the content encoding from the original utterance and the non-content embedding from the reference speech. Two systems are submitted.

The rest of the paper is organized as follows. Section 2 describes the anonymization framework. The implementation details are provided in Section 3. The experiments are presented in Section 4. Finally, the paper is concluded in Section 5.

2. System description

As shown in Fig. 1, our anonymization framework is composed of four modules: a module for speech emotion recog-

*Those authors contributed equally to this work as first authors.
Corresponding author: *Liping Chen*.

[1]https://suno.com/

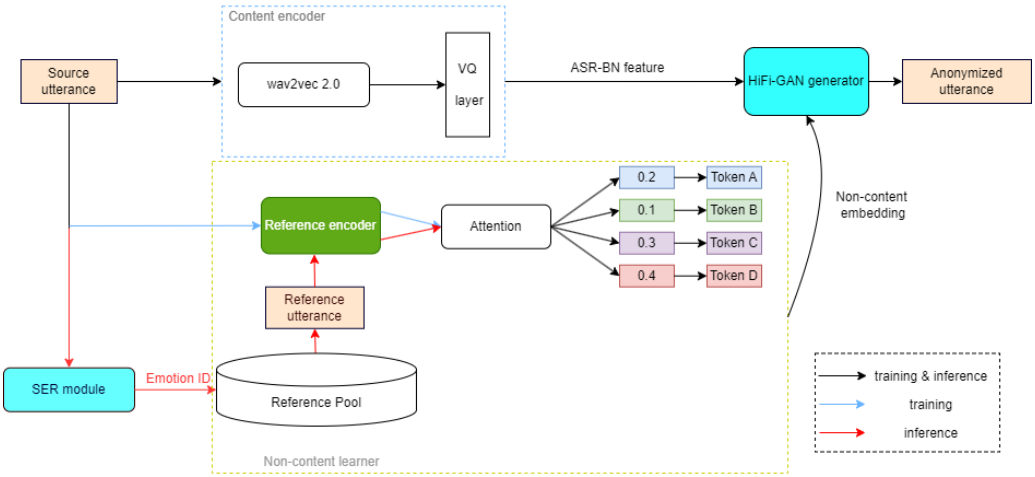

Figure 1: *Diagram of proposed anonymization system. The modules in the rectangular box of yellow dotted lines form the non-content learner. The content encoder is illustrated in the rectangular box of blue dotted lines. The black line is valid in both the training and inference processes. The red and blue lines are only valid in the inference and the training processes, respectively.*

nition (SER), a content encoder, a non-content learner, and a HiFi-GAN generator. Taking an original speech as input, the content information is encoded using the content encoder. Meanwhile, its emotion type is predicted according to the SER module, represented with the emotion ID in Fig. 1. A reference speech with the same emotion is selected from a predetermined pool of reference speech. Subsequently, the non-content information, including speaker characteristics, style, emotion, etc, is estimated using a non-content learner from the reference speech. Finally, the content and reference style representations are sent to the end-to-end HiFi-GAN generator [16], generating the anonymized speech.

2.1. Content encoder

In our work, the content encoder adopted in the baseline **B5** provided by VPC2024 is used for content encoding extraction, referred to as the content encoder in Fig.1. In detail, the pre-trained wav2vec 2.0 is used for speech embedding extraction. Following that, the vector quantization (VQ) technique, described in [15], is applied to discretize the embedding vector. This discretized representation serves as the content representation for the original speech utterance.

2.2. Non-content learner

The concept of style transfer is incorporated into our framework for the non-content attributes modeling where the reference encoder proposed in [17] is adopted. It compresses the prosody of a variable-length audio signal into a fixed-length vector called *reference embedding*. Meanwhile, a bank of vectors which is called global style tokens (GSTs) [18], is randomly initialized. An attention module learns a similarity measure between the reference embedding and each token in the bank, and combines GSTs into the non-content embedding with the learned weights.

2.3. HiFi-GAN generator

To effectively utilize the content features from ASR-BN and the non-content features from GST, the HiFi-GAN [16] is employed with multi-scale discriminators as our decoder. By concatenating the content features from the ASR-BN extractor with the non-content embeddings derived from GST, the combined representation is fed into the HiFi-GAN generator to synthesize high-quality and high-resolution speech.

2.4. SER

The SER module is applied in the anonymization process. It is employed to predict the emotion present in the original speech, whereby a reference utterance matching the same emotion is chosen from a pre-established reference pool.

3. Implementation Details

The details of the training data and the model structure are presented in this section.

3.1. Training data

Our model training data consists of the *LibriTTS train-clean-100* [19] dataset supplemented with English utterances from the *Emotional Speech Dataset* (ESD) [20], which contains a diverse range of expressive styles.

3.2. SER

The SER module classifies among four emotions: angry, happy, neutral, and sad. Among the five folds of pre-trained models provided by the challenge organizers, the SER model of the first fold was used to in our experiments. For the first fold, data from females in session one of the IEMOCAP dataset is designated as the development set, while data from males in the same session is used as the evaluation set. The remaining sessions' data are utilized as the training set.

3.3. Non-content learner

The reference encoder takes the Mel filterbank features with 80 channels as input. It is made up of 10 SE-ResNet layers followed by a GRU layer [21]. The Mel filterbank features first pass through the SE-ResNet layers. Then the GRU layer is used to compress the information. The attention module has 8 heads and 10 GSTs are used in our non-content learner. The dimension of the non-content embedding is 256.

The Non-content learner is trained on LibriTTS train-clean-100 datasets supplemented with English utterances from the ESD with HiFi-GAN generator.LibriTTS train-clean-100 datasets consists of 100 hours of speech recordings from 245 different speakers whice provide high audio quality suitable for speech synthesis. ESD is a multilingual emotional speech database which provides a rich resource for developing and evaluating.

3.4. Model training details

The training setup is detailed as follows:

- **Optimizer:** Adam with a learning rate of 0.001, decayed by 0.998 every epoch
- **Batch Size:** 16
- **Training Steps:** 239,700 steps on three NVIDIA GTX 3090ti GPU

4. Results

Our results of objective evaluation can be affected by the reference utterances in each emotion corpus. Given that over half of the utterances from Libri datasets are labeled as angry instead of neutral, it's necessary to enlarge the angry and neutral corpus because we found that insufficient utterances degrade the EER metric. We filled the angry and neutral corpus with data from LibriTTS train-clean-100 and ESD. However, since LibriTTS data sound not that 'angry', the recognizing accuracy of angry class can suffer degradation after style transfer. Therefore, We provide two sets of evaluation results with different numbers of utterances in angry corpus, which are denoted as large and small. Large corpus contains utterance labeled angry from ESD and LibriTTS100h, while small corpus only have those from ESD. The whole process of evaluation follows the challenge's pipeline.

4.1. Privacy evaluation

Table 1: *EER(%) on the semi-informed scenario for Libri testsets. The results on libri_dev_f, libri_dev_m, libri_test_f and libri_test_m are included.*

dataset	B1	B2	B3	B4	B5	B6	large	small
lbiri_dev_f	10.937	12.910	28.426	34.378	35.816	25.141	45.186	37.925
libri_dev_m	7.454	2.045	22.044	31.056	32.918	20.961	41.482	33.074
libri_test_f	7.474	7.483	27.920	29.378	33.946	21.146	44.006	36.821
libri_test_m	4.675	1.557	26.724	31.155	34.729	21.137	41.417	34.994

Table1 shows our EER results in the semi-informed scenario. Our system outperforms all baseline methods on privacy metric. As we expected, the system works better when larger corpus with more speakers and utterances.

4.2. Utility evaluation

Table 2: *WER(%) on anonymized Libri testsets. The results on the dev and test sets are included.*

dataset	B1	B2	B3	B4	B5	B6	large	small
dev	3.069	10.442	4.288	6.146	4.731	9.693	4.670	5.135
test	2.911	9.954	4.352	5.899	4.369	9.092	4.572	4.972

In Table2, we display WERs in the the ASR tests on the anonymized Libri_dev and Libri_test sets. Our WER results are close to b6 because we use its forepart as our content representation extractor. The model generate speech with more stability when the reference speech has lower emotional intensity. That explain the better performance on the large corpus.

Table 3: *UAR(%) on anonymized IEMOCAP testsets. The results in the dev and test sets are included.*

dataset	B1	B2	B3	B4	B5	B6	large	small
dev	42.71	55.61	38.09	41.97	38.08	36.39	52.38	55.31
test	42.77	53.49	37.57	42.78	38.17	36.13	48.99	55.39

Table3 displays the the unweighted average recall (UAR) for anonymized IEMOCAP test sets. Our method shows superiority over other methods using speech synthesis technique. Due to the evident style of utterances from ESD, speech anonymized with small corpus gets more hard to be misclassified.

5. Conclusions

In this paper, we describe our submitted anonymization system for the VoicePrivacy 2024 Chanllenge. Our experiment results showed that the proposed system achieves a better privacy-utility trade-off compared with the provided baseline methods.

6. 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] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis," *Applied Sciences*, vol. 9, no. 19, p. 4050, 2019.

[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, 2021.

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

[5] 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.

[6] 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*. PMLR, 2022, pp. 2709–2720.

[7] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[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] C. O. Mawalim, K. Galajit, J. Karnjana, and M. Unoki, "X-vector singular value modification and statistical-based decomposition with ensemble regression modeling for speaker anonymization system," in *Proc. Interspeech*, 2020, pp. 1703–1707.

[10] X. Chen, G. Li, H. Huang, W. Zhou, and et al, "System description for Voice Privacy Challenge 2022," in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022.

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

[12] 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 *Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 912–919.

[13] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in *Proc. ICASSP*. IEEE, 2024, pp. 4725–4729.

[15] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," *arXiv preprint arXiv:2308.04455*, 2023.

[16] 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.

[17] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4693–4702.

[18] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[19] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[20] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. ICASSP*. IEEE, 2021, pp. 920–924.

[21] Y. Jiang, L. Chen, H. Zhang, and X. Xiao, "Breast cancer histopathological image classification using convolutional neural networks with small se-resnet module," *PloS one*, vol. 14, no. 3, p. e0214587, 2019.