FaCodec-based Anonymization Solution Enhanced with Prosody Anonymization

Jiabei He[1]*, Jiaming Zhou*[1]*, Haoqin Sun*[1]*, Hui Wang*[1]*, Yong Qin*[1]

[1] College of Computer Science, Nankai University, Tianjin, China

hejiabei@mail.nankai.edu.cn
qinyong@mail.nankai.cn

Abstract

As the concern about voice privacy leakage from speech keeps rising, the Voice Privacy Challenge (VPC) 2024 sets higher demands than VPC 2022, changing anonymization granularity privacy to utterance level and requiring the anonymized speech to retain the original emotional expression. In this paper, we designed an anonymization solution based on FaCodec from NatrualSpeech3 [1]. To improve its privacy performance, a prosody anonymization network based on conditional variational autoencoder (CVAE), PANO, is designed to convert the prosody to the target speaker according to the prompt. PANO prevents identity leakage from the prosodic features of FaCodec, recovers the original emotional expression, and minimizes the interference to the clarity caused by mismatching the prosody and the content features. By cloning the pseudo speaker in timbre and prosody, FaCodec+PANO executes a dual anonymization method more comprehensively. Besides, PANO is low-cost and efficient as it doesn't need joint training and is plug-and-play in FaCodec. Trained on libri-light small, PANO can converge within only 5 epochs, restoring the prosody Mel spectrogram with high similarity. With the closest center distance anonymization strategy, FaCodec+PANO achieves over 51% unweighted average recall (UAR) on both IEMOCAP test/dev, 2.95%/3.02% word error rate (WER) on libri dev/test, and average 33.20% equal error rate (EER) on libri dev/test, ranking 2nd/1st/2nd in emotional expression, content clarity, and privacy performance respectively, achieving the best performance in average rank among baseline systems.

Index Terms: Voice Anonymization, FaCodec, Prosody Anonymization, Voice Privacy Challenge 2024

1. Introduction

With the popularity of Artificial Intelligence Generated Content (AIGC), speech synthesis technology has rapidly developed, but it both protects and threatens voice privacy. Complex scenarios require voice data identity privacy preservation and specific needs. In high-end services, analyzing paralinguistic info helps, but customer identity is sensitive. How to balance identity privacy and voice data usability? VPC 2024 poses a practical and research-valuable challenge.

The requirements of VPC 2024 mainly have the following two differences compared with the previous competition (VPC 2022):

1. The granularity of anonymization changes from the speaker level to the utterance level. The requirements to preserve voice distinctiveness and intonation are removed.
2. In evaluation metrics, apart from EER for privacy and the WER for Automatic Speech Recognition (ASR) used in previous years, the UAR for Speech Emotion Recognition (SER) has been added.

A hypothesis: converting prosody to target style may improve performance. We propose PANO, a lightweight CAVE-based network for FaCodec. PANO reconstructs prosody for anonymization. It uses FaCodec for timbre conversion and PANO for fine-grained prosody conversion. FaCodec+PANO is a dual anonymization framework that synthesizes timbre and prosody, preventing speaker info leakage for thorough anonymization.

The contributions of this paper can be summarized as follows:

1. Prosody anonymization network based on CVAE, PANO, has been proposed for FaCodec to transform the prosody from the original to the target speaker and retain the emotional expression as much as possible. Besides, a gradient reverse layer (GRL) speaker classifier has been applied to guide the prosody encoder extracting speaker-dependent latent prosodic features, further preventing speaker information leakage from prosody features.
2. A dual anonymization framework, FaCodec+PANO has been designed to anonymize the speech more completely in both timbre and prosody, alleviating the conflict between privacy and high utility in voice privacy.
3. The FaCodec+PANO's performances in 3 aspects have been evaluated with VPC 2024 metrics: SER, WER, and EER on dataset IEMOCAP/LibriSpeech/LibriSpeech. With the closest center distance anonymization strategy, FaCodec+PANO ranks 2ed/1st/2ed respectively compared to 7 official solutions and 1st in the average rank.

2. System Overview

The dual anonymization solution comprises two components: one is the native FaCodec, and the other is PANO, specifically designed for FaCodec. Each part is given an overall introduction in this section.

FaCodec has several components: encoder, decoder, timbre extractor, and three VQs (prosody VQ, content VQ, and acoustic detail VQ) as shown in Fig 1. The encoder and decoder handle encoding and decoding original audio to/from latent space features. The timbre extractor gets speaker embeddings. Each VQ extracts corresponding discretized tokenized features. Notably, the input for prosody VQ is from the low-frequency of 80-dim mel-spectrogram, not latent space features. It is the prosody that PANO aims to reconstruct for anonymized speech as shown in Fig 2, and this prosody statement is adopted in the following paper.

PANO includes a prosody encoder, a prosody decoder, a

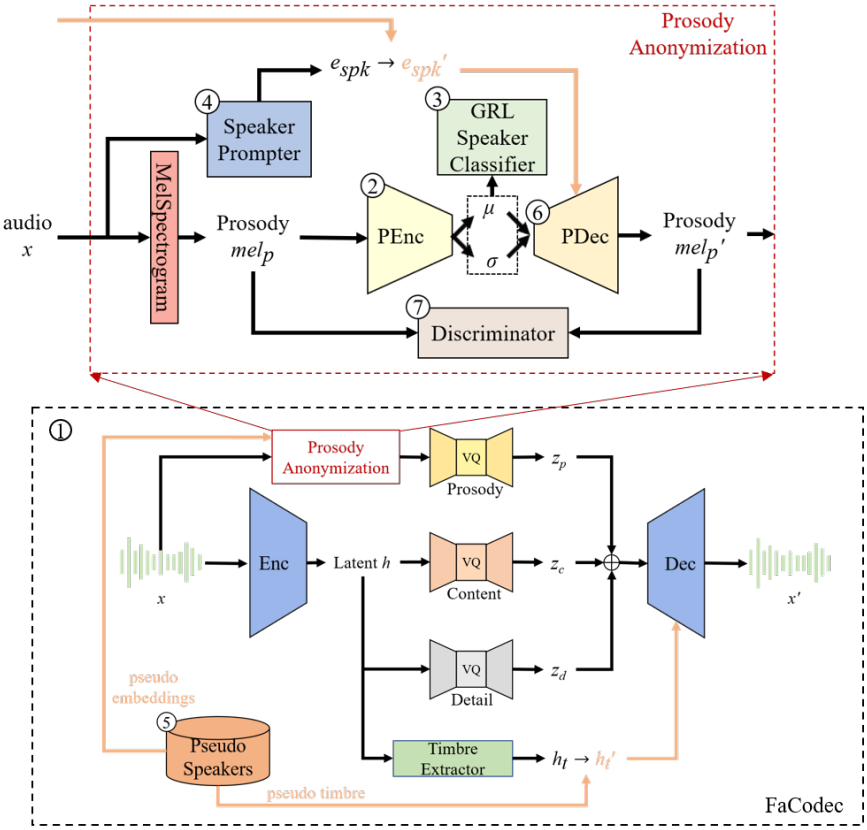

Figure 1: *The Network Architecture of FaCodec+PANO dual Anonymization Solution*

#	Module	Descripion	Output features	Data
1	FaCodec	Codec in NaturalSpeech3 implemented from Amphion `https://huggingface.co/amphion/naturalspeech3_FaCodec`	speech waveform	Libri-Light large
2	Prosody Encoder	Prior Encoder implemented from VITS `https://github.com/jaywalnut310/vits`	latent prosody feature^{192} per time step	Libri-Light small
3	GRL Speaker Classifier	ECAPA-TDNN implemented from speechbrain `https://github.com/speechbrain/speechbrain`	GRL loss	Libri-Light small
4	Speaker Prompter	Pretrained Voice Encoder implemented from resemblyzer `https://github.com/resemble-ai/Resemblyzer`	Speaker embedding^{256}	
5	Speaker Anonymization	Timbre and speaker embeddings extracted by timbre extractor from FaCodec and resemblyzer respectively.	speaker embedding^{256} timbre embedding^{256} of 1166 speakers	train-other-500 from LibriSpeech
6	Prosody Decoder	Prior Encoder implemented from VITS `https://github.com/jaywalnut310/vits`	the low-frequency part of mel spectrogram$^{80}[:, : 20]$	Libri-Light small
7	Discriminator	Multi-kernel discriminator inspired by multi-period discriminator from VITS `https://github.com/jaywalnut310/vits`	discriminator loss	Libri-Light small

Table 1: *Modules and training corpora for FaCodec+PANO. The module indexes are the same as in Figure 1. Superscript numbers represent feature dimensions.*

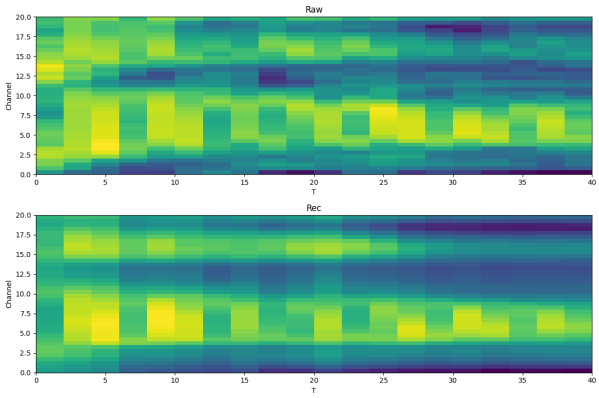

Figure 2: *The reconstruction performance of PANO, Y-axis channels presenting frequency bins and x-axis T presenting time steps. The upper case is raw prosody and the lower case is reconstructed prosody.*

discriminator, a GRL speaker classifier, and a speaker prompter. The prosody encoder extracts 192-dim latent prosody features from 20-dim prosody mel_p. The prosody decoder synthesizes msl'_p with latent prosody features and speaker prompt. The discriminator determines mel'_p authenticity and calculates feature distance. The GRL speaker classifier guides the encoder to extract speaker-independent features.

3. Implementation Details

After introducing the dual anonymization framework in Section 2, the subsequent sections delve into the detailed description of its specific implementation details with the accordingly Table 1. The implementation details are divided into 4 parts: FaCodec, PANO, anonymization strategy, and loss functions.

3.1. FaCodec

FaCodec, from Amphion [2], uses V2 checkpoints of Encoder and Decoder, trained on Libri-Light [3] with 60,000 hours of utterances as in [1]. In this paper, FaCodec needs no additional training and is used for model inference only. Its excellent disentanglement performance trained on the large corpus in supervised learning can avoid info leakage problems. Choosing FaCodec as the prototype is mainly due to its high expressiveness in speech emotion. Also, its prosody handling is more fine-grained and suitable for complex ops and extensional transformation. Before prosody mel_p is tokenized by prosody VQ, PANO must transform the prosody from the original to another pseudo-speaker style.

3.2. PANO

From the overall perspective, prosody synthesis saves more time and computational resources as the prosody of the same duration costs far less storage compared to audio synthesis Trained on Libri-Light small within only 5 epochs, PANO can converge and doesn't need joint training with FaCodec. From the partial perspective, PANO consists of 5 components: a prosody encoder, a GRL speaker classifier, a prosody decoder, a speaker prompter, and a discriminator. Each of them is introduced in the following part of this subsection.

1. **Prosody encoder** is the prior encoder implemented in VITS [4]. It takes raw prosody $mel_p \in \mathbb{R}^{T\times C}, C = 20$ as input, and outputs latent prosodic features $h_p \in \mathbb{R}^{T\times P}, P = 192$.
2. A **GRL speaker classifier**, based on ECAPA-TDNN [5] implemented in speechbrain [6] and AAM loss [7] is adapted to avoid the original speaker's information remaining in h_p, guiding the prosody encoder to extract speaker-independent h_t. It inputs h_p and its outputs one-hot classification results are used for computing AAM loss.
3. **Prosody decoder** uses the same component as the prosody encoder. It inputs the speaker-independent latent prosodic features h_p and a speaker prompt and outputs the generated prosody mel'_p of the target speaker's style.
4. **Speaker prompter** provides 256-dimensional speaker embeddings as prompts for the prosody decoder. The Voice Encoder from resemblyzer [8] plays this role.
5. **Discriminator** helps the prosody decoder to generate more genuine prosody. In our paper, the multi-kernel discriminator adapted here originated from the multi-period discriminator in VITS [4], using multiple kernels convolution to measure the distance between mel_p and mel'_p.

3.3. Loss Functions

Five loss functions are used to guide PANO in achieving the functions above.

1. KL divergence loss aligns the distributions $p(h_p|x), q(h_p)$ from the prosody encoder and the decoder respectively, and regularizes the latent prosody features h_p approximate Gaussian distribution.

$$L_{KL} = \sum_{i\in T}\{-D_{KL}[p(h_{p,i}|x)||q(h_{p,i})]+\mathbb{E}_{q(h_p|x_i)}[ln\, q(x_i|h_{p,i})]\}$$

2. MSE loss is adopted as the reconstruction loss presenting the capability to recover the prosody.

$$L_{rec} = MSE(mel_p, mel'_p)$$

3. GAN-related loss consists of the adversarial loss [9] $L_{adv}(D), L_{adv}(PD)$ for the discriminator and prosody decoder, and the feature matching loss $L_{fm}(PD)$ [10] for the prosody decoder.
4. AAM-softmax loss L_{aam} enhanced helps the prosody encoder extract more speaker-independent features.

The training loss for the PANO can be summarized as:

$$L = L_{KL} + Lrec + L_{adv}(PD) + L_{fm}(PD) + \alpha \cdot L_{aam},$$

where $\alpha = 0.1$ is set to control the GRL speaker loss to the same scale as other losses.

3.4. Anonymization Strategy

As for the anonymization strategy, the description can be divided into 2 parts: the procedures to execute FaCodec+PANO framework and the method to select the pseudo-speakers.

We select one utterance sample for 1166 speakers in train-other-500 from LibriSpeech [11]. From these samples, we extract the timbre features by the timbre extractor in FaCodec and the speaker embeddings using the voice encoder in resemblyzer as speaker prompts. Assuming that the pseudo speaker is chosen, follow the steps to execute anonymization solutions.

1. Encode the source audio with FaCodec encoder and obtain prosody mel_p and latent tokenized feature h.

2. PANO generates the pseudo prosody mel'_p according to the prompt of the pseudo speaker from the speaker prompter. Replace the original prosody mel_p with the pseudo prosody mel'_p.
3. Extract z'_p, z_c, h_t with prosody VQ, content VQ, and timbre extractor in FaCodec. Replace h_t with the pseudo speaker timbre h'_t.
4. Decode (z'_p, z_c, h'_t) with the decoder in FaCodec. The output of the decoder is the anonymized speech.

Both pseudo timbre and pseudo prosody are used for synthesis, leaving less original speaker information in the anonymized speech.

For each dataset, only one pseudo speaker from train-other-500 is selected to anonymize all the speech. The steps of the strategy for choosing the pseudo-speaker are described as follows:

1. Compute the center of all the timbre embeddings $h_t^c = \text{average}(\{h_t\})$ extracted by the timbre extractor in FaCodec.
2. Then, compute the distance between the pseudo timbre embedding from train-other-500 with cosine similarity and find the pseudo-speaker whose timbre embedding is the closest to the center h_t^c. This pseudo-speaker is assigned to be the constant speaker for the anonymization of that dataset.

This method used for FaCodec+PANO dual anonymization solution is named the closest center distance anonymization strategy.

4. Evaluation and Results

The performance of FaCodec+PANO dual anonymization was evaluated in 3 aspects on IEMOCAP and LibriSpeech datasets according to VPC 2024 requirements: emotion expression, content clarity, and privacy protection. Table 2 shows the emotional expressiveness performance of FaCodec+PANO on IEMOCAP, Table 3 shows the content clarity on LibriSpeech's dev-clean/test-clean, and Table 4 details the anonymization performance for privacy protection on LibriSpeech's dev-clean/test-clean.

For emotional expression, FaCodec+PANO's performance on IEMOCAP dev/test sets was second only to B2, showing high preservation in each emotional category of IEMOCAP. Regarding content clarity, FaCodec+PANO was compared with 6 baseline systems on LibriSpeech's dev-clean/test-clean in Table 3. It achieved the best performance in WER on dev-clean and maintained it. Anonymization performance was evaluated on LibriSpeech's dev-clean and test-clean. FaCodec+PANO achieved the best male anonymization performance on dev-clean with an EER of 41.615%. It maintained high performance consistently across all levels, with an average EER of 33.1975%, ranking second only to B5-wav2vec2.

The comprehensive performance of 7 anonymization solutions is shown in Table 5 calculated in two ways and ranked. Ranking by the average rank, FaCodec+PANO achieves the best of all the solutions. Ranking by the weighted rank, if equally 50% for privacy and 50% for utility performance, FaCodec+PANO can also achieve the first.

In summary, FaCodec+PANO has demonstrated highly competitive performance across all metrics evaluated. Comparing its rankings in emotional expression, content clarity, and anonymization performance with various baseline systems, FaCodec+PANO achieved rankings of 2nd/1st/2nd respectively. This comprehensive performance places FaCodec+PANO at the forefront across all evaluated aspects.

5. Conclusion

VPC 2024 introduces more granular and diversified anonymization metrics, focusing on utterance level and emotional expression. Facing the new requirement of emotional expression, officially provided baselines fail to deal with prosody processing effectively, leading to interference with the performance declines in other aspects. This paper proposes the PANO network, an optimized extension of FaCodec tailored to these needs, to convert the prosody to the target speaker's style, ensuring harmony in the following anonymization speech synthesis. The FaCodec+PANO dual anonymization framework utilizes FaCodec's outstanding performance in speech disentanglement, preserving both emotion and content features in the original speech, with PANO network enhancing the anonymization of prosody features. Among the anonymization solutions evaluated with respect to emotion expression, content clarity, and anonymization, FaCodec+PANO ranks 2nd/1st/2nd respectively, reaching the optimum performance in the average rank and weighted rank.

6. References

[1] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang, Z. Wu, T. Qin, X.-Y. Li, W. Ye, S. Zhang, J. Bian, L. He, J. Li, and S. Zhao, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," *ArXiv*, vol. abs/2403.03100, 2024.

[2] X. Zhang, L. Xue, Y. Wang, Y. Gu, X. Chen, Z. Fang, H. Chen, L. Zou, C. Wang, J. Han *et al.*, "Amphion: An open-source audio, music and speech generation toolkit," *arXiv preprint arXiv:2312.09911*, 2023.

[3] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[4] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 5530–5540.

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *ArXiv*, vol. abs/2005.07143, 2020.

[6] M. Ravanelli, T. Parcollet, P. W. V. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. N. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," *ArXiv*, vol. abs/2106.04624, 2021.

[7] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1652–1656.

[8] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.

[9] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

	IEMOCAP dev						IEMOCAP test						
	UAR ↑ mean	sad ↑ mean	neu ↑ mean	ang ↑ mean	hap ↑ mean	rank	UAR ↑ mean	sad ↑ mean	neu ↑ mean	ang ↑ mean	hap ↑ mean	rank	SER rank
original	69.08	63.63	65.97	79.78	66.95	-	71.06	72.57	71.66	72.82	67.19	-	-
B1	42.71	0.26	34.03	**78.88**	57.67	3	42.78	2.78	37.97	**72.51**	57.85	3	3
B2	**55.61**	32.96	**57.97**	64.44	67.09	**1**	**53.49**	32.78	**66.23**	56.97	57.98	**1**	1
B3	38.09	0.73	34.45	70.54	46.63	5	37.57	0.65	41.83	66.09	41.72	6	5
B4-NAS	41.97	9.03	41.88	63.22	53.74	4	42.78	11.26	46.68	61.54	51.64	3	4
B5-wav2vec2	38.08	7.54	49.11	62.05	33.62	6	38.17	5.07	55.3	56.2	36.1	5	5
B6-tdnnf	36.39	2.58	15.25	49.77	**77.96**	7	36.13	1.59	24.49	46.72	**71.71**	7	8
FaCodec+PANO	51.65	**55.09**	55.54	42.62	53.36	2	51.41	**55.5**	62.93	34.74	52.49	2	2

Table 2: *Emotion performance UAR on IEMOCAP processed by 7 anonymization systems*

	libri dev	libri test	avg	WER rank
original	1.81	1.84	1.825	-
B1	3.07	**2.91**	2.99	2
B2	10.44	9.95	10.195	7
B3	4.29	4.35	4.32	3
B4-NAS	6.15	5.9	6.025	5
B5-wav2vec2	4.73	4.37	4.55	4
B6-tdnnf	9.69	9.09	9.39	6
FaCodec+PANO	**2.95**	3.02	**2.985**	**1**

Table 3: *WER↓ performance on anonymized utterances in LibriSpeech processed by 7 anonymization systems*

	libri dev		libri test			
	eer-f ↑	eer-m ↑	eer-f ↑	eer-m ↑	avg	EER rank
original	10.51	0.93	8.76	0.42	5.16	-
B1	10.94	7.45	7.47	4.68	7.64	6
B2	12.91	2.05	7.48	1.56	6	7
B3	28.43	22.04	27.92	26.72	26.28	4
B4-NAS	34.38	31.06	29.38	31.16	31.5	3
B5-wav2vec2	**35.82**	32.92	**33.95**	**34.73**	34.36	**1**
B6-tdnnf	**25.14**	20.96	**21.15**	**21.14**	22.1	5
PANO	31.4	**41.62**	33.76	26.01	33.2	2

Table 4: *EER↑ performance evaluated by ECAPA-TDNN trained on anonymized utterances in LibriSpeech processed by 7 anonymization systems*

	SER rank	**WER rank**	**EER rank**	**AVG rank**	**WTD rank**
B1	3	2	6	2	4
B2	**1**	8	7	6	6
B3	5	4	4	4	4
B4-NAS	4	6	3	4	3
B5-wav2vec2	5	5	**1**	2	2
B6-tdnnf	8	7	5	7	7
FaCodec+PANO	2	**1**	2	1	1

Table 5: *The Average rank and Weighted (25%/25%/50% for SER/WER/EER) rank of 7 anonymization systems)*

[10] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 1558–1566.

[11] 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.