ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890338

Emotion-Preserving Prosody Anonymization Network for Voice Privacy Protection

Jiabei He[1], Shiwan Zhao[1], Jiaming Zhou[1], Haoqin Sun[1], Hui Wang[1], and Yong Qin[1*]
[1]*TMCC, College of Computer Science*, Nankai Unversity, Tianjin, China
Email: hejiabei@mail.nankai.edu.cn

***Abstract*—Balancing emotion preservation and privacy protection in voice anonymization presents a significant challenge, particularly due to the difficulty of effectively handling prosody, a key feature in speech. While preserving prosodic features in anonymized speech enhances emotional expression, it also increases the risk of leaking speaker information. To address this conflict, we propose a lightweight Emotion-Preserving Prosody Anonymization (EPPA) network, which extracts speaker-independent prosodic features to preserve speech emotion while converting them into another speaker's style for anonymization. By combining EPPA with timbre cloning for anonymization while retaining speech content, we achieve a more balanced voice conversion. Evaluated using the Voice Privacy Challenge (VPC) 2024 metrics, our proposed EPPA, utilizing the closest center distance (CCD) anonymization strategy, demonstrates strong performance across emotional expression, content clarity, and privacy protection, achieving the highest ranking in both average and weighted ranks compared to the six baseline solutions.**

***Index Terms*—Voice Anonymization, Emotion Preservation, Prosody Anonymization, Voice Privacy Challenge 2024**

I. Introduction

With the rise of Artificial Intelligence Generated Content (AIGC), speech synthesis technology has rapidly advanced, bringing both innovation and new challenges [1]. A significant concern is the growing threat to voice privacy, as the public becomes increasingly vulnerable to malicious voice cloning. Voice anonymization has emerged as a promising solution, offering a way to protect individuals' original voices through speech-to-speech voice conversion models [2].

Maintaining speech utility, particularly the emotional state, in voice anonymization is critical in applications where high levels of service are essential. In industries such as online healthcare diagnosis and financial consultancy, the ability to analyze paralinguistic information, such as the emotional state conveyed in speech, is vital, enabling a deeper understanding of customer needs and significantly enhancing service quality.

However, balancing emotion preservation and privacy protection in voice anonymization presents a significant trade-off. While retaining prosodic features in anonymized speech enhances emotional expression, it also increases the risk of exposing speakers' identities. To maintain high speech utility, some solutions attempt to preserve the original prosodic features in anonymized speech. For instance, the approach by Fang et al. [3] effectively preserves prosody but performs poorly in privacy protection, as its prosodic features, such as F_0, still contain identifiable speaker information. Similarly, Patino et al. [4] anonymize speech by modifying the formants using the McAdams coefficient [5], a pure speech processing technique, showing good performance in emotional expression. However, this method compromises content clarity and privacy.

Another line of research involves sacrificing utility to improve privacy performance. Panariello et al. [6] employ a transformer decoder to convert concatenated tokenized acoustic features from a neural audio codec [7] and semantic features from HuBERT [8], achieving robust privacy protection. However, experiments reveal poor performance in Speech Emotion Recognition (SER) and Automatic Speech Recognition (ASR), likely due to mismatches between the prosodic features remaining in the semantic tokens and those in the acoustic tokens [9], [10]. Champion et al. [11] utilize wav2vec2 [12] to extract acoustic features, achieving competitive results in privacy performance; however, its performance in SER and ASR remains suboptimal.

Meyer et al. [13] attempt to retain the original emotion while obscuring speaker information by slightly modifying the prosody. However, randomly and manually modifying prosodic features proves ineffective for emotion preservation, and disharmonious prosody can even degrade ASR performance in anonymized speech. Nonetheless, it suggests that modifying prosody could be a promising direction.

Prosody is both context-dependent [14] and speaker-dependent [15], [16], with emotional expression primarily derived from the context-dependent aspects of prosody, as different speakers can convey the same emotions using similar tones. If the original speaker's prosodic style is transformed into that of a target speaker, the emotion in the synthesized speech can be preserved while adopting the target's style [17]. Building on this assumption, this paper introduces a lightweight Emotion-Preserving Prosody Anonymization (EPPA) network, based on a Conditional Variational Autoencoder (CVAE) [18], specifically designed for integration with FACodec from NaturalSpeech3 [19]. EPPA anonymizes prosody by converting it within FACodec to the style of a pseudo speaker. This approach leverages FACodec for speaker timbre conversion and EPPA for more fine-grained prosody conversion. The dual anonymization framework, combining FACodec and EPPA, performs both timbre and prosody synthesis, effectively preserving the original emotion while preventing speaker in-

*Corresponding author. This work was supported by the National Key R&D Program of China (Grant No.2022ZD0116307) and NSF China (Grant No.62271270).

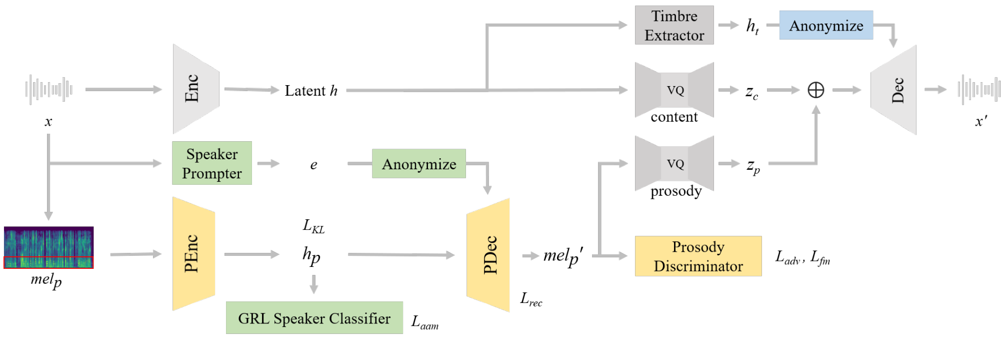

Fig. 1. The architecture of the FACodec+EPPA dual anonymization framework.

formation leakage from prosodic features, ensuring a more comprehensive and robust anonymization.

Our main contributions are summarized as follows:

1) We propose a CVAE-based prosody anonymization network, EPPA, designed for FACodec to convert prosodic style while preserving emotional expression as much as possible.
2) We introduce a dual anonymization framework, FACodec+EPPA, which anonymizes speech more comprehensively in both timbre and prosody, effectively mitigating the conflict between privacy protection and the utility of anonymized speech.
3) We evaluate the performance of our solution across three key aspects—SER, ASR, and privacy protection—using VPC 2024 metrics on the IEMOCAP [20] and LibriSpeech [21] datasets. Our solution achieves first place in both average and weighted ranks, outperforming six existing solutions.

II. System Overview

The dual anonymization framework consists of two main components: FACodec and EPPA. Each component is introduced in detail in this section, as illustrated in Fig. 1.

A. FACodec

FACodec [19] decomposes speech into subspaces representing different attributes and reconstructs high-quality waveforms from these attributes. It consists of an encoder, a decoder, a timbre extractor, and three vector quantizers (VQ): prosody VQ, content VQ, and acoustic detail VQ, represented as gray blocks. The acoustic detail VQ is omitted from the figure as it does not participate in the anonymization process. The encoder and decoder are responsible for encoding the original audio into latent features h and decoding the tokenized features back into audio, respectively. The timbre extractor generates the speaker embedding h_t from the input latent features, while the prosody VQ and content VQ extract their corresponding discretized tokenized features. Notably, in FACodec, prosody mel_p is defined as the low-frequency 20 dimensions of the 80-dimensional mel-spectrogram, serving as the input for the prosody VQ. This definition of prosody is applied throughout this paper.

In our solution, FACodec is used solely for inference and requires no additional training. It was chosen primarily for its exceptional emotional expressiveness and reconstruction capabilities.

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

B. EPPA

The Emotion-Preserving Prosody Anonymization (EPPA) network consists of a prosody encoder, a prosody decoder, a prosody discriminator (PD), a GRL speaker classifier, and a speaker prompter. Each component functions as follows:

1) **Prosody Encoder**: A prior encoder [22] that takes prosody mel_p as input and outputs latent prosodic features h_p.
2) **GRL Speaker Classifier**: Based on the gradient-reversed ECAPA-TDNN [23], adapted from the SpeechBrain classifier [24], this component encourages the prosody encoder to extract speaker-independent h_p, producing one-hot classification results.
3) **Prosody Decoder**: Utilizes the same architecture as the prosody encoder. It takes h_p and a speaker prompt e as inputs and generates reconstructed prosody mel_p'.
4) **Speaker Prompter**: Uses the voice encoder from Resemblyzer [25] to provide speaker embeddings e as prompts for the prosody decoder.
5) **Prosody Discriminator**: A multi-kernel convolutional discriminator, inspired by the multi-period discriminator [22], that helps the prosody decoder generate more realistic prosody by distinguishing between real and generated prosody and measuring the distance between them.

Note that EPPA can converge within 5 epochs of training on the Libri-Light [26] small dataset and does not require joint training with FACodec.

C. Loss Functions

Five loss functions, categorized into three groups, are used to train EPPA:

1) **VAE-related Loss**: This includes the KL divergence loss and the MSE loss. The KL divergence loss aligns the distributions $p(h_p|x)$ from the prosody encoder and

$q(h_p)$ from the decoder, regularizing the latent prosody features h_p to approximate a Gaussian distribution.

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

where D_{KL} is the KL divergence, and $\mathbb{E}_q$ is the expectation with respect to distribution q.

The MSE loss is adopted as the reconstruction loss, representing the model's ability to recover the prosody:

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

2) **GAN-related Loss**: This includes the adversarial loss $L_{adv}(PD)$ [27] and the feature matching loss $L_{fm}(PD)$ [28], both output by the discriminator to encourage the prosody decoder to generate more realistic prosody.
3) **AAM-Softmax Loss**: The AAM-Softmax loss L_{aam} [29] of the GRL Speaker Classifier helps the prosody encoder extract more speaker-independent features.

The training loss of EPPA is summarized as:

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

where $\alpha = 0.1$ is set to scale the GRL speaker loss to match the magnitude of other losses.

D. Anonymization Strategy

The anonymization strategy is explained in two parts: first, the method for selecting pseudo-speakers, and second, the procedures for executing our dual anonymization framework.

Firstly, we build a pseudo-speaker pool consisting of 1,166 speakers from the train-other-500 subset of LibriSpeech [21]. For each speaker, one utterance is selected, and timbre features are extracted using the timbre extractor in FACodec, while speaker embeddings are obtained via the speaker prompter. For each dataset to be anonymized, a single pseudo-speaker is chosen from the pool to anonymize all speech in that dataset.

The steps for selecting the pseudo-speaker are as follows:

1) For the dataset to be anonymized, compute the center of all timbre embeddings, $h_t^c = \text{average}(h_t)$, extracted by the timbre extractor in FACodec.
2) Compute the distance between each pseudo-speaker's timbre embedding and the center h_t^c using cosine similarity. The pseudo-speaker with the timbre embedding closest to the center is selected as the constant speaker for anonymizing the entire dataset.

This method for selecting pseudo-speakers is referred to as the Closest Center Distance (CCD) anonymization strategy.

Assuming the pseudo-speaker has been selected, follow these steps to execute the anonymization solution:

1) Encode the source audio using the FACodec encoder to obtain the prosody mel_p and latent tokenized feature h.
2) EPPA generates the pseudo prosody mel'_p based on 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) using the prosody VQ, content VQ, and timbre extractor in FACodec. Replace h_t with the pseudo-speaker's timbre h'_t.
4) Decode (z'_p, z_c, h'_t) using the FACodec decoder. The output of the decoder is the anonymized speech.

By utilizing both pseudo timbre and pseudo prosody for synthesis, the original speaker's information is effectively minimized in the anonymized speech while preserving the emotional state.

III. Evaluation and Results

The performance of the FACodec+EPPA dual anonymization framework is evaluated according to VPC 2024 requirements using the IEMOCAP and LibriSpeech datasets across three aspects: emotion expression (Table I), content clarity (Table II), and privacy protection (Table III) with three corresponding metrics unweighted average recall (UAR), word error rate (WER), and equal error rate (EER) respectively. Among the VPC 2024 baselines, B1 [3] and B2 [4] sacrifice privacy for better utility, while B4 [6] and B5 [11] take the opposite approach, prioritizing privacy over utility. Unfortunately, B3 [13] and B6 [11] fail to achieve strong performance in all evaluation aspects.

In the experiments, FACodecused in our framework is from Amphion [30]. To further validate the effectiveness of prosody discrimination [31] in EPPA, two versions of EPPA are trained: one with prosody discrimination (denoted as ours) and one without (denoted as ours-PD, where "-" indicates the removal of PD). In the evaluation, **boldface** and <u>underline</u> are used to denote the 1^{st} and 2^{nd} best performances, respectively, in each table.

TABLE I
SER PERFORMANCE EVALUATED BY UAR↑(%) ON ANONYMIZED IEMOCAP

	dev	test	avg
original	69.08	71.06	70.07
B1	42.71	42.78	42.75
B2	**55.61**	**53.49**	**54.55**
B3	38.09	37.57	37.83
B4	41.97	42.78	42.38
B5	38.08	38.17	38.13
B6	36.39	36.13	36.26
Ours-PD	48.99	46.40	47.70
Ours	<u>51.65</u>	<u>51.41</u>	<u>51.53</u>

A. Emotion Performance

As shown in Table I, our approach (with PD) achieves the second-best performance, surpassed only by B2, demonstrating the strong emotion preservation capabilities of EPPA. The version without the PD (Ours-PD) ranks 3^{rd} in SER, maintaining emotional expression at a relatively usable level. However, without the PD supervision, the reconstruction ability of EPPA decreases slightly, resulting in a reduction in emotion preservation.

https://github.com/open-mmlab/Amphion/tree/main/models/codec/ns3_codec

TABLE II
ASR PERFORMANCE EVALUATED BY WER↓(%) ON ANONYMIZED LIBRISPEECH

	dev	test	avg
original	1.81	1.84	1.825
B1	3.07	**2.91**	2.99
B2	10.44	9.95	10.20
B3	4.29	4.35	4.32
B4	6.15	5.90	6.025
B5	4.73	4.37	4.55
B6	9.69	9.09	9.39
Ours-PD	3.54	3.30	3.42
Ours	**2.95**	3.02	**2.99**

B. Content Performance

In terms of content clarity (Table II), our approach (with PD) achieves the best performance, while Ours-PD ranks 3^{rd} in average WER, thanks to FACodec's strong disentanglement capability, which helps maintain clear linguistic content. The comparison between Ours and Ours-PD in WER highlights that rougher prosody recovery can interfere with content clarity in anonymized speech.

TABLE III
PRIVACY PERFORMANCE EVALUATED BY EER↑(%) ON ANONYMIZED LIBRISPEECH. FOR EACH SOLUTION, THE EVALUATION SPEAKER MODEL ECAPA-TDNN HAS TRAINED ON ITS ANONYMIZED TRAIN-CLEAN-360 FROM LIBRISPEECH TO RECOGNIZE THE ORIGINAL SPEAKER.

	libri dev		libri test		
	EER-f ↑	EER-m ↑	EER-f ↑	EER-m ↑	EER-avg ↑
original	10.51	0.93	8.76	0.42	5.16
B1	10.94	7.45	7.47	4.68	7.64
B2	12.91	2.05	7.48	1.56	6.00
B3	28.43	22.04	27.92	26.72	26.28
B4	34.38	31.06	29.38	31.16	31.50
B5	35.82	32.92	33.95	**34.73**	34.36
B6	25.14	20.96	21.15	21.14	22.10
Ours-PD	**38.35**	**45.34**	**40.69**	16.48	**35.22**
Ours	31.40	41.62	33.76	26.01	33.20

C. Anonymization Performance

As shown in Table III, Ours-PD achieves the best performance with an average EER of 35.22%. Ours ranks 3^{rd} but still maintains a competitive average EER of 33.20%, demonstrating strong anonymization capabilities.

TABLE IV
THE AVERAGE (AVG) RANK AND WEIGHTED (WTD, 25%/25%/50% FOR UAR/WER/EER) RANK OF 8 ANONYMIZATION SOLUTIONS

	SER rank	**WER rank**	**EER rank**	**AVG rank**	**WTD rank**
B1	4	2	7	3	5
B2	**1**	8	8	7	7
B3	7	4	5	6	6
B4	5	6	4	5	4
B5	6	5	2	3	3
B6	8	7	6	8	8
Ours-PD	3	3	**1**	2	**1**
Ours	2	**1**	3	**1**	2

D. Overall Performance

The comprehensive performance of the eight anonymization solutions is ranked in two ways, as shown in Table IV. Both Ours and Ours-PD achieve the highest rankings in average and weighted ranks, with the weighted rank assigning 50% to privacy and 50% to utility (25% for emotion expression and 25% for content clarity). Rather than trading off between these metrics, EPPA maintains relatively high speech utility while effectively protecting privacy.

In summary, FACodec+EPPA has demonstrated highly competitive performance across all three evaluated aspects. Compared to the six baseline solutions, two versions of FACodec+EPPA achieve top-3 out of 8 rankings in all categories, further solidifying its strong overall performance and positioning it at the forefront of anonymization technologies.

TABLE V
THE EMOTION, CLARITY, AND ANONYMIZATION PERFORMANCE OF ABLATION STUDY

Model	Strategy	UAR ↑		WER ↓		EER ↑	
		dev	test	dev	test	dev	test
FACodec	constant	55.64	**59.16**	**2.59**	2.47	11.70	6.47
Ours	constant	48.09	51.19	3.43	3.22	10.80	7.61
FACodec	CCD	**59.14**	58.74	**2.59**	**2.46**	31.76	17.41
Ours	CCD	51.65	51.41	2.95	3.02	**36.51**	**29.89**

IV. ABLATION STUDY

The ablation study is conducted to demonstrate the effectiveness of EPPA and the CCD strategy by comparing network architectures and anonymization strategies in four combinations, as shown in Table V. Since the use of a constant pseudo-speaker for each dataset is allowed in VPC 2024, we use this as the comparison strategy against the CCD strategy.

The CCD strategy improves performance across nearly all aspects for both FACodec and FACodec+EPPA. A likely explanation is that selecting a pseudo-speaker closer to the average timbre of the dataset helps generate speech with higher speaker similarity and naturalness.

While EPPA's improvement in EER is not significant when using the constant speaker strategy, it becomes much more noticeable with the CCD strategy. This highlights the contributions of both EPPA and the CCD strategy to the dual anonymization framework.

V. CONCLUSION

This paper introduces the EPPA network, which extracts speaker-independent prosodic features to preserve speech emotion while converting them into another speaker's style for anonymization. By integrating EPPA with timbre cloning, the FACodec+EPPA dual anonymization framework preserves both emotional and content features in the original speech while effectively protecting voice privacy.

In evaluations of emotion expression, content clarity, and anonymization, our proposed FACodec+EPPA framework achieves the highest ranking in both average and weighted ranks, demonstrating its strong overall performance and its ability to balance privacy protection with high speech utility.

References

[1] Danhuai Guo, Huixuan Chen, Ruoling Wu, and Yangang Wang, "Aigc challenges and opportunities related to public safety: a case study of chatgpt," *Journal of Safety Science and Resilience*, vol. 4, no. 4, pp. 329–339, 2023.

[2] Brij Mohan Lal Srivastava, *Speaker anonymization: representation, evaluation and formal guarantees*, Ph.D. thesis, Université de Lille, 2021.

[3] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre, "Speaker anonymization using x-vector and neural waveform models," *arXiv preprint arXiv:1905.13561*, 2019.

[4] Jose Patino, Natalia Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans, "Speaker anonymisation using the mcadams coefficient," *arXiv preprint arXiv:2011.01130*, 2020.

[5] Stephen Edward McAdams, *Spectral fusion, spectral parsing and the formation of auditory images*, Stanford university, 1984.

[6] Michele Panariello, Francesco Nespoli, Massimiliano Todisco, and Nicholas Evans, "Speaker anonymization using neural audio codec language models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4725–4729.

[7] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al., "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[8] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[9] Jiaming Zhou, Shiwan Zhao, Ning Jiang, Guoqing Zhao, and Yong Qin, "Madi: Inter-domain matching and intra-domain discrimination for cross-domain speech recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[10] Jiaming Zhou, Shiwan Zhao, Yaqi Liu, Wenjia Zeng, Yong Chen, and Yong Qin, "Knn-ctc: Enhancing asr via retrieval of ctc pseudo labels," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11006–11010.

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

[12] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael 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.

[13] Sarina Meyer, Florian Lux, Julia Koch, Pavel Denisov, Pascal Tilli, and Ngoc Thang Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] Max Morrison, Lucas Rencker, Zeyu Jin, Nicholas J Bryan, Juan-Pablo Caceres, and Bryan Pardo, "Context-aware prosody correction for text-based speech editing," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7038–7042.

[15] Berrak Sisman and Haizhou Li, "Wavelet analysis of speaker dependent and independent prosody for voice conversion.," in *Interspeech*, 2018, pp. 52–56.

[16] Kun Zhou, Berrak Sisman, Mingyang Zhang, and Haizhou Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion," in *Interspeech*, 2020.

[17] Berrak Sisman, Haizhou Li, and Kay Chen Tan, "Transformation of prosody in voice conversion," *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1537–1546, 2017.

[18] Kihyuk Sohn, Honglak Lee, and Xinchen Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[19] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," *ArXiv*, vol. abs/2403.03100, 2024.

[20] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev 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.

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

[23] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *ArXiv*, vol. abs/2005.07143, 2020.

[24] Mirco Ravanelli, Titouan Parcollet, Peter William VanHarn Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, E. N. Rastorgueva, Franccois Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio, "Speechbrain: A general-purpose speech toolkit," *ArXiv*, vol. abs/2106.04624, 2021.

[25] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez 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.

[26] Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian 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.

[27] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[28] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 1558–1566.

[29] Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai 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.

[30] Xueyao Zhang, Liumeng Xue, Yuancheng Wang, Yicheng Gu, Xi Chen, Zihao Fang, Haopeng Chen, Lexiao Zou, Chaoren Wang, Jun Han, et al., "Amphion: An open-source audio, music and speech generation toolkit," *arXiv preprint arXiv:2312.09911*, 2023.

[31] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua, "Cvae-gan: fine-grained image generation through asymmetric training," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2745–2754.