# **NPU-NTU System for Voice Privacy 2024 Challenge**

Jixun Yao<sup>1</sup>, Nikita Kuzmin<sup>2</sup>, Qing Wang<sup>1</sup>, Pengcheng Guo<sup>1</sup>, Ziqian Ning<sup>1</sup>, Dake Guo<sup>1</sup>, Kong Aik Lee<sup>3</sup>, Eng-Siong Chng<sup>2</sup>, Lei Xie<sup>1</sup>

> <sup>1</sup>Audio, Speech and Language Processing Group (ASLP@NPU) School of Computer Science, Northwestern Polytechnical University <sup>2</sup>Nanyang Technological University <sup>3</sup>The Hong Kong Polytechnic University

### Abstract

Speaker anonymization is an effective privacy protection solution that conceals the speaker's identity while preserving the linguistic content and paralinguistic information of the original speech. To establish a fair benchmark and facilitate comparison of speaker anonymization systems, the VoicePrivacy Challenge (VPC) was held in 2020 and 2022, with a new edition planned for 2024. In this paper, we describe our proposed speaker anonymization system for VPC 2024. Our system employs a disentangled neural codec architecture and a serial disentanglement strategy to gradually disentangle the global speaker identity and time-variant linguistic content and paralinguistic information. We introduce multiple distillation methods to disentangle linguistic content, speaker identity, and emotion. These methods include semantic distillation, supervised speaker distillation, and frame-level emotion distillation. Based on these distillations, we anonymize the original speaker identity using a weighted sum of a set of candidate speaker identities and a randomly generated speaker identity. Experimental results demonstrate that our proposed system outperforms all VPC 2024 baseline systems in privacy protection and paralinguistic preservation.

Index Terms: speaker anonymization, voice privacy 2024 challenge, voice conversion

# 1. Introduction

Speech data on the Internet have proliferated exponentially due to the advent of social media, which encapsulates a wealth of sensitive personal information such as the speaker's identity, age, gender, health status, personality, racial or ethnic origin, geographical background, social identity, and socio-economic status. This sensitive information can be recognized by speaker identification [1], pathological condition detection [2], or other speech attribute recognition systems. With new regulations like the General Data Protection Regulation (GDPR) in the European Union [3], the privacy protection of personal data has gained more attention.

Speaker anonymization serves as a proactive privacy protection solution, implemented before users share their speech data. Its fundamental objective is to effectively remove a speaker's identity while preserving the para-linguistic information of the original speech. To establish a fair benchmark and make speaker anonymization systems comparable, the VoicePrivacy Challenge (VPC) was introduced by the speech community and held in 2020 and 2022 [4–6]. Previous VPC clearly defines the speaker anonymization task, benchmark, metrics, and datasets to drive the development and innovation of techniques dedicated to preserving the privacy of speech data.

Unlike previous VPC editions, the third edition focuses on

preserving the emotional state, a key paralinguistic attribute in many real-world voice anonymization scenarios, such as call centers using third-party speech analytics [7]. To evaluate the performance of emotion preservation, the unweighted average recall (UAR) for speech emotion recognition (SER) replaces the previous voice distinctiveness and intonation metrics. Pre-trained models for utility evaluation (linguistic content and emotion) are trained on original data rather than anonymized data to ensure linguistic and emotional content remains undistorted. Additionally, all data are anonymized at the utterance level. The utterance-level anonymization means that the voice anonymization system must assign a pseudo-speaker to each utterance independently of the other utterances [7].

In this paper, we propose a speaker anonymization system based on a disentangled neural codec. We introduce a serial disentanglement strategy to perform step-by-step disentangling from a global time-invariant representation (speaker identity) to a temporal time-variant representation (linguistic content and fundamental frequency). Since the duration of anonymized speech does not change after anonymization, we assume that the emotional characteristics correspond closely to the fundamental frequency. Therefore, we introduce frame-level emotion distillation to disentangle the emotion-related representation, enhancing emotion preservation during the anonymization process. Additionally, we employ a semantic teacher and selfsupervised speaker distillation to disentangle linguistic content and speaker identity information. During the anonymization process, the anonymized speaker identity is the weighted sum of a set of candidate speaker identities and a randomly generated speaker identity. Experimental results on VPC 2024 Challenge datasets demonstrate that our proposed system effectively protects speaker identity while maintaining the original linguistic content and paralinguistic information.

# 2. Related Works

The VPC series has shown that most current speaker anonymization systems can be broadly categorized into two classes: (1) signal-processing-based approaches [8, 9] and (2) neural voice conversion based approaches [10–13]. We provide an overview of these two classes of speaker anonymization approaches in this section.

#### 2.1. Signal-processing Based Speaker Anonymization

Most signal-processing-based approaches do not require training data and directly manipulate instantaneous speech characteristics. The current mainstream method focuses on adjusting formant frequency and F0. A typical approach uses McAdams coefficients to introduce randomized shifts to the formant frequencies [8], effectively distorting the spectral enve-



Figure 1: An overview architecture of our proposed speaker anonymization system. The pre-trained WavLM model is frozen during training and the red dashed lines are only used in the training process.

lope to anonymize the original speaker's identity. Building on this McAdams coefficients approach, Gupta *et al.* further extends the distortion of the spectral envelope by amplifying the width of formant peaks [9].

#### 2.2. Neural Voice Conversion Based Speaker Anonymization

Neural voice conversion-based speaker anonymization generally outperforms signal-processing-based approaches in terms of privacy and utility metrics. A typical approach involves using a pre-trained ASV model to extract speaker identity representations [10], like x-vectors or d-vectors, while an ASR model extracts linguistic content representations. The original speaker representation is then replaced by averaging a set of candidate speaker representations from an external pool [11, 12]. Candidates are selected based on cosine distance to ensure the farthest representation, protecting privacy. Finally, the averaged speaker representation, linguistic content, and F0 are processed by the neural voice conversion model to generate an anonymized mel-spectrogram. A vocoder then converts the anonymized mel-spectrogram into anonymized speech.

# 3. Proposed System

# 3.1. Overview

As illustrated in Figure 1, our proposed speaker anonymization system utilizes an auto-encoder architecture, which contains a speech encoder, a speaker encoder, a decoder and a residual bottleneck module. The speech encoder converts the speech samples into frame-level representations including linguistic content, speaker identity, and emotional state. We then serially disentangle each factor, ultimately reconstructing the original waveform from the disentangled components. Specifically, the speaker identity is extracted by the speaker encoder, while the residual bottleneck module disentangles the linguistic content and emotional state. Each disentangled factor is concatenated and used to reconstruct the original speech waveform through the decoder.

During inference, we first disentangle the linguistic content and emotional state from the speech waveform. We then average a set of randomly selected speaker identities and combine the averaged speaker identity with a randomly generated speaker identity to obtain the final anonymized speaker identity. The decoder then combines the original linguistic content and emotional state with anonymized speaker identity to reconstruct the final anonymized speech waveform.

# 3.2. Factor Distillation

In this section, we will introduce the disentanglement details and three key distillation methods.

Speaker distillation. The speaker identity extraction is a crucial step in achieving speaker anonymization. The speaker's identity is a voice's inherent and time-invariant characteristic. We first transform the speech waveform to a mel-spectrogram and feed it into our speaker encoder to produce a global representation  $s \in \mathbb{R}^d$ , while d represents the dimension of the speaker representation. To ensure the speaker encoder captures speaker identity information rather than other time-invariant information, we employ two speaker distillation losses to constrain the speaker encoder. We randomly sample two segments from the same utterance and produce two global representations with the speaker encoder. Since the speaker's identity is time-invariant, representations extracted from different segments should be similar and belong to the same speaker identity. We use cosine similarity and an additional speaker classifier with explicit labels to align the extracted global representation with the speaker's identity. The overall speaker distillation loss  $\mathcal{L}_{spk}$  can be defined as follows:

$$\mathcal{L}_{\text{spk}} = \mathbb{E}[-log(C(I \mid s_*))] - \cos(s_1, s_2), \tag{1}$$

where  $s_1$  and  $s_2$  represent the global representation extracted from two segments, respectively. I is the speaker identity label and  $\cos(\cdot)$  denotes cosine similarity.  $C(\cdot)$  denote a speaker identity classifier tasked with determining whether  $s_*$  is associated with the corresponding speaker identity. Finally, we obtain the speaker identity representation and can further disentangle the speaker identity from the output of the speach encoder by direct subtraction. Thus, we obtain the speaker-independent representation  $r_1$ .

**Linguistic distillation**. We utilize a residual bottleneck module for hierarchical disentanglement to further disentangle

linguistic content and emotion state from  $r_1$ . The residual bottleneck module consists of N Vector Quantization (VQ) layers [14], with each VQ layer cascading in a residual manner. Each quantizer contains only residual information from the preceding quantizer, making it inherently suitable for serial disentanglement. Therefore, we distill the first quantizer with a linguistic teacher. We extract the 6th layer's output from the pre-trained WavLM model [15] and use a K-means cluster to transfer WavLM's hidden features into discrete tokens, serving as the linguistic teacher. The linguistic distillation loss  $\mathcal{L}_{\text{lin}}$  can be described as follows:

$$\mathcal{L}_{\text{lin}} = \mathbb{E}[-\log(K(w) \mid q_1))], \tag{2}$$

where  $k(\cdot)$  denote as K-means cluster,  $q_1$  and w represent the quantized token from the first quantizer and WavLM's hidden feature, respectively. The linguistic distillation process is inspired by SpeechTokenizer [16].

**Emotion distillation**. After aligning the first quantizer to linguistic content, we further constrain the residual quantizer with the fundamental frequency (F0). Since the duration of the anonymized speech does not change after anonymization, we believe the emotional state corresponds closely to the F0. Therefore, we directly use F0 to distill the residual quantizer as follows:

$$\mathcal{L}_{\text{emo}} = \cos(f, \operatorname{Proj}(q_2)), \tag{3}$$

where  $Proj(\cdot)$  represent the linear projection. f and  $q_2$  represent the F0 and quantized embedding of second quantizer.

#### 3.3. Training Objectives

The training objective of our proposed anonymization system includes a reconstruction task with adversarial training and three distillation tasks. The reconstruction loss consists of both time and frequency domain losses. The total reconstruction losses are shown as follows:

$$\mathcal{L}_{\text{rec}} = \|X - \hat{X}\|_{1} + \|\text{mel}(X) - \text{mel}(\hat{X})\|_{1} + \|\text{mel}(X) - \text{mel}(\hat{X})\|_{2}, \quad (4)$$

where X and  $\hat{X}$  represent the original waveform and reconstructed waveform, respectively. Meanwhile, mel(·) is an 80bin mel-spectrogram process using a short-time Fourier transform (STFT). For adversarial loss  $\mathcal{L}_{adv}$ , we follow the same configuration with HiFi-GAN. Similar to the conventional VQbased models, we employ a straight-through estimator to optimize the commitment loss between the input feature and quantized feature as follows:

$$\mathcal{L}_{\rm com} = \sum_{i=1}^{N} ||x_i - q_i||_2^2,$$
(5)

where *i* represents the number of quantizer layers.

Generally, our proposed anonymization system is optimized by the mixture of the following losses:

$$\mathcal{L} = \lambda_r \mathcal{L}_{\text{rec}} + \lambda_a \mathcal{L}_{\text{adv}} + \lambda_c \mathcal{L}_{\text{com}} + \lambda_s \mathcal{L}_{\text{spk}} + \lambda_m \mathcal{L}_{\text{sem}}, + \lambda_e \mathcal{L}_{\text{emo}},$$
(6)

where  $\lambda_r$ ,  $\lambda_a$ ,  $\lambda_c$ ,  $\lambda_s$ ,  $\lambda_m$  and  $\lambda_e$  are hyper-parameters used to balance each loss term.

#### 3.4. Anonymization Strategy

Our proposed anonymization strategy involves combining the averaged speaker identity and a randomly generated speaker identity using weighted sums. To generate the averaged speaker identity, we randomly select a set of speaker identities from the speaker vector pool, following the same configuration as the conventional VPC baseline setting [4,6]. Additionally, we sample a random speaker identity from a Gaussian distribution. The original speaker identity is anonymized as follows:

$$s_{\text{anon}} = \alpha \bar{s} + (1 - \alpha) \hat{s},\tag{7}$$

where  $\alpha$  denotes the weight parameter,  $\bar{s}$  and  $\hat{s}$  represent the averaged and randomly sampled speaker identities, respectively.

# 4. Experiments Setup

#### 4.1. Datasets

For model training, our anonymization system is trained on both the LibriSpeech [17] and LibriTTS [18] datasets. For evaluation, we use LibriSpeech-dev-clean and LibriSpeech-test-clean for privacy and utility evaluation, while the IEMOCAP [19] development and evaluation sets are used for emotion evaluation.

#### 4.2. Metrics

The equal error rate (EER) is used as the privacy metric, and the word error rate (WER) and unweighted average recall (UAR) are used as utility metrics. A higher EER indicates better privacy protection, a lower WER indicates better intelligibility, and a higher UAR indicates better emotion preservation. All evaluation setups follow the same configuration as the VPC 2024 pipelines [7].

# 4.3. Configuration

The speech encoder is structured with four convolution blocks, each integrating a residual unit followed by a down-sampling layer. The number of channels doubles during downsampling, and the strides for the four convolution blocks are set as (2, 4, 5, 8). After the convolution blocks, there is a two-layer LSTM for sequence modeling and a concluding 1D convolution layer with a kernel size of 7 and 512 output channels. The architecture of the decoder mirrors the encoder, utilizing transposed convolutions in place of stride convolutions, with the strides in the reverse order of those in the encoder. The speaker encoder and the residual bottleneck module follow the same architecture in [20] and [21]. We employ 8 quantizers in the residual bottleneck module.

For model training, we use the AdamW optimizer with parameters  $\beta_1 = 0.8$ ,  $\beta_2 = 0.99$ , and weight decay  $\lambda = 1e-5$ . The learning rate decay followed a schedule with a decay factor of 0.99 per epoch, starting from an initial learning rate of  $2 \times 10^{-4}$ . The training process comprised a total of 50k steps, utilizing 4 NVIDIA 3090 GPUs with a batch size of 64 utterances. For the hyper-parameters of training loss, we empirical set  $\lambda_r = 45$ ,  $\lambda_a = 1$ ,  $\lambda_c = 0.1$ ,  $\lambda_s = 1$ ,  $\lambda_m = 1$  and  $\lambda_e = 1$ . Regarding the anonymization process, the speaker identity pool is selected from the VCTK dataset. We randomly select 20 speakers and set  $\alpha = 0.9$  to meet the EER threshold for condition 3, while condition 4 employs  $\alpha = 0.8$ .

Table 1: Averaged results over baseline systems and our proposed anonymization system on VPC 2024 development and test datasets. F and M represent the female and male, respectively. Avg denotes average results between female and male EER results. B\* denotes different baseline systems described in VPC 2024 and C\* represents the different EER conditions.

|       | EER, % (†)      |       |       |                  |       |       | WER, % (↓)      |                  | UAR, % (†)      |                  |
|-------|-----------------|-------|-------|------------------|-------|-------|-----------------|------------------|-----------------|------------------|
|       | LibriSpeech-dev |       |       | LibriSpeech-test |       |       | LibriSpeech-dev | LibriSpeech_test | LibriSpeech-dev | LibriSpeech-test |
|       | F               | Μ     | Avg   | F                | Μ     | Avg   | Lionopeen-dev   | Lionspecentest   | Libitopeeen-dev | Lionopeeen-test  |
| Orig. | 10.51           | 0.93  | 5.72  | 8.76             | 0.42  | 4.59  | 1.80            | 1.85             | 69.08           | 71.06            |
| B1    | 10.94           | 7.45  | 9.20  | 7.47             | 4.68  | 6.07  | 3.07            | 2.91             | 42.71           | 42.78            |
| B2    | 12.91           | 2.05  | 7.48  | 7.48             | 1.56  | 4.52  | 10.44           | 9.95             | 55.61           | 53.49            |
| B3    | 28.43           | 22.04 | 25.24 | 27.92            | 26.72 | 27.32 | 4.29            | 4.35             | 38.09           | 37.57            |
| B4    | 34.37           | 31.06 | 32.71 | 29.37            | 31.16 | 30.26 | 6.15            | 5.90             | 41.97           | 42.78            |
| B5    | 35.82           | 32.92 | 34.37 | 33.95            | 34.73 | 34.34 | 4.73            | 4.37             | 38.08           | 38.17            |
| B6    | 25.14           | 20.96 | 23.05 | 21.15            | 21.14 | 21.14 | 9.69            | 9.09             | 36.39           | 36.13            |
| C3    | 44.18           | 31.20 | 37.69 | 37.96            | 36.03 | 36.99 | 2.56            | 2.66             | 65.98           | 64.48            |
| C4    | 45.31           | 39.60 | 42.45 | 40.66            | 40.26 | 40.46 | 3.51            | 3.19             | 62.93           | 60.87            |

# 5. Experimental Results

#### 5.1. Privacy Protection

As shown in Table 1, our proposed anonymization system achieves 37.69% and 36.99% on dev and test datasets in condition 3, while the EER results for condition 4 are 42.45% and 40.46%, respectively. These EER results outperform all baseline systems, especially when compared with B1 and B2. Additionally, our proposed speaker anonymization system shows a significant advantage in EER results compared to other baseline systems. This indicates that our proposed system can effectively conceal the original speaker's identity and protect personal privacy.

#### 5.2. Utility Preservation

Regarding intelligibility utility, the WER results of the original speech are 1.80% and 1.85% on the dev and test sets, respectively. The speech anonymized by our proposed system achieves 2.56% and 2.66% in condition 3, that only slightly higher than the original speech, demonstrating that our proposed speaker anonymization system can effectively preserve the linguistic content of the original speech. Compared to other baseline systems, our proposed speaker anonymization system also achieves the lowest WER results, especially when compared to B2. The WER results indicate that B2 is undesirable as it significantly degrades intelligibility. Even though the WER results of condition 4 are higher than condition 3, they are still lower than those of other baseline systems.

For emotional evaluation, our proposed system achieves the highest UAR results compared to other baseline systems, with 65.98% and 64.48% for the dev and test sets, respectively. In condition 4, the UAR results of our proposed speaker anonymization system are slightly lower than in condition 3 but still higher than those of other baseline systems.

These results show that our proposed speaker anonymization system achieves better utility preservation than baseline systems, effectively preserving both linguistic content and paralinguistic information. Additionally, the utility results for condition 3 are better than those for condition 4, indicating that there is still a trade-off between privacy and utility.

# 6. Conclusion

In this paper, we propose our speaker anonymization system for VPC 2024. The anonymization system is based on the disentangled neural codec and we employ a serial disentanglement strategy to disentangle the linguistic content, speaker identity and emotion state step-by-step. This serial disentanglement strategy disentangles speech representation from a global time-invariant representation to a temporal time-variant representation. To achieve this disentanglement, we propose three specifically designed distillation methods that rely on semantic teachers or explicit labels to guide the disentanglement process. During the anonymization process, the anonymized speaker identity is the weighted sum of a set of candidate speaker identities and a randomly generated speaker identity. Experimental results on VPC 2024 Challenge datasets demonstrate that our proposed system effectively protects speaker identity while maintaining the original linguistic content and paralinguistic information.

## 7. References

- T. Zhou, Y. Zhao, and J. Wu, "Resnext and res2net structures for speaker verification," in 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021, pp. 301–307.
- [2] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, 2013, p. 395–399.
- [3] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The gdpr & speech data: Reflections of legal and technology communities, first steps towards a common understanding," in *Proceedings INTERSPEECH 2019, 20th Annual Conference of the International Speech Communication Association*, 2019, p. 3695–3699.
- [4] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien et al., "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, 2022.
- [5] N. A. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. W. D. Evans, J. Patino, J. Bonastre, P. Noé, and M. Todisco, "Introducing the voiceprivacy initiative," in *Proceedings INTERSPEECH 2020, 21st Annual Conference of the International Speech Communication Association*, 2020, pp. 1693–1697.

- [6] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, "The voiceprivacy 2022 challenge evaluation plan," *arXiv* preprint arXiv:2203.12468, 2022.
- [7] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, "The voiceprivacy 2024 challenge evaluation plan," *arXiv preprint arXiv:2404.02677*, 2024.
- [8] J. Patino, N. A. Tomashenko, M. Todisco, A. Nautsch, and N. W. D. Evans, "Speaker anonymisation using the mcadams coefficient," in *Proceedings INTERSPEECH 2021, 22nd Annual Conference of the International Speech Communication Association*, 2021, pp. 1099–1103.
- [9] P. Gupta, G. P. Prajapati, S. Singh, M. R. Kamble, and H. A. Patil, "Design of voice privacy system using linear prediction," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2020*. IEEE, 2020, pp. 543–549.
- [10] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi, "Privacy and utility of x-vector based speaker anonymization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2383–2395, 2022.
- [11] X. Miao, X. Wang, E. Cooper, J. Yamagishi, and N. A. Tomashenko, "Language-independent speaker anonymization approach using self-supervised pre-trained models," in *Odyssey* 2022: The Speaker and Language Recognition Workshop, 2022, pp. 279–286.
- [12] J. Yao, Q. Wang, L. Zhang, P. Guo, Y. Liang, and L. Xie, "Nwpuaslp system for the voiceprivacy 2022 challenge," *arXiv preprint* arXiv:2209.11969, 2022.
- [13] J. Yao, Q. Wang, Y. Lei, P. Guo, L. Xie, N. Wang, and J. Liu, "Distinguishable speaker anonymization based on formant and fundamental frequency scaling," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE, 2023, pp. 1–5.
- [14] A. Van Den Oord, O. Vinyals et al., "Neural discrete representation learning," Advances in neural information processing systems, vol. 30, 2017.
- [15] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale selfsupervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [16] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech large language models," arXiv preprint arXiv:2308.16692, 2023.
- [17] 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.
- [18] 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 textto-speech," *arXiv preprint arXiv:1904.02882*, 2019.
- [19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.
- [20] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7748–7759.
- [21] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "Hificodec: Group-residual vector quantization for high fidelity audio codec," arXiv preprint arXiv:2305.02765, 2023.