CTEFM-VC: Zero-Shot Voice Conversion Based on Content-Aware Timbre Ensemble Modeling and Flow Matching

Anonymous submission to Interspeech 2025

Abstract

Despite impressive advances in recent zero-shot voice conversion (VC), achieving speaker similarity and naturalness comparable to ground truth recordings remains a significant challenge. In this paper, we propose ***CTEFM-VC***, a zero-shot VC framework that integrates **C**ontent-aware **T**imbre **E**nsemble modeling and **F**low **M**atching. Specifically, CTEFM-VC decouples utterances into content and timbre representations and leverages a conditional flow matching model to reconstruct the Mel-spectrogram of the source speech. To enhance its timbre modeling capability and the naturalness of generated speech, we introduce a context-aware timbre ensemble approach that adaptively integrates diverse speaker verification embeddings and enables the joint utilization of source content and target timbre through a cross-attention module. Experimental results show that CTEFM-VC outperforms state-of-the-art VC systems, achieving better speaker similarity, inference speed, and superior speech naturalness.

Index Terms: zero-shot voice conversion, content-aware timbre ensemble modeling, cross-attention, flow matching

1. Introduction

As a pivotal task within the field of speech signal processing, zero-shot voice conversion (VC) aims to transfer the timbre of a source utterance to an arbitrary unseen speaker while maintaining the original phonetic content, with applications spanning various practical domains such as voice anonymization [1] and audiobook production [2].

In general, the core difficulties of zero-shot VC lie in effectively modeling, decoupling, and utilizing various speech attributes, including content, timbre, etc. Previous approaches [3, 4, 5] often use pre-trained automatic speech recognition (ASR) methods [6, 7] and speaker verification (SV) models [8, 9] to extract linguistic content and timbre information from the source and target speech, respectively. Nevertheless, due to many factors such as the inherent complexity of speech signals [10, 11] and limitations in timbre and content modeling methods [12, 13], they present significant opportunities for performance enhancement. With the progressions of self-supervised learning (SSL)-based speech models [14, 15], many works [16, 12] have sought to use them to extract semantic features from speech. However, the extracted features inevitably contain certain source timbre characteristics [17], ultimately affecting their VC quality. To this end, [18, 13, 17] incorporated the model quantization technique to minimize non-content elements. Nonetheless, this operation incurs additional training overhead, and variations in training objectives may lead to token format inconsistencies across different models, thus limiting the broader applicability. Moreover, existing zero-shot VC methods [3, 5, 12, 17] normally employ a single pre-trained SV model to capture target timbre embeddings. Although speaker embedding techniques have advanced significantly [9, 19, 20], relying exclusively on a single model is insufficient to deliver optimal VC performance, resulting in sub-par speaker similarity compared to authentic recordings.

Inspired by the powerful zero-shot capabilities of recent large-scale language models (LLMs) [21], several studies [22, 23] have attempted to discretize waveforms using neural codecs [24, 25] and then leverage LLMs to generate target waveforms in an autoregressive manner. [22] proposed a two-stage language model that first generates coarse acoustic tokens to capture the source content and target timbre elements, and then refines the acoustic details for VC. [23] implemented a single-stage VC framework based on the context-sensitive language model and acoustic predictor, which facilitated zero-shot voice conversion in a streamable way. Despite great results, such methods commonly encounter stability problems due to their auto-regressive fashion and may experience error accumulation, leading to a gradual decline in VC performance.

To address the aforementioned problems, we propose ***CTEFM-VC***, a novel zero-shot VC framework based on content-aware timbre ensemble modeling and flow matching. We first employ HybridFormer, a pretrained ASR model [7], to capture the precise source linguistic content elements. To improve timbre modeling capabilities and real-time performance, we introduce an ensemble strategy that leverages multiple pre-trained SV models to extract target timbre embeddings, concatenating them as a conditional input for a conditional flow matching (CFM) [26, 17] module. Additionally, to further enhance the speaker similarity and naturalness of the entire system, we propose an effective and easily scalable content-aware timbre ensemble modeling (CTE) approach that integrates all captured SV embeddings and uses a cross-attention module to achieve adaptive fusion between timbre embeddings and linguistic content features, thereby generating higher-quality representations to serve as another conditional input for CFM. Finally, we incorporate the CFM and pre-trained vocoder [27] to reconstruct the Mel-spectrogram of the source utterance and generate the converted speech, respectively. Experimental results demonstrate that our CTEFM-VC surpasses several state-of-the-art (SOTA) zero-shot VC approaches in terms of speaker similarity and real-time performance, while obtaining impressive speech naturalness.

2. METHODOLOGY

2.1. System Architecture

As depicted in Fig. 1, the proposed CTEFM-VC is an end-to-end zero-shot VC framework. Assume the input speech signal

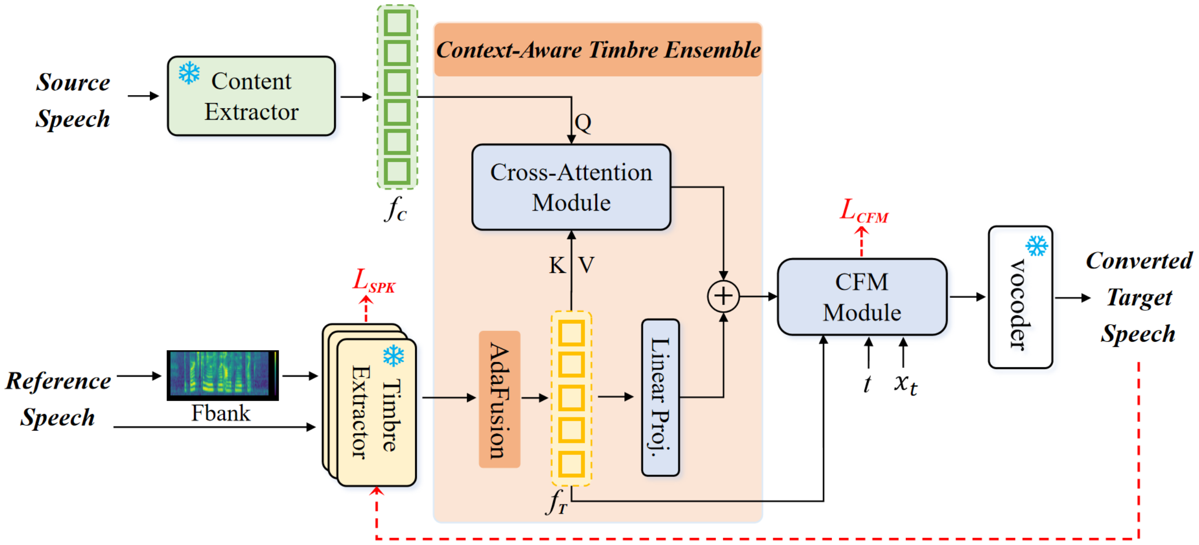

Figure 1: *Overall training architecture of the proposed CTEFM-VC framework.*

is represented as $X = [x_1, x_2, ..., x_T] \in R^T$, our CTEFM-VC initially adopts a pre-trained ASR method, HybridFormer, to extract the linguistic content $f_C \in R^{T_1 \times D}$. Detailed, the used HybridFormer consists of 12 blocks, each with a convolution kernel size of 31 and 4 attention heads. The hidden dimensions of the attention layer and feedforward network (FFN) are set to 256 and 1024, respectively. For timbre modeling, unlike previous studies, we use multiple pretrained SV models [9, 19, 20] to extract their corresponding timbre embeddings $f_{T_i} \in R^{d_i}$ of the reference waveform. To further enhance its speaker similarity and naturalness, we propose CTE, a context-aware timbre ensemble modeling approach that uses a straightforward yet effective AdaFusion method to fuse all SV embeddings f_{T_i} as global representations f_T and employ a cross-attention module to facilitate joint utilization of f_C and f_T. Last, the outputs F of CTE and the ensembled target timbre characteristics f_T are fed into a conditional flow matching model to reconstruct the Mel-spectrogram of source waveform, followed by a pretrained Bigvgan vocoder to generate the desired target waveform.

2.2. CTE: Context-aware Timbre Ensemble Modeling

Essentially, the core challenge of zero-shot VC stems from timbre modeling, as it necessitates the model's ability to generalize effectively to arbitrary unseen speakers without requiring supplementary training or fine-tuning.

Consequently, to enhance the overall timbre modeling capacity of the proposed method, we first present a model ensemble approach that employs multiple pretrained speaker verification (SV) models to extract timbre features. We hypothesize that integrating diverse timbre embeddings enables our approach to capture a richer array of timbre characteristics, thereby enhancing its adaptability to variations in speaker identity. Besides, to better exploit these features, we propose a simple yet effective AdaFusion method that applies learnable hyperparameters to weight the SV embeddings before concatenating them into a unified global timbre representation f_T. Subsequently, we construct a cross-attention module comprising six multi-head cross-attention blocks to facilitate the adaptive utilization of the captured source linguistic content and target timbre, as depicted in Fig. 2. Each block consists of a stack of linear projection layer, layer normalization layer, and FFN, where f_C serves as the query in the attention mechanism and f_T functions as the key and value. Finally, the captured target timbre f_T is projected to the same dimension as the output of the cross-attention module through a linear layer and an element-wise addition operation, as shown in Fig. 1.

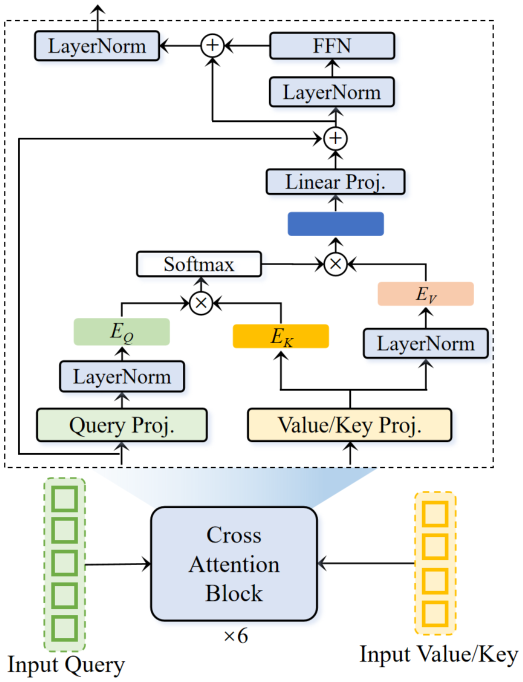

Figure 2: *Schematic of the proposed cross-attention module.*

2.3. Conditional Flow Matching

To achieve an optimal balance between the generation quality and real-time performance, we incorporate an optimal transport (OT)-based CFM module to reconstruct the target mel-spectrogram $x_1{=}p_1(x)$ from standard Gaussian noise $x_0 = p_0(x){=}\mathcal{N}(x;0,I)$. Concretely, an OT flow $\psi_t : [0,1] \times R^d \rightarrow R^d$ is employed to train our proposed OT-CFM that is composed of multiple UNet [28] blocks with timestep fusion. By leveraging the ordinary differential equation to estimate a learnable time-dependent vector field $v_t{:}[0,1] \times R^d \rightarrow R^d$, it can approximate the optimal transport probability path from $p_0(x)$ to the target distribution $p_1(x)$:

$$\frac{d}{d_t}\psi_t(x) = v_t(\psi_t(x), t) \tag{1}$$

where $\psi_0(x) = x$, and $t \in [0,1]$. Additionally, inspired by [29, 17] that recommend straighter trajectories, we simplify the formula of OT flow as follow:

$$\psi_{t,z}(x) = \mu_t(z) + \sigma_t(z)x \tag{2}$$

where $\mu_t(z) = tz$, $\sigma_t(z) = (1{-}(1{-}\sigma_{min})t)$, z represents the random conditioned input, σ_{min} signifies the minimum standard deviation of white noise introduced to perturb individual samples, empirically set to 0.0001. As a result, the training loss of our OT-CFM module is formulated as:

$$\mathcal{L}_{CFM}{=}\mathbb{E}_{t,p(x_0),q(x_1)}||(x_1{-}(1{-}\sigma)x_0){-}v_t(\psi_{t,x_1}(x_0)|h)||^2 \tag{3}$$

Here, $t \sim U[0,1]$, $x_0 \sim p(x_0)$, $x_1 \sim q(x_1)$, with $q(x_1)$ representing the true but potentially non-Gaussian distribution of the data, and h denotes the captured conditional inputs of the CFM model.

2.4. Training Objectives

In addition to the L_{CFM}, we incorporate a structural similarity-based loss function [25] to minimize the disparity between the timbre embeddings of the reference and converted speech, thereby further enhancing the speaker similarity performance of CTEFM-VC:

$$\begin{gathered} L_{spk} = \sum_{i=1}^{N} L_{spk_i} = \text{SSIM}(T_{r_i}, T_{c_i}) \\ \text{SSIM}(T_{r_i}, T_{c_i}){=}\frac{(2\mu_{T_{r_i}}\mu_{T_{c_i}}{+}c_1)(2\sigma_{T_{r_i}T_{c_i}}{+}c_2)}{(\mu^2_{T_{r_i}}{+}\mu^2_{T_{c_i}}{+}c_1)(\sigma^2_{T_{r_i}}{+}\sigma^2_{T_{c_i}}{+}c_2)} \end{gathered} \tag{4}$$

where, T_{r_i} and T_{c_i} are the corresponding features of the employed SV models, μ_{T_r} and μ_{T_c} denote the means of T_r and T_c, while $\sigma^2_{T_r}$ and $\sigma^2_{T_c}$ represent their respective variances. In addition, $\sigma_{T_rT_c}$ indicates the covariance between T_r and T_c. The constants c_1 and c_2 are employed to ensure numerical stability during division, with values set to 0.01 and 0.03, respectively.

Therefore, the overall training objective of the proposed CTEFM-VC can be expressed as:

$$L_{Total} = \mathcal{L}_{CFM} + \lambda L_{spk} \tag{5}$$

where λ is a tuning hyperparameter empirically set to 0.05.

3. EXPERIMENTS

3.1. Experimental Setups

3.1.1. Datasets

We conduct all experiments on the LibriTTS [30] dataset, which comprises 585 hours of English recordings from 2,456 speakers. All data is downsampled to 16kHz. We use all training subsets for model training, while the dev-clean subset is employed for validation. To assess their zero-shot VC performance, we select the VCTK [31] and ESD [32] corpora. For each corpus, we randomly select 100 samples from 10 unseen speakers, ensuring that there is no overlap with the training data.

3.1.2. Implementation Details

We adopt the AdamW optimizer to train the proposed CTEFM-VC approach over 600K iterations using four NVIDIA A10 GPUs, with an initial learning rate of 1e-4 and a batch size of 64. During training, we randomly segment 4s of the source speech as the reference waveform. During the inference stage, the CFM model operates with 20 Euler steps to generate the target outputs.

3.1.3. Evaluation Metric

To thoroughly assess our CTEFM-VC method, we perform objective and subjective evaluations.

In the objective evaluation, we compute the speaker embedding cosine similarity (SECS), character error rate (CER), and UTMOS between the converted speech and the reference speech using a pre-trained WavLM-TDCNN model[1], a CTC-based ASR system[2], and a mean opinion score (MOS) prediction model[3], respectively. Regarding subjective evaluation, we invite 15 professional raters to assign MOS scores for both naturalness (NMOS) and similarity (SMOS) on a scale ranging from 1 to 5. This scoring reflects the naturalness of the generated speech and the speaker similarity between the converted and source waveforms. To be specific, a score of '5' indicates excellent quality, '4' denotes good quality, '3' represents fair quality, '2' indicates poor quality, and '1' signifies bad quality. Real-time factor (RTF) is adopted to evaluate the real-time performance.

3.2. Main Results

To examine the performance of the proposed CTEFM-VC, we compare it with six SOTA zero-shot VC baselines, with the results reported in Table 1.

Our experimental results show that CTEFM-VC achieves superior performance in both subjective and objective evaluations. In terms of objective assessment, CTEFM-VC outperforms all other methods in speaker similarity and real-time performance, while attaining the suboptimal results in terms of speech naturalness and intelligibility. In detail, CTEFM-VC attains an SECS of 0.78 and an RTF of 0.148, exceeding other SOTA methods by at least 9.9% and 3.9%. These findings showcase the powerful capabilities of the proposed method in accurately converting the target timbre and high inference speed. In addition, our method achieves the secondary lowest WER score, indicating that the converted utterance by CTEFM-VC exhibits superior intelligibility.

As for subjective evaluation, the proposed CTEFM-VC system consistently exceeds comparative zero-shot VC baselines in speaker similarity as well. In particular, CTEFM-VC achieves the highest SMOS of 4.16, which correlates with the SECS metric and further validates its robust capabilities in converting the

[1] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification

[2] https://huggingface.co/facebook/hubert-large-ls960-ft

[3] https://github.com/tarepan/SpeechMOS

Table 1: *Comparative results of subjective and objective evaluations between CTEFM-VC and the baseline systems in zero-shot VC. Subjective metrics are computed with 95% confidence intervals and "GT" denotes the ground truth recordings. The best results are highlighted in bold, while the sub-optimal results are underlined.*

	NMOS (↑)	SMOS (↑)	WER (↓)	UTMOS (↑)	SECS (↑)	RTF (↓)
GT	4.18±0.05	-	2.01	4.19	-	
DiffVC [33]	3.75±0.05	3.66±0.07	3.08	3.68	0.61	0.294
NS2VC [34]	3.65±0.07	3.51±0.06	2.94	3.64	0.53	0.347
VALLE-VC [35]	3.80±0.06	3.79±0.04	2.77	3.72	0.65	3.678
SEFVC [13]	3.68±0.05	3.76±0.06	3.75	3.51	0.63	0.187
StableVC [26]	3.83±0.04	3.88±0.06	2.77	3.92	0.66	0.267
Takin-VC [17]	**3.98**±0.04	4.11±0.05	**2.35**	**4.08**	0.71	0.154
CTEFM-VC	3.92±0.05	**4.16**±0.04	2.41	3.99	**0.78**	**0.148**

Table 2: *Experimental results on ablation studies. 'w/o' represents removing the corresponding module. 'SV1', 'SV2', and 'SV3' represent the pretrained CAM++, ERes2Net, and ReDimNet models, respectively.*

	NMOS (↑)	SMOS (↑)	WER (↓)	UTMOS (↑)	SECS (↑)
CTEFM-VC	**3.92**±0.05	**4.16**±0.04	**2.41**	3.99	**0.78**
w/o *SV1*	3.66±0.05	3.79±0.05	2.85	3.82	0.65
w/o *SV2*	3.71±0.04	3.85±0.05	2.78	3.86	0.68
w/o *SV3*	3.73±0.05	3.92±0.04	2.76	3.85	0.69
w/o *AdaFusion*	3.69±0.05	3.94±0.05	2.68	3.72	0.71
w/o L_{spk}	**3.92**±0.04	3.75±0.05	2.44	**4.01**	0.61

target timbre. Regarding NMOS, our proposed approach gets the secondary best result, slightly worse than Takin-VC. Overall, these findings provide compelling corroboration for the subjective results, underscoring the effectiveness and robustness of the proposed approach.

3.3. Ablation Study

To evaluate the contributions and validity of each component of the proposed CTEFM-VC method, ablation studies are conducted. All results are summarized in Table 2.

Initially, we assess the effectiveness of the timbre ensemble modeling strategy. As indicated in Table 2, the omission of any pretrained SV model leads to a marked decrease in performance across all metrics, particularly evident in the SECS score. This phenomenon supports our hypothesis that the integration of diverse timbre embeddings can facilitate the capture of a broader range of timbre characteristics, thereby enhancing the timbre modeling capabilities of the proposed method. Furthermore, the experimental results indicate that within the proposed CTEFM-VC framework, CAM++ demonstrates the most effective timbre modeling capability, followed closely by ERes2Net, whereas ReDimNet exhibits the least effectiveness.

Next, we examine the influence of removing the AdaFusion method. From Table 2, we can easily observe that in the absence of AdaFusion, both subjective and objective metrics display varying degrees of decline. Specifically, subjective NMOS and SMOS scores decrease by 6.2% and 5.6%, while objective scores drop by 7.3% to 11.2%. This reduction proves the effectiveness of the proposed AdaFusion method in adaptively fusing different timbre embeddings. By learning the importance of individual SV model, AdaFusion facilitates the joint utilization of linguistic content and timbre features, thus enhancing overall performance of CTEFM-VC.

Last, we study the impact of the SSIM-based speaker similarity loss L_{spk}. From the above table, we notice that without the L_{spk}, the SECS and SMOS scores drop significantly, while the UTMOS and NMOS exhibit slightly improvements. This observation suggests that the proposed L_{spk} is crucial for capturing the characteristics of speakers. However, the slight increase in the UTMOS and NMOS scores without L_{spk} indicates that the model may slightly enhance naturalness and intelligibility at the cost of speaker similarity. This trade-off highlights the importance of carefully balancing the loss components to achieve optimal performance in multiple evaluation metrics.

4. CONCLUSIONS

In this work, we propose CTEFM-VC, an innovative and scalable zero-shot VC workflow based on content-aware timbre ensemble approach and conditional flow matching. To elaborate, CTEFM-VC utilizes a pretrained ASR model and multiple SV models to extract linguistic content and timbre features. Subsequently, a content-aware timbre ensemble modeling method is proposed to integrate diverse timbre embeddings and facilitate adaptive utilization of the source content and target timbre representations. To enable stable training and fast inference, we incorporate a CFM model to reconstruct the source mel-spectrogram, followed by a pretrained vocoder to generate the converted speech. Extensive experiments conducted on the LibriTTS corpus indicate that compared to recent SOTA zero-shot VC methods, our proposed CTEFM-VC realizes better speaker similarity, naturalness, and superior speech anturalness.

5. References

[1] J. Yao, Q. Wang, P. Guo, Z. Ning, Y. Yang, Y. Pan, and L. Xie, "Musa: Multi-lingual speaker anonymization via serial disentanglement," *arXiv preprint arXiv:2407.11629*, 2024.

[2] S. Chen, Y. Feng, L. He, T. He, W. He, Y. Hu, B. Lin, Y. Lin, P. Tan, C. Tian *et al.*, "Takin: A cohort of superior quality zero-shot speech generation models," *arXiv preprint arXiv:2409.12139*, 2024.

[3] Z. Tan, J. Wei, J. Xu, Y. He, and W. Lu, "Zero-shot voice conversion with adjusted speaker embeddings and simple acoustic features," in *ICASSP 2021-2021 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 5964–5968.

[4] X. Zhao, F. Liu, C. Song, Z. Wu, S. Kang, D. Tuo, and H. Meng, “Disentangling content and fine-grained prosody information via hybrid asr bottleneck features for voice conversion,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7022–7026.

[5] S. Kovela, R. Valle, A. Dantrey, and B. Catanzaro, “Any-to-any voice conversion with f 0 and timbre disentanglement and novel timbre conditioning,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.

[7] Y. Yang, Y. Pan, J. Yin, J. Han, L. Ma, and H. Lu, “Hybridformer: Improving squeezeformer with hybrid attention and nsr mechanism,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[8] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *arXiv preprint arXiv:2005.07143*, 2020.

[9] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, “Cam++: A fast and efficient network for speaker verification using context-aware masking,” *arXiv preprint arXiv:2303.00332*, 2023.

[10] Y. Pan, Y. Hu, Y. Yang, W. Fei, J. Yao, H. Lu, L. Ma, and J. Zhao, “Gemo-clap: Gender-attribute-enhanced contrastive language-audio pretraining for accurate speech emotion recognition,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 021–10 025.

[11] Y. Pan, Y. Yang, H. Lu, L. Ma, and J. Zhao, “Gmp-atl: Gender-augmented multi-scale pseudo-label enhanced adaptive transfer learning for speech emotion recognition via hubert,” *arXiv preprint arXiv:2405.02151*, 2024.

[12] S. Hussain, P. Neekhara, J. Huang, J. Li, and B. Ginsburg, “Ace-vc: Adaptive and controllable voice conversion using explicitly disentangled self-supervised speech representations,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] J. Li, Y. Guo, X. Chen, and K. Yu, “Sef-vc: Speaker embedding free zero-shot voice conversion with cross attention,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 296–12 300.

[14] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. 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.

[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 self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[16] T. Dang, D. Tran, P. Chin, and K. Koishida, “Training robust zero-shot voice conversion models with self-supervised features,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6557–6561.

[17] Y. Yang, Y. Pan, J. Yao, X. Zhang, J. Ye, H. Zhou, L. Xie, L. Ma, and J. Zhao, “Takin-vc: Zero-shot voice conversion via jointly hybrid content and memory-augmented context-aware timbre modeling,” *arXiv preprint arXiv:2410.01350*, 2024.

[18] J. Yao, Y. Yang, Y. Lei, Z. Ning, Y. Hu, Y. Pan, J. Yin, H. Zhou, H. Lu, and L. Xie, “Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 571–10 575.

[19] Y. Chen, S. Zheng, H. Wang, L. Cheng, Q. Chen, and J. Qi, “An enhanced res2net with local and global feature fusion for speaker verification,” *arXiv preprint arXiv:2305.12838*, 2023.

[20] I. Yakovlev, R. Makarov, A. Balykin, P. Malov, A. Okhotnikov, and N. Torgashov, “Reshape dimensions network for speaker recognition,” *arXiv preprint arXiv:2407.18223*, 2024.

[21] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, “Gpt understands, too,” *AI Open*, 2023.

[22] Z. Wang, Y. Chen, L. Xie, Q. Tian, and Y. Wang, “Lm-vc: Zero-shot voice conversion via speech generation based on language models,” *IEEE Signal Processing Letters*, 2023.

[23] Z. Wang, Y. Chen, X. Wang, Z. Chen, L. Xie, Y. Wang, and Y. Wang, “Streamvoice: Streamable context-aware language modeling for real-time zero-shot voice conversion,” *arXiv preprint arXiv:2401.11053*, 2024.

[24] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *arXiv preprint arXiv:2210.13438*, 2022.

[25] Y. Pan, L. Ma, and J. Zhao, “Promptcodec: High-fidelity neural speech codec using disentangled representation learning based adaptive feature-aware prompt encoders,” *arXiv preprint arXiv:2404.02702*, 2024.

[26] J. Yao, Y. Yan, Y. Pan, Z. Ning, J. Ye, H. Zhou, and L. Xie, “Stablevc: Style controllable zero-shot voice conversion with conditional flow matching,” *arXiv preprint arXiv:2412.04724*, 2024.

[27] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, “Bigvgan: A universal neural vocoder with large-scale training,” *arXiv preprint arXiv:2206.04658*, 2022.

[28] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[29] A. Tong, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, K. Fatras, G. Wolf, and Y. Bengio, “Conditional flow matching: Simulation-free dynamic optimal transport,” *arXiv preprint arXiv:2302.00482*, vol. 2, no. 3, 2023.

[30] 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.

[31] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92),” *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pp. 271–350, 2019.

[32] K. Zhou, B. Sisman, R. Liu, and H. Li, “Emotional voice conversion: Theory, databases and esd,” *Speech Communication*, vol. 137, pp. 1–18, 2022.

[33] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, “Diffusion-based voice conversion with fast maximum likelihood sampling scheme,” *arXiv preprint arXiv:2109.13821*, 2021.

[34] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, “Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers,” *arXiv preprint arXiv:2304.09116*, 2023.

[35] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, “Neural codec language models are zero-shot text to speech synthesizers,” *arXiv preprint arXiv:2301.02111*, 2023.