Takin-VC: Expressive Zero-Shot Voice Conversion via Adaptive Hybrid Content Encoding and Enhanced Timbre Modeling

Anonymous ACL submission

Abstract

Expressive zero-shot voice conversion (VC) is a critical and challenging task that aims to transform the source timbre into an arbitrary unseen speaker while preserving the original content and expressive qualities. Despite recent progress in zero-shot VC, there remains considerable potential for improvements in speaker similarity and speech naturalness. Moreover, existing zero-shot VC systems struggle to fully reproduce paralinguistic information in highly expressive speech, such as breathing, crying, and emotional nuances, limiting their practical applicability. To address these issues, we propose Takin-VC, a novel expressive zero-shot VC framework via adaptive hybrid content encoding and memory-augmented context-aware timbre modeling. Specifically, we introduce an innovative hybrid content encoder that incorporates an adaptive fusion module, capable of effectively integrating quantized features of the pre-trained WavLM and HybridFormer in an implicit manner, so as to extract precise linguistic features while enriching paralinguistic elements. For timbre modeling, we propose advanced memory-augmented and context-aware modules to generate high-quality target timbre features and fused representations that seamlessly align source content with target timbre. To enhance real-time performance, we advocate a conditional flow matching model to reconstruct the Mel-spectrogram of the source speech. Experimental results show that our Takin-VC consistently surpasses state-of-the-art VC systems, achieving notable improvements in terms of speech naturalness, speech expressiveness, and speaker similarity, while offering enhanced inference speed.

1 Introduction

Zero-shot voice conversion (VC) aims to modify the timbre of a source speech to match that of a previously unseen speaker, while maintaining the original phonetic content, has found broad applications in various practical domains (Gan et al., 2022; Tomashenko et al., 2022; Liu et al., 2021).

The advancement of deep learning techniques has significantly propelled the development of zero-shot VC, with numerous methods (Li et al., 2023; Hussain et al., 2023; Choi et al., 2023; Anastassiou et al., 2024; Luo and Dixon, 2024) exhibiting impressive results in converting natural and realistic speech. The key idea behind is factorizing speech into distinct elements, such as content and timbre elements, and leveraging the source speech content alongside the target timbre to synthesize the desired output. In this paradigm, the quality of content and timbre features, as well as the quality of their disentanglement, critically influences performance. Consequently, various studies have focused on developing advanced modules (Wu et al., 2020; Wu and Lee, 2020; Tang et al., 2022; Wang et al., 2021; Yang et al., 2022a; Huang et al., 2023) and information disentanglement approaches (Zhao et al., 2022; Tang et al., 2022; Dang et al., 2022; Yao et al., 2024c) to enhance zero-shot VC. However, achieving high-quality decoupling of utterances into distinct components remains challenging (Pan et al., 2023, 2024a,c; Yao et al., 2024a), with existing systems still exhibiting subpar performance for unseen speakers. Two main issues are prevalent. First, current methods cannot fully mitigate the impact of source timbre during source content extraction, a problem referred to as "timbre leakage". Second, these approaches often use pre-trained speaker-verification (SV) models to capture target timbre features as globally time-invariant representations. Nonetheless, such SV embeddings cannot ensure robust timbre modeling and vary with linguistic content (Jiang et al., 2024; Pan et al., 2024d) which may diminish their effectiveness.

Recently, the progressions in large-scale speech language models (Wang et al., 2023b; Borsos et al., 2023) have tried to tackle this issue by leveraging robust in-context learning capabilities for converting target speech from concise utterances as prompts. Nevertheless, these methods suffer from stability issues and error accumulation due to their auto-regressive nature, which can gradually degrade conversion quality. Moreover, current state-of-the-art (SOTA) zero-shot VC systems still struggle to simultaneously transfer the paralinguistic characteristics in highly expressive speech, such as crying, breathing, and emotional nuances, thus limiting their effectiveness and practical applicability.

In this paper, we propose *Takin-VC*, a novel expressive zero-shot VC framework that delivers advanced modeling of content, timbre, and speech quality in a zero-shot fashion. To be specific, we introduce an adaptive fusion-based hybrid content encoder that seamlessly combines the strengths of phonetic posterior-grams (PPGs) and self-supervised learning (SSL)-based representations derived from pre-trained HybridFormer (Yang et al., 2023b) and WavLM (Chen et al., 2022). This integration enables the precise extraction of linguistic content while simultaneously enriching paralinguistic elements. For timbre modeling, we first advocate a memory-augmented module capable of generating high-quality conditional target timbre inputs for our conditional flow matching (CFM) model. To further enhance speaker similarity, a context-aware timbre modeling module based on an efficient cross-attention (CA) mechanism is presented. This module effectively aligns and fuse the extracted source content and target timbre features, rather than solely using the source linguistic content as the conditional input for CFM. Conditioned on these features, the predicted outputs of the CFM model are ultimately fed into a pre-trained vocoder (Lee et al., 2022) to synthesize the target speech.

Experiments conducted on both large-scale 500k-hour multilingual (Mandarin and English) and small-scale LibriTTS (Zen et al., 2019) datasets demonstrate that Takin-VC consistently outperforms several SOTA zero-shot VC methods in speech naturalness, expressiveness, speaker similarity, and real-time performance. Notably, Takin-VC achieves significant improvements in both subjective and objective metrics compared to all baseline systems, further validating its effectiveness and robustness. For more detailed speech samples, please visit our **demo page** [1]. In summary, the primary contributions of this work are as follows:

[1]https://anonymous.4open.science/w/takin-vc-0CD8/

- We present Takin-VC, a novel expressive zero-shot VC framework. To the best of our knowledge, this is the first approach capable of simultaneously transforming the source timbre to arbitrary unseen speakers while effectively maintaining the paralinguistic characteristics of highly expressive speech.

- We introduce an adaptive hybrid content encoder that employs an adaptive feature fusion module to implicitly integrate PPGs and quantized SSL features in a learnable manner, thereby capturing precise linguistic elements with enriched paralinguistic characteristics.

- We propose memory-augmented and content-aware modules to enhance timbre modeling. The former aims to extract high-quality target timbre conditions, while the latter focuses on generating fused features that align and leverage target timbre embeddings with source content for the conditional flow matching model.

2 Background

2.1 Zero-shot Voice Conversion

Recent progressions in deep learning techniques, such as SSL-based speech models (Hsu et al., 2021; Chen et al., 2022; Baevski et al., 2020) and diffusion models (Ho et al., 2020; Lu et al., 2022), have greatly advance zero-shot VC. SEF-VC (Li et al., 2024) utilizes a CA mechanism to extract timbre features and reconstruct waveforms from HuBERT (Hsu et al., 2021) tokens, while (Choi et al., 2023) proposes a diffusion-based hierarchical VC method using XLS-R (Babu et al., 2021) for content extraction and dual diffusion models for generating pitch and Mel-spectrograms. Despite these innovations, SSL-based zero-shot VC methods (Dang et al., 2022; Hussain et al., 2023; Li et al., 2023) are likely to encounter the timbre leakage challenge, as SSL features do not explicitly disentangle timbre features. Likewise, diffusion-based approaches (Popov et al., 2021; Choi et al., 2024) suffer from suboptimal real-time performance. Another emerging paradigm (Zhang et al., 2023; Wang et al., 2023b; Baade et al., 2024) involves decoupling speech into semantic and acoustic tokens using neural codecs (Défossez et al., 2022; Yang et al., 2023a; Pan et al., 2024b) and SSL-based models, subsequently using language models to generate converted speech. While these approaches mark impressive results, current SOTA VC methods still have considerable room for improvement in achieving better speaker similarity and naturalness. Besides, they continue to face difficulties in faithfully and simultaneously reproducing the paralinguistic characteristics of highly expressive speech.

2.2 Flow Matching-based Generative Models

Flow matching-based generative models (Lipman et al., 2022; Tong et al., 2023c,a) have recently emerged as a powerful solution for generative tasks. By estimating vector fields to approximate the transport path from noise to the target distribution, these models employ neural ordinary differential equations (ODEs) to learn optimal transport trajectories. Compared to diffusion-based methods (Bartosh et al., 2023; Zhou et al., 2023), flow matching offers improved training stability and real-time performance by enabling direct noise-to-sample mapping while significantly reducing sampling steps. In the speech processing domain, flow matching-based systems are emerging as a promising paradigm. SpeechFlow (Liu et al., 2023) uses a pre-trained flow matching model with masked conditions on large-scale untranscribed speech data, facilitating speech enhancement and separation tasks. P-Flow (Kim et al., 2024) adopts speech prompts for speaker adaptation, integrating a speech-prompted text encoder and a flow matching decoder to enable high-quality and real-time speech synthesis. Despite these advancements, the application of flow matching in zero-shot VC remains nascent, underscoring the need for developing a stable and efficient flow matching-based zero-shot VC framework.

3 TakinVC

3.1 Overivew

As shown in Fig. 1, our Takin-VC system primarily comprises three key components: an adaptive hybrid content encoder, a memory-augmented context-aware timbre modeling approach, and a conditional flow matching-based decoder.

In detail, the objective of the adaptive hybrid content encoder is to precisely capture linguistic characteristics enriched with paralinguistic elements, denoted as $X_{s_{cont}}$. To achieve this, an adaptive feature fusion module on top of the hybrid content encoder is presented to effectively leverage the complementary strengths of PPG and quantized SSL representations in a learnable fashion. Regarding timbre modeling, we first propose a memory-augmented module that incorporates a stack of convolution, activation, and self-attention layers to extract high-quality target timbre conditions $X_{t_{tcond}}$ for the CFM model. To further improve timbre modeling capabilities, a cross-attention-based context-aware module is presented to generate fused representations $X_{s_{ct_t}}$ that effectively integrate $X_{s_{cont}}$ with target timbre. Finally, to enable stable training and accelerate the reference speed, we design a CFM model that consists of multiple UNet (Ronneberger et al., 2015) blocks to reconstruct the source Mel-spectrograms conditioned on $X_{s_{ct_t}}$ and $X_{t_{tcond}}$, followed by a pretrained Bigvgan vocoder to synthesize the desired target speech.

3.2 Adaptive Hybrid Content Encoder

Current mainstream zero-shot VC systems typically use pretrained automatic speech recognition (ASR) (Gulati et al., 2020; Yang et al., 2022b; Kim et al., 2022) or SSL-based speech models to capture linguistic content from the original waveform. However, they both have inherent limitations: ASR-derived PPGs lack sufficient paralinguistic elements, whereas SSL-based models do not explicitly disentangle timbre information. To address these flaws, we propose an adaptive fusion-based hybrid content encoder within the Takin-VC framework, integrating the merits of both approaches.

Formally, given an input source speech X, our adaptive hybrid content encoder separately encodes its corresponding PPG and SSL features, denoted as X_p and X_s, using pre-trained HybridFormer and WavLM, respectively. To alleviate potential timbre leakage, a residual vector quantization (RVQ) based quantizer of EnCodec (Défossez et al., 2022) is applied to discretize X_s, resulting in $\tilde{X}_s$. Additionally, we introduce a gradient-driven adaptive feature fusion module to further reduce timbre leakage and effectively integrate the complementary benefits of PPG and SSL features. Unlike conventional element-wise addition for feature fusion, the proposed strategy first processes the quantized WavLM features through a multi-layer projection module comprising a one-dimensional convolutional (Conv1d) layer followed by a LeakyReLU activation function, with the negative slope empirically set to 0.2. Temporal interpolation is then applied to ensure dimensional alignment with the PPG features, and the resulting WavLM representations are employed as coefficients for element-wise

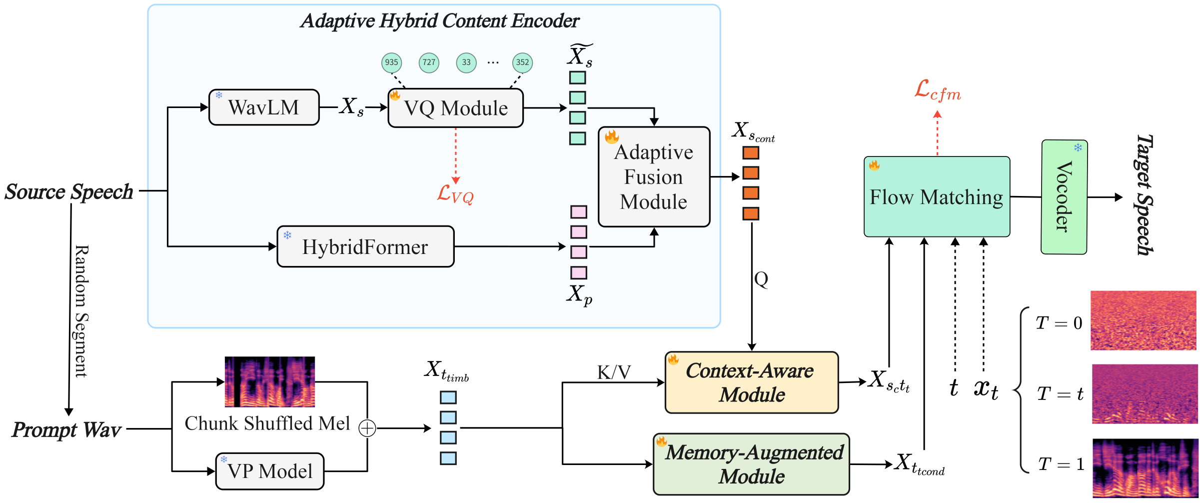

Figure 1: The overall framework of Takin-VC.

multiplication with the PPGs:

$$X_{s_{cont}} = \text{LeakyReLU}(\text{Conv1d}(\tilde{X}_s)) \cdot X_p \quad (1)$$

where Conv1d denotes the 1D convolutional layer. By this means, as gradients propagate back to Formula 1 during training, the limited representation of paralinguistic nuances within the PPG features results in larger gradient magnitudes for these elements. Since the PPGs are fixed before training, the gradients primarily affect the adaptive fusion module associated with the quantized WavLM features. As a consequence, this gradient-driven adjustment dynamically optimizes the weights of the quantized WavLM features in an implicit way, thereby amplifying the representation of paralinguistic elements in the combined feature space, improving overall content modeling capabilities, and significantly reducing the risk of voiceprint leakage.

3.3 Enhanced Timbre Modeling

3.3.1 Memory-augmented Timbre Modeling

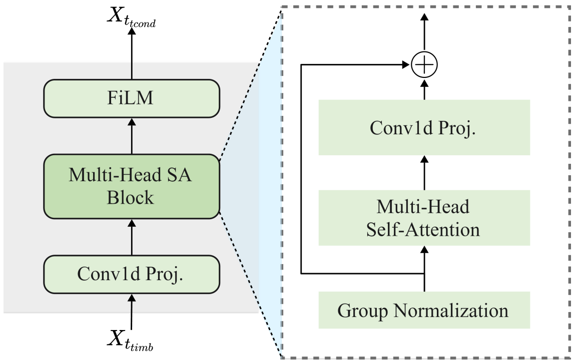

Figure 2: Schematic of the memory-augmented module.

To capture high-quality target timbre conditions for the CFM model, we propose an efficient memory-augmented module that adaptively integrates the shuffled Mel-spectrogram and VP features of the reference speech, as outlined in Fig. 2.

Detailed, we extract the Mel-spectrograms from randomly segmented reference waveforms originating from the same speaker as the source speech. The individual frames of these Mel-spectrograms are then shuffled to preserve essential timbre characteristics while minimizing the influence of the source speech content. Subsequently, a lightweight pre-trained SV model[2] is utilized to extract timbre embeddings from the reference speech. These embeddings are then concatenated with the shuffled Mel-spectrograms, resulting in the target timbre representations, referred to as $X_{t_{timb}}$. To refine these concatenated features, our proposed memory-augmented module that begins by employing a Conv1d layer to project the captured features and then incorporates four SA blocks, each comprising a group normalization layer, multi-head SA mechanism, a Conv1d layer, and a shortcut connection operation. The resulting features are then subjected to a temporal averaging operation, followed by the application of a FiLM layer (Perez et al., 2018) to perform affine feature-wise transformation, producing the conditional target timbre inputs $X_{t_{cond}}$.

3.3.2 Context-aware Timbre Modeling

Speaker timbre features have long been viewed as global and time-invariant representations (Lin et al., 2021; Li et al., 2024; Pan et al., 2024d). However, recent studies (Jiang et al., 2024) have revealed a

[2] https://modelscope.cn/models/iic/speech_campplus_sv_zh_en_16k-common_advanced

close interdependence between timbre modeling and content information. Hence, drawing inspiration from this insight, we propose an innovative context-aware timbre modeling approach based on advanced cross-attention mechanism.

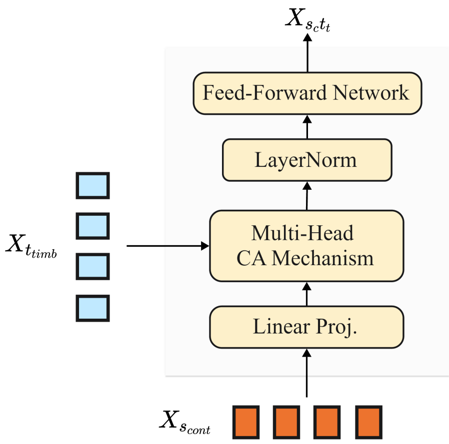

Figure 3: Schematic of the context-aware module.

As illustrated in Fig. 3, the CA-based module is designed to generate semantically aligned timbre features that harmonize the source linguistic content with the target timbre. Concretely, the CA-based module consists of a series of linear projection layers, multi-head cross-attention layers, layer normalization, and position feed-forward network (FFN), which can effectively facilitate the integration of $X_{s_{cont}}$ and $X_{t_{timb}}$. The source content $X_{s_{cont}}$ is used as the query, while the target timbre $X_{t_{timb}}$ serves as both the key and value. Finally, the extracted features $X_{s_c t_t}$ are interpolated to ensure dimensional compatibility with the ground truth, i.e., the source Mel-spectrogram, facilitating the subsequent training of the CFM model.

3.4 Conditional Flow Matching Model

In Takin-VC, to facilitate more efficient training and faster inference, we use a CFM model with optimal-transport (OT-CFM) to approximate the distribution of source Mel-spectrograms and generate predicted outputs conditioned on $X_{s_{ctt}}$ and $X_{t_{tcond}}$, all in a simulation-free manner.

Assume that the standard distribution and the target distribution are denoted as $p_0(x)$ and $p_1(x)$, respectively. The OT flow $\phi : [0,1] \times R^d \rightarrow R^d$ establishes the mapping between two density functions through the use of an ordinary differential equation (ODE):

$$\frac{d}{d_t}\phi_t(x) = v_t(\phi_t(x), t)$$
$$\phi_0(x) \sim p_0(x) = \mathcal{N}(x; 0, I), \phi_1(x) \sim p_1(x) \tag{2}$$

where v_t is a learnable time-dependent vector field, and $t \in [0, 1]$. Since multiple flows can generate this probability path, making it challenging to determine the optimal marginal flow, we adopt a simplified formulation, as proposed in (Tong et al., 2023b):

$$\phi^{OT}_{t,z}(x) = \mu_t(z) + \sigma_t(z)x$$
$$\sigma_t = 1-(1-\sigma_{min})t, \quad \mu_t = tz \tag{3}$$

where z represents the random variable, σ_{min} is a hyper-parameter set to 0.0001. Therefore, the training objective of the proposed CFM model can be formulated as:

$$\mathcal{L}_{cfm} = E_{t,p(x_0),q(x_1)} \cdot$$
$$\|(x_1-(1-\sigma)x_0) - v_t(\phi^{OT}_{t,x_1}(x_0)|\theta,h)\|^2 \tag{4}$$

where θ represents the parameters of the flow matching model, and h denotes the conditional set comprising $X_{t_{tcond}}$ and $X_{s_c t_t}$.

3.5 Training Objective

The training objective of the proposed Takin-VC is composed of two components, i.e., the RVQ commitment loss $\mathcal{L}_{vq}$ of the VQ module and $\mathcal{L}_{cfm}$.

$$\mathcal{L}_{total} = \mathcal{L}_{cfm} + \lambda\mathcal{L}_{vq}$$
$$\mathcal{L}_{vq}(X_s, \tilde{X}_s) = \sum_{i=1}^{N} \left\| X_{s_i} - \hat{X}_{s_i} \right\|_2^2 \tag{5}$$

Here, λ is a hyper-parameter that controls the weight of L_{vq}, and N represents the number of RVQ-based quantizers. In our implementation, λ is empirically set to 0.01, N is set to 1, and the codebook size of the RVQ-based quantizer is empirically determined to be 8200.

4 Experimental Setup

4.1 Baseline System

We conduct a comparative experiment of the performance in zero-shot voice conversion between our proposed Takin-VC approach and baseline systems, encompassing the following system: 1) DiffVC (Popov et al., 2021): A zero-shot VC system based on diffusion probabilistic modeling, which employs an averaged mel spectrogram aligned with phoneme to disentangle linguistic content and timbre information; 2) NS2VC[3]: A modified

[3] https://github.com/adelacvg/NS2VC

voice conversion edition of NaturalSpeech2 (Shen et al., 2023), which employ both diffusion and codec model to achieve zero-shot VC; 3) VALLE-VC (Wang et al., 2023a): We replace the original phoneme input with the semantic token extracted from the supervised model to make VALLE convert the timbre of source speech to the target speaker; 4) SEFVC (Li et al., 2024): A speaker embedding free voice conversion model, which is designed to learn and incorporate speaker timbre from reference speech. 5) StableVC (Yao et al., 2024b): A style controllable zero-shot voice conversion system, which employs dual adaptive gate attention to capture timbre and style information. 6) SeedVC (Liu, 2024): A zero-shot voice conversion system with an external timbre shifter and diffusion transformer.

4.2 Evaluation Metrics

Both subjective and objective metrics are employed to evaluate the performance of our Takin-VC and baseline systems. For **subjective metrics**, we employ naturalness mean opinion score (NMOS) to evaluate the naturalness of the generated samples and similarity mean opinion scores (SMOS) to evaluate the speaker similarity. We invite 20 professional participants to listen to the generated samples and provide their subjective perception scores on a 5-point scale: '5' for excellent, '4' for good, '3' for fair, '2' for poor, and '1' for bad. For **objective metrics**, we employ word error rate (WER), UTMOS, and speaker embedding cosine similarity (SECS) to evaluate the intelligibility, quality, and speaker similarity. Specifically: 1) We use a pre-trained CTC-based ASR model[4] to transcribe the generated speech and compare with ground-truth transcription; 2) We use a MOS prediction system that ranked first in the VoiceMOS Challenge 2022[5] to estimate the speech quality of the generated samples; 3) We use the WavLM-TDCNN SV model[6] to measure speaker similarity between generated speech and target speech. Furthermore, we introduce real-time factor (RTF) to evaluate the efficiency of Takin-VC.

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

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

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

4.3 Dataset

4.3.1 Small Scale Dataset

We employ the LibriTTS dataset to train our system and baseline systems, which contain 585 hours of recordings from 2,456 English speakers. We follow the official data split, using all training datasets for model training and "dev-clean" for model selection. The "test-clean" dataset is used to construct the evaluation set. All samples are processed at a 16kHz sampling rate.

4.3.2 Large Scale Dataset

To train a robust Takin VC model, we collected a dataset of approximately 500k hours. During the data collection process, we used an internally constructed data pipeline specifically designed for audio large model tasks. This pipeline includes signal-to-noise ratio (SNR) filtering, audio spectrum filtering (filtering out 24k audio with insufficient high-frequency information and pseudo 24k audio), VAD (Voice Activity Detection), LiD+ASR (Language Identification + Automatic Speech Recognition), speaker separation and identification, punctuation prediction, and background noise filtering. Regarding the test set, to validate the effectiveness of the Takin-VC model, we collected speech data from the internet that includes 100 non-preset speakers for evaluation. These speakers represent a variety of attributes such as gender, age, language, and emotion to ensure a comprehensive evaluation of the model's performance.

4.4 Model Configuration

For the content encoder part, in the first stage, we used the 12-layer HybridFormer-base model trained on a large dataset of 500K hours. For the WavLM part, we used the output features of the 6th layer. In the VQ part, we adopted a single-layer 8200 codebook with a hidden dimension of 1024, trained for 1 million steps on 100K hours of data. The fusion layer, as described in Sec. 3.2, is a simple module with several convolutional layers, an activation layer, and weighted summation. The Decoder adopts the same structure and configuration as HiFi-codec (Yang et al., 2023a).

In the part of timbre modeling and flow matching model, both the context-aware and memory-augmented modules use a transformer block with 8 heads, 6 layers, and a hidden size of 1024, with only the form of attention being different. The main structure of CFM uses a design of 10-layer U-net

Table 1: Comparison results of subjective and objective metrics between Takin-VC and the baseline systems in zero-shot voice conversion. Subjective metrics are computed with 95% confidence intervals and “GT" refers to ground truth samples.

	NMOS (↑)	SMOS (↑)	WER (↓)	UTMOS (↑)	SECS (↑)	RTF (↓)
GT	4.17±0.04	-	2.04	4.21	-	
DiffVC	3.75±0.05	3.66±0.07	3.08	3.68	0.61	0.294
NS2VC	3.65±0.07	3.51±0.06	2.94	3.64	0.53	0.347
VALLE-VC	3.80±0.06	3.79±0.04	2.77	3.72	0.65	3.678
SEFVC	3.68±0.05	3.76±0.06	3.75	3.51	0.63	0.187
StableVC	3.83±0.04	3.88±0.06	2.77	3.92	0.66	0.267
SeedVC	3.87±0.05	3.74±0.06	2.51	3.81	0.68	0.341
Takin-VC	**3.98**±0.04	**4.11**±0.05	**2.35**	**4.08**	**0.71**	**0.154**

plus 3 layers of ResNet block (He et al., 2016), with a hidden size of 1280. A Memory Fusion Block is inserted into the 10-layer U-net to enhance the speaker similarity of the generated audio.

For the small-data experiments, we use four A800 GPUs, whereas the large-data experiments are conducted on eight A800 servers. The batch size on each GPU is set to 16 with the AdamW optimizer using 1e-4 as the learning rate. In the inference section, experiments typically took 5 to 20 steps, with the final table uniformly adopting the results of 10 steps. The Classifier-Free Guidance (CFG) coefficient ranged from 0.1 to 1.0, with 0.7 used in the table. The specific experimental results will be detailed later.

5 Experimental Results

5.1 Experiments on small dataset

We first evaluate the performance of our proposed Takin-VC using subjective metrics. These metrics capture human perception of the enhanced speech's naturalness, intelligibility, and speaker similarity. As shown in Table 1, we can find that 1) our proposed system achieves the highest NMOS of 3.98, which is significantly higher than baseline systems; 2) the speaker similarity of our proposed system also outperforms all baseline systems. These results demonstrate that Takin-VC can achieve superior performance than the baseline system in the perceived aspect.

Furthermore, we evaluate the performance using objective metrics. The WER of our proposed system is 2.35, only slightly higher than the ground truth samples, indicating that the samples generated by Takin-VC exhibit better intelligibility. Moreover, Takin-VC achieves a UTMOS of 4.08 and an SECS of 0.71, demonstrating superior quality and similarity performance. Overall, the objective results of our proposed Takin-VC outperform all baseline systems and further corroborate the subjective findings. For inference efficiency, Takin-VC achieves the lowest RTF over all baseline systems, demonstrates superior real-time performance.

5.2 Experiments on large dataset

We employ the large scale dataset to train our proposed Takin-VC and investigate the performance in different conversion scenarios across different gender. As shown in Table 2, we divide the experiments into four groups: female to female (F2F), female to male (F2M), male to male (M2M), and male to female (M2F) to investigate performance differences. The results show that all metrics outperform Takin-VC trained on a smaller dataset, demonstrating that our proposed approach scales effectively. Additionally, the conversion results for same-gender conversions are slightly better than cross-gender conversions in both SMOS and SECS, while other metrics remain similar across all four group settings.

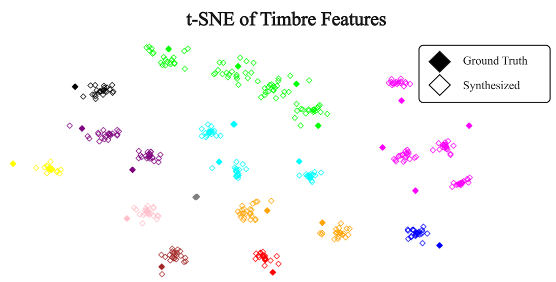

Figure 4: The t-SNE result of speaker similarity between ground truth samples and converted speech.

To further investigate the speaker similarity performance of our Takin-VC, we use the t-SNE method (Van der Maaten and Hinton, 2008) to visualize the speaker embeddings of 13 speakers, comparing the ground truth samples with the con-

Table 2: Detailed results of Takin-VC on different conversion scenarios. “F” and “M” represent the female and male, respectively.

	NMOS (↑)	SMOS (↑)	WER (↓)	UTMOS (↑)	SECS (↑)
GT	4.21±0.05	-	2.11	4.18	-
F2F	4.16±0.04	4.18±0.03	2.11	4.11	0.74
F2M	4.14±0.05	4.09±0.05	2.24	4.13	0.71
M2M	4.12±0.04	4.11±0.04	2.20	4.20	0.73
M2F	4.13±0.05	4.04±0.06	2.31	4.09	0.70

verted samples generated by Takin-VC. As shown in Figure 4, the embeddings of real and converted speech from the same speaker are closely clustered. This demonstrates that the speech generated by Takin-VC closely matches real human speech in both quality and speaker similarity.

Table 3: The ablation results for linguistic content extraction modules. “w/o ppg” and “w/o SSL” represent removing the HybridFormer or WavLM branch in our proposed hybrid content encoder, respectively.

	NMOS	SMOS	WER	UTMOS	SECS
Takin-VC	3.98±0.04	4.11±0.05	2.35	4.08	0.71
w/o ppg	3.74±0.04	3.07±0.04	2.79	3.91	0.45
w/o SSL	3.63±0.04	3.81±0.04	2.64	3.84	0.67

5.3 Ablation Study

We conduct two ablation experiments to evaluate the effectiveness of each proposed component in linguistic content extraction and timbre modeling. As shown in Table 3, SMOS results are significantly degraded, suggesting that only using the SSL model to extract linguistic content will result in timbre leakage. When we remove the SSL model in the hybrid content encoder and only use HybridFormer to extract linguistic content, we can find that NMOS and WER results degrade. This suggests that the conventional ASR encoder is less capable of disentangling linguistic content from the necessary paralinguistic information, underscoring the importance and effectiveness of our hybrid encoder in extracting linguistic content.

Additionally, we conduct an ablation study for timbre-related modules, results are shown in Table 4. We find significant degradation across all metrics when removing context-aware timbre modeling. It suggests that the system can not capture timbre information as well without the module, resulting in poor generation results. We observe a notable decline in speaker similarity when the voice print is removed from the attention module. We believe the voice print introduces a stronger timbre bias, which helps the attention module focus on capturing timbre information. Furthermore, when we remove the memory-augmented timbre modeling module, SMOS and SECS scores show significant degradation compared to the original Takin-VC, demonstrating the critical role of the memory module in improving timbre modeling. These ablation results demonstrate the effectiveness of each component proposed in our Takin-VC.

Table 4: The ablation results for timbre-related modules. “w/o con” represents removing content-aware timbre modeling and only employing voice print to extract timbre information. “w/o vp” represents removing the voice print, and “w/o mem” means removing the memory-augmented timbre modeling module.

	NMOS	SMOS	WER	UTMOS	SECS
Takin-VC	3.98±0.04	4.11±0.05	2.35	4.08	0.71
w/o con	3.77±0.04	3.61±0.04	3.01	3.85	0.58
w/o vp	3.94±0.05	3.89±0.04	2.51	3.98	0.61
w/o mem	3.92±0.04	3.75±0.05	2.44	4.01	0.52

6 Conclusion

In this study, we introduce Takin-VC, an effective framework for expressive zero-shot VC. Leveraging an adaptive fusion-based hybrid content encoder, Takin-VC integrates the complementary strengths of PPGs and quantized WavLM features in a learnable manner, thereby enhancing the naturalness and expressiveness of the converted speech. To improve speaker similarity, we propose an advanced memory-augmented module capable of extracting fine-grained conditional target timbre features. Additionally, we design a context-aware timbre modeling module to capture fused representations that effectively align and exploit the source content with target timbre elements. To enable stable training and fast inference, a conditional flow-matching model is presented reconstruct the Mel-spectrogram of the source speech. Experimental results demonstrate that Takin-VC outperforms all baseline systems regarding naturalness, expressiveness, speaker similarity, and real-time performance. Ablation studies further validate the effectiveness of each proposed component in our framework.

Limitations

This work primarily focuses on expressive zero-shot capabilities for speech generation, while zero-shot capabilities for speech editing remain limited and are a subject for future exploration. Additionally, while high-quality zero-shot VC has great potential, it can also lead to negative social impacts, such as voice impersonation of public figures and non-consenting individuals. We highlight this as a potential misuse of the technology to raise awareness of its ethical implications.

References

Philip Anastassiou, Zhenyu Tang, Kainan Peng, Dongya Jia, Jiaxin Li, Ming Tu, Yuping Wang, Yuxuan Wang, and Mingbo Ma. 2024. Voiceshop: A unified speech-to-speech framework for identity-preserving zero-shot voice editing. *arXiv preprint arXiv:2404.06674*.

Alan Baade, Puyuan Peng, and David Harwath. 2024. Neural codec language models for disentangled and textless voice conversion.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Grigory Bartosh, Dmitry Vetrov, and Christian A Naesseth. 2023. Neural diffusion models. *arXiv preprint arXiv:2310.08337*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. 2023. Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. *International Speech Communication Association*, pages 2283–2287.

Ha-Yeong Choi, Sang-Hoon Lee, and Seong-Whan Lee. 2024. Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17862–17870.

Trung Dang, Dung Tran, Peter Chin, and Kazuhito Koishida. 2022. 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)*, pages 6557–6561. IEEE.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Wendong Gan, Bolong Wen, Ying Yan, Haitao Chen, Zhichao Wang, Hongqiang Du, Lei Xie, Kaixuan Guo, and Hai Li. 2022. Iqdubbing: Prosody modeling based on discrete self-supervised speech representation for expressive voice conversion. *arXiv preprint arXiv:2201.00269*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

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

Liangjie Huang, Tian Yuan, Yunming Liang, Zeyu Chen, Can Wen, Yanlu Xie, Jinsong Zhang, and Dengfeng Ke. 2023. Limi-vc: A light weight voice conversion model with mutual information disentanglement. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shehzeen Hussain, Paarth Neekhara, Jocelyn Huang, Jason Li, and Boris Ginsburg. 2023. 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)*, pages 1–5.

Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Zhenhui Ye, Shengpeng Ji, Qian Yang, Chen Zhang, Pengfei Wei, Chunfeng Wang, et al. 2024. Mega-tts 2: Boosting prompting mechanisms for zero-shot speech synthesis. In *The Twelfth International Conference on Learning Representations*.

Sehoon Kim, Amir Gholami, Albert Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W Mahoney, and Kurt Keutzer. 2022. Squeezeformer: An efficient transformer for automatic speech recognition. *Advances in Neural Information Processing Systems*, 35:9361–9373.

Sungwon Kim, Kevin Shih, Joao Felipe Santos, Evelina Bakhturina, Mikyas Desta, Rafael Valle, Sungroh Yoon, Bryan Catanzaro, et al. 2024. P-flow: a fast and data-efficient zero-shot tts through speech prompting. *Advances in Neural Information Processing Systems*, 36.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2022. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*.

Dayong Li, Xian Li, and Xiaofei Li. 2023. Dvqvc: An unsupervised zero-shot voice conversion framework. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Junjie Li, Yiwei Guo, Xie Chen, and Kai Yu. 2024. 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)*, pages 12296–12300. IEEE.

Jheng-hao Lin, Yist Y Lin, Chung-Ming Chien, and Hung-yi Lee. 2021. S2vc: A framework for any-to-any voice conversion with self-supervised pretrained representations. *arXiv preprint arXiv:2104.02901*.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Alexander H Liu, Matt Le, Apoorv Vyas, Bowen Shi, Andros Tjandra, and Wei-Ning Hsu. 2023. Generative pre-training for speech with flow matching. *arXiv preprint arXiv:2310.16338*.

Songting Liu. 2024. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*.

Songxiang Liu, Yuewen Cao, Disong Wang, Xixin Wu, Xunying Liu, and Helen Meng. 2021. Any-to-many voice conversion with location-relative sequence-to-sequence modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1717–1728.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787.

Yin-Jyun Luo and Simon Dixon. 2024. Posterior variance-parameterised gaussian dropout: Improving disentangled sequential autoencoders for zero-shot voice conversion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11676–11680. IEEE.

Yu Pan, Yanni Hu, Yuguang Yang, Wen Fei, Jixun Yao, Heng Lu, Lei Ma, and Jianjun Zhao. 2024a. 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)*, pages 10021–10025. IEEE.

Yu Pan, Lei Ma, and Jianjun Zhao. 2024b. Promptcodec: High-fidelity neural speech codec using disentangled representation learning based adaptive feature-aware prompt encoders. *arXiv preprint arXiv:2404.02702*.

Yu Pan, Yuguang Yang, Yuheng Huang, Jixun Yao, Jingjing Yin, Yanni Hu, Heng Lu, Lei Ma, and Jianjun Zhao. 2023. Msac: Multiple speech attribute control method for reliable speech emotion recognition. *arXiv preprint arXiv:2308.04025*.

Yu Pan, Yuguang Yang, Heng Lu, Lei Ma, and Jianjun Zhao. 2024c. Gmp-atl: Gender-augmented multi-scale pseudo-label enhanced adaptive transfer learning for speech emotion recognition via hubert. *arXiv preprint arXiv:2405.02151*.

Yu Pan, Yuguang Yang, Jixun Yao, Jianhao Ye, Hongbin Zhou, Lei Ma, and Jianjun Zhao. 2024d. Ctefm-vc: Zero-shot voice conversion based on content-aware timbre ensemble modeling and flow matching. *arXiv preprint arXiv:2411.02026*.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov, and Jiansheng Wei. 2021. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. *arXiv preprint arXiv:2109.13821*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. 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*, pages 234–241. Springer.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian.

2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Huaizhen Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. 2022. Avqvc: One-shot voice conversion by vector quantization with applying contrastive learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4613–4617. IEEE.

Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, et al. 2022. The voiceprivacy 2020 challenge: Results and findings. *Computer Speech & Language*, 74:101362.

Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. 2023a. Simulation-free schr\" odinger bridges via score and flow matching. *arXiv preprint arXiv:2307.03672*.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. 2023b. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2(3).

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. 2023c. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

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

Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. 2021. Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. *arXiv preprint arXiv:2106.10132*.

Zhichao Wang, Yuanzhe Chen, Lei Xie, Qiao Tian, and Yuping Wang. 2023b. Lm-vc: Zero-shot voice conversion via speech generation based on language models. *IEEE Signal Processing Letters*.

Da-Yi Wu, Yen-Hao Chen, and Hung-Yi Lee. 2020. Vqvc+: One-shot voice conversion by vector quantization and u-net architecture. *arXiv preprint arXiv:2006.04154*.

Da-Yi Wu and Hung-yi Lee. 2020. One-shot voice conversion by vector quantization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7734–7738. IEEE.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023a. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*.

SiCheng Yang, Methawee Tantrawenith, Haolin Zhuang, Zhiyong Wu, Aolan Sun, Jianzong Wang, Ning Cheng, Huaizhen Tang, Xintao Zhao, Jie Wang, et al. 2022a. Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion. *arXiv preprint arXiv:2208.08757*.

Yuguang Yang, Yu Pan, Jingjing Yin, Jiangyu Han, Lei Ma, and Heng Lu. 2023b. Hybridformer: Improving squeezeformer with hybrid attention and nsr mechanism. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yuguang Yang, Yu Pan, Jingjing Yin, and Heng Lu. 2022b. Lmec: Learnable multiplicative absolute position embedding based conformer for speech recognition. *arXiv preprint arXiv:2212.02099*.

Jixun Yao, Qing Wang, Pengcheng Guo, Ziqian Ning, Yuguang Yang, Yu Pan, and Lei Xie. 2024a. Musa: Multi-lingual speaker anonymization via serial disentanglement. *arXiv preprint arXiv:2407.11629*.

Jixun Yao, Yuguang Yan, Yu Pan, Ziqian Ning, Jiaohao Ye, Hongbin Zhou, and Lei Xie. 2024b. Stablevc: Style controllable zero-shot voice conversion with conditional flow matching. *arXiv preprint arXiv:2412.04724*.

Jixun Yao, Yuguang Yang, Yi Lei, Ziqian Ning, Yanni Hu, Yu Pan, Jingjing Yin, Hongbin Zhou, Heng Lu, and Lei Xie. 2024c. 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)*, pages 10571–10575. IEEE.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.

Xintao Zhao, Feng Liu, Changhe Song, Zhiyong Wu, Shiyin Kang, Deyi Tuo, and Helen Meng. 2022. 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)*, pages 7022–7026. IEEE.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. 2023. Denoising diffusion bridge models. *arXiv preprint arXiv:2309.16948*.