TransferSinger: Zero-Shot Singing Voice Synthesis with Style Transfer

Anonymous ACL submission

Abstract

Zero-shot Singing Voice Synthesis (SVS) with style transfer aims to generate high-quality singing voices of unseen timbres and styles (including singing methods, rhythm, techniques, and pronunciation) from the prompt audio. However, the multifaceted nature of singing voice styles poses a significant challenge for comprehensive modeling and effective transfer. Furthermore, existing SVS models often fail to generate singing voices with a wealth of stylistic nuances for unseen singers. In this paper, we introduce TransferSinger, a novel zero-shot SVS model that primarily employs three modules to address these challenges: 1) the style encoder that employs a Vector Quantization (VQ) model to condense style information into a compact latent space, thus facilitating subsequent predictions; 2) the Style and Duration Language Model (S&D-LM), which concurrently predicts style information and phoneme duration, thereby enhancing both; and 3) the style adaptive decoder that uses a novel style adaptive normalization method to generate singing voices with enhanced details. Experimental results show that TransferSinger outperforms baseline models in terms of both synthesis quality and singer similarity across various tasks, including zero-shot SVS, controllable style synthesis, cross-lingual style transfer, and speech-to-singing style transfer. Singing voice samples can be accessed at https://transfersinger.github.io/.

1 Introduction

Singing Voice Synthesis (SVS) is dedicated to generating high-quality singing voices by utilizing lyrics and musical notations. The pipeline of traditional SVS systems involves an acoustic model to transform musical notations and lyrics into F0 and mel-spectrogram, which are then synthesized into the target singing voice by a vocoder.

Recent years have seen significant advancements in SVS technology, with remarkable results being generated (Zhang et al., 2022b; Kim et al., 2023; Cho et al., 2022; Liu et al., 2022a). However, the increasing demand for personalized timbre and styles in singing voices presents a challenge to current SVS models. Unlike traditional SVS tasks, the zero-shot SVS with style transfer seeks to generate high-quality singing voices with unseen timbres and styles of the prompt audio. Personal singing styles mainly include **singing methods** (like bel canto and pop), **rhythm** (including the stylistic handling of individual notes and transitions between them), **techniques** (such as vibrato and falsetto), and **pronunciation** (like articulation and accent). Despite this, traditional SVS methods lack necessary mechanisms to model and transfer these personal styles effectively. Their performance tends to decline for unseen singers, as these methods generally assume that target singers are identifiable during the training phase (Zhang et al., 2023).

Presently, the zero-shot SVS with style transfer task primarily faces two major challenges: 1) The multifaceted nature of singing styles presents a substantial challenge for comprehensive modeling and effective transfer. Previous models employ pre-trained models to model global styles (Cooper et al., 2020). StyleSinger (Zhang et al., 2023) uses a Residual Quantization (RQ) model to capture detailed styles. However, these models focus on limited aspects of styles, neglecting styles like singing methods. Moreover, they fail to extend to cross-lingual speech and singing styles and do not conduct controllable style synthesis. 2) Existing SVS models often fail to generate singing voices rich in stylistic nuances for unseen singers. Diffsinger (Liu et al., 2022a) employs a diffusion decoder to capture the intricacies of singing voices. RMSSinger (He et al., 2023) uses a post-net to enhance synthesis quality. However, these methods do not adequately incorporate style information into the synthesis of singing voices, leading to results that lack style variations in zero-shot tasks.

To address these challenges, we introduce TransferSinger, a model designed to transfer unseen timbre and styles (like singing methods, rhythm, techniques, and pronunciation) from prompts to synthesize high-quality target singing voices. To model styles of the prompt audio, we propose a style encoder that uses a vector quantization (VQ) model with ℓ_2 normalization for enhancing training stability and reconstruction quality. To transfer styles to the target, we put forth the Style and Duration Language Model (S&D-LM). The S&D-LM incorporates a multi-task language module to concurrently predict both style information and phoneme duration, thereby enhancing both predictions. To generate singing voices rich in stylistic nuances, we introduce the style adaptive decoder, which employs a novel style adaptive normalization method to refine mel-spectrograms with style information. Our experimental results illustrate that TransferSinger outperforms baseline models in terms of both synthesis quality and singer similarity across various tasks, including zero-shot SVS, controllable style synthesis, cross-lingual style transfer, and speech-to-singing style transfer. The main contributions of this work can be summarized as follows:

- We introduce the style encoder using a VQ model with ℓ_2 normalization, and the Style and Duration Language Model (S&D-LM) to predict style information and phoneme duration, addressing style modeling and transfer.
- We propose the style adaptive decoder to generate intricately detailed singing voices using a novel style adaptive normalization method.
- TransferSinger is the first method for the SVS with style transfer task that successfully models styles of cross-lingual speech and singing data, and achieves controllable style synthesis.
- Our experimental results demonstrate that TransferSinger surpasses baseline models in both synthesis quality and singer similarity across various tasks: zero-shot SVS, controllable style synthesis, cross-lingual style transfer, and speech-to-singing style transfer.

2 Related Works

2.1 Singing Voice Synthesis

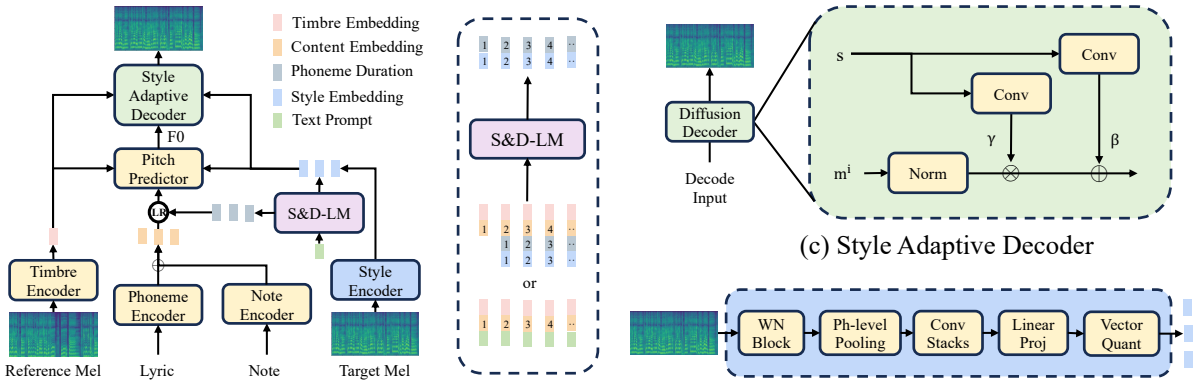
Singing Voice Synthesis (SVS) has emerged as a dynamic field focused on generating high-quality singing voices from provided lyrics and musical scores. VISinger (Zhang et al., 2022b) introduces a comprehensive, end-to-end SVS system, building upon the VITS model (Kim et al., 2021). Choi and Nam (2022) presents a melody-unsupervised model that only requires pairs of audio and lyrics, thus eliminating the need for temporal alignment. For multi-singer tasks, both M4Singer (Zhang et al., 2022a) and Multi-Singer (Huang et al., 2021) make substantial contributions by releasing multi-singer Chinese song datasets. Recently, RMSSinger (He et al., 2023) has proposed a diffusion pitch predictor to forecast F0 and UV, and a diffusion-based post-net to improve synthesis quality. Nonetheless, these methods are based on the assumption that the target singer is visible during the training phase and they do not adequately incorporate style information into synthesis, with few style variations in generated audio for zero-shot SVS tasks.

2.2 Style Modeling and Transfer

Modeling and transferring styles remains a pivotal area of research within the audio domain, with past models predominantly leveraging pre-trained models to capture a limited array of styles (Kumar et al., 2021). Atmaja and Sasou (2022) evaluates the performance of wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022) in speech emotion recognition tasks. Generspeech (Huang et al., 2022a) integrates global and local style adaptors to capture speech styles. YourTTS (Casanova et al., 2022) conditions the affine coupling layers of the flow-based decoder to handle zero-shot tasks. Mega-TTS (Jiang et al., 2023) decomposes speech into multiple attributes and models prosody using a language model. Recently, StyleSinger (Zhang et al., 2023) has employed a Residual Quantization (RQ) model to capture detailed styles in singing voices. Although these approaches have made strides in capturing style, there remains a notable gap in fully modeling styles like singing methods and extending these capabilities to cross-lingual speech and singing styles, as well as in controllable style synthesis.

3 TransferSinger

In this section, we first overview the proposed TransferSinger. Then, we introduce several critical components including the style encoder, the style adaptive decoder, and the Style and Duration Language Model (S&D-LM). Finally, we elaborate on the training and inference procedures.



(a) Overall (b) S&D-LM (c) Style Adaptive Decoder (d) Style Encoder

Figure 1: The architecture of TransferSinger. In Figure (a), S&D-LM represents the Style and Duration Language Model, while LR stands for length regulator. In Figure (b), the S&D-LM autoregressively predicts both style information and phoneme duration. In Figure (c), intermediate mel-spectrograms are refined with style information in the style adaptive decoder. In Figure (d), the style encoder extracts style information from mel-spectrograms.

3.1 Overview

The architecture of TransferSinger is depicted in Figure 1(a). We disentangle singing voices into separate representations for content, style (including singing methods, rhythm, techniques, and pronunciation), and timbre. Regarding content representation, lyrics are encoded through a phoneme encoder, while a note encoder captures musical notes. For style representation, we use a VQ module within the style encoder to condense style information into a compact latent space, thus facilitating subsequent predictions. We use $\ell 2$ normalization in the VQ model to enhance training stability and reconstruction quality. In terms of timbre representation, we feed a prompt mel-spectrogram, sampled from different audio of the same singer, into the timbre encoder to disentangle the timbre and content information. We then temporally average the output to obtain a one-dimensional global timbre vector. Then, we utilize the Style and Duration Language Model (S&D-LM) to simultaneously predict style information and phoneme duration since styles and duration of singing voices are closely related, and a composite module benefits both. Next, we use the content, style, and timbre representations as inputs to the pitch predictor with diffusion-based architecture (He et al., 2023) for F0 prediction. Finally, we use the style adaptive decoder to generate the target mel-spectrogram. The style adaptive decoder generates intricately detailed singing voices using a novel style adaptive normalization method. During inference, we use the content from the given lyrics and notes, the timbre extracted from the prompt audio, and style information, phoneme duration predicted by S&D-LM to synthesize the target singing voice. Additionally, we can substitute a text prompt (like alto pop vibrato) as input to S&D-LM for style information and phoneme duration prediction. Please refer to Appendix A for more details.

3.2 Style Encoder

To comprehensively capture styles (such as singing methods, rhythm, techniques, and pronunciation) from mel-spectrograms, we introduce the style encoder. As illustrated in Figure 1 (d), the input mel-spectrogram is initially refined through WaveNet blocks before being condensed into phoneme-level hidden states by a pooling layer based on the phoneme boundary. Subsequently, the convolution stacks capture phoneme-level correlations. Then, we employ a linear projection from the convolution stacks' output into a low-dimensional latent variable space for code index lookup, which could significantly increase the codebook's usage (Yu et al., 2021). The vector quantization (VQ) layer (Van Den Oord et al., 2017) then employs these inputs x to generate phoneme-level style representations, establishing an information bottleneck that effectively eliminates non-style information. Through the dimensionality reduction of the linear projection and the bottleneck of VQ, we achieve a decoupling of styles from the timbre and content information. To enhance training stability and improve reconstruction quality, we apply ℓ_2 normalization to the encoded latent variables $z_e(x)$ and the code-

book latent variables e. This approach has proven useful in the VQ-related tasks of image domain (Yu et al., 2021). By mapping all latent variables onto a sphere, the Euclidean distance of ℓ_2-normalized latent variables $|\ell_2(z_e(x)) - \ell_2(e_j)|_2^2$ is transformed into the cosine similarity between the two vectors $z_e(x)$ and e. To train the style encoder, we use the VQ loss with ℓ_2 normalization:

$$\mathcal{L}_{VQ} = \|sg[\ell_2(z_e(x))] - \ell_2(e)\|_2^2 + \beta\|\ell_2(z_e(x)) - \ell_2(sg[e])\|_2^2, \quad (1)$$

where $\mathrm{sg}(\cdot)$ is the stop-gradient operator, β is a commitment loss hyperparameter.

3.3 Style Adaptive Decoder

The dynamic nature of singing voices poses a substantial challenge to traditional mel-decoders, which often fail to capture the intricacies of mel-spectrograms effectively. Furthermore, using VQ to extract style information is inherently lossy (Razavi et al., 2019), and closely related styles can easily be encoded into identical codebook indices. Consequently, if we employ traditional mel decoders here, our synthesized singing voices may become rigid and lacking in stylistic variation. To address these challenges, we introduce the style adaptive decoder, which utilizes a novel style adaptive normalization method. While the adaptive instance normalization method has been widely used in image synthesis tasks (Zheng et al., 2022; Dumoulin et al., 2016), our work is pioneering in refining mel-spectrograms using style information in the singing field. Our approach can infuse stylistic variations into mel-spectrograms, thereby generating more believable and diverse audio results, even when the same style quantization index is used for closely related styles in decoder inputs.

As depicted in Figure 1 (c), our style adaptive decoder is fundamentally based on an 8-step diffusion-based decoder (Huang et al., 2022b). We utilize FFT as the denoiser and enhance it by incorporating multiple layers of our style adaptive normalization. In our model, we denote the intermediate mel-spectrogram of the i-th layer in the diffusion decoder denoiser as m^i. In i-th layer, m^{i-1} is initially normalized using a normalization method and then adapted by the scale and bias that are computed from the style embedding s. To be more detailed, m^i is given by:

$$m^i = \phi_\gamma(s)\frac{m^{i-1} - \mu(m^{i-1})}{\sigma(m^{i-1})} + \phi_\beta(s), \quad (2)$$

where the functions $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation calculation. We employ Layer Normalization (Ba et al., 2016) as the normalization method here. $\phi_\gamma(\cdot)$ and $\phi_\beta(\cdot)$ are two learned affine transformations for converting the style representation s to the scaling and bias values. As $\phi_\gamma(\cdot)$ and $\phi_\beta(\cdot)$ inject the stylistic variant information, it encourages similar decoder inputs entry to generate plausible and diverse mel-spectrograms.

In the training phase, we first apply Mean Absolute Error (MAE) loss. Let x_0 be the original clean data, while x_θ denotes the denoised data sample:

$$\mathcal{L}_{mae} = \left\| x_\theta\left(\alpha_t x_0 + \sqrt{1-\alpha_t^2}\epsilon\right) - x_0 \right\|, \quad (3)$$

where $\alpha_t = \prod_{i=1}^{t}\sqrt{1-\beta_i}$. β_t represents the predefined fixed noise schedule at diffusion step t. Additionally, ϵ is randomly sampled from a normal distribution $\mathcal{N}(0, I)$. Furthermore, we also incorporate the Structural Similarity Index (SSIM) loss (Wang et al., 2004) to the reconstruction loss:

$$\mathcal{L}_{ssim} = 1 - SSIM\left(x_\theta\left(\alpha_t x_0 + \sqrt{1-\alpha_t^2}\epsilon\right), x_0\right). \quad (4)$$

3.4 S&D-LM

Singing styles (like singing methods, rhythm, techniques, and pronunciation) usually exhibit both local and long-term dependencies, and they change rapidly over time with a weak correlation to content. This makes the conditional language model inherently ideal for generating style information. Meanwhile, phoneme duration is rich in variations and closely related to singing styles. Therefore, we propose the Style and Duration Language Model (S&D-LM) to simultaneously predict style information and phoneme duration, serving as a multi-task module to enhance both. Through S&D-LM, we can generate high-quality target singing voices with unseen timbre and styles of the prompt audio.

To be more specific, given the lyrics $\tilde{l}$, notes $\tilde{n}$ of the target, and lyrics l, notes n, mel-spectrogram m of the prompt audio, our goal is to synthesize the high-quality target singing voice's mel-spectrogram $\tilde{m}$ with unseen timbre and styles of the prompt audio. Initially, we use different encoders to extract the timbre information t, content information c, and style information s of the prompt audio and the target content information $\tilde{c}$:

$$s = E_{style}(m), t = E_{timbre}(m), \quad c = E_{content}(l, n), \tilde{c} = E_{content}(\tilde{l}, \tilde{n}), \quad (5)$$

where E denotes encoders for each attribute. Given that the target timbre $\tilde{t}$ is anticipated to mirror the prompt audio, we also require the target style information $\tilde{s}$ to generate the target mel-spectrogram $\tilde{m}$. Utilizing the powerful in-context learning capabilities of language models, we design the S&D-LM to predict $\tilde{s}$. Concurrently, we also use the S&D-LM to predict the target phoneme duration $\tilde{d}$, leveraging the strong correlation between phoneme duration and styles in singing voices to enhance both predictions. Our S&D-LM is based on a decoder-only transformer-based architecture (Brown et al., 2020). We concatenate the prompt phoneme duration d, the prompt style information s, along with prompt content c, target content $\tilde{c}$, and target timbre $\tilde{t}$ to form the input. The autoregressive prediction process can be formulated as follows:

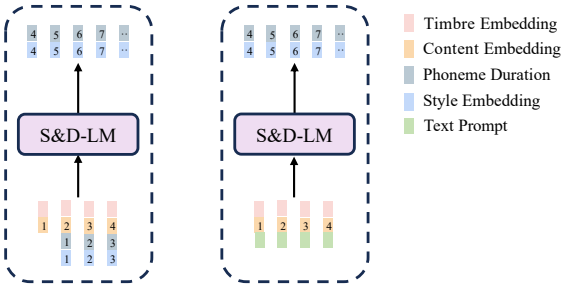
$$p\left(\tilde{s}, \tilde{d} \mid s, d, c, \tilde{t}, \tilde{c}; \theta\right) = \prod_{t=0}^{T} p\left(\tilde{s}_t, \tilde{p}_t \mid \tilde{s}_{<t}, \tilde{d}_{<t}, s, d, c, \tilde{t}, \tilde{c}; \theta\right), \tag{6}$$

where θ is the parameter of our S&D-LM. We train the S&D-LM in the teacher-forcing mode using the cross-entropy loss for the predicted style information and the Mean Squared Error (MSE) loss for the phoneme duration. Finally, we use P to denote the pitch predictor and D to represent our style adaptive decoder, the formula for synthesizing the target F0 and mel-spectrogram is:

$$\begin{aligned} F0 &= P(\tilde{s}, \tilde{d}, \tilde{t}, \tilde{c}), \\ \tilde{m} &= D(\tilde{s}, \tilde{d}, \tilde{t}, \tilde{c}, F0). \end{aligned} \tag{7}$$

3.5 Training and Inference Procedures

Training Procedures The final loss terms of TransferSinger in the training phase consist of the following parts: 1) VQ loss $\mathcal{L}_{VQ}$: the VQ loss with $\ell 2$ normalization for the style encoder; 2) Pitch reconstruction loss $\mathcal{L}_{gdiff}, \mathcal{L}_{mdiff}$: the Gaussian diffusion loss and the multinomial diffusion loss between the predicted and the GT pitch spectrogram for the pitch predictor; 3) Mel reconstruction loss $\mathcal{L}_{mae}, \mathcal{L}_{ssim}$: the MAE loss and the SSIM loss between the predicted and the GT mel-spectrogram for the style adaptive decoder. 4) Duration prediction loss $\mathcal{L}_{dur}$: the MSE loss between the predicted and the GT phoneme-level duration in log scale for S&D-LM in the teacher-forcing mode; 5) Style prediction loss $\mathcal{L}_{style}$: the cross-entropy loss between the predicted and the GT style information for S&D-LM in the teacher-forcing mode.



(a) Zero-Shot SVS (b) Text Prompt

Figure 2: Inference procedure of TransferSinger. In Figure (a), the S&D-LM extracts information from the prompt audio to predict the target style information and phoneme duration, while in Figure (b), the S&D-LM utilizes the text prompt to predict them.

Inference of Zero-Shot SVS Refer to Figure 2 (a) and Equation 6, during the inference phase of zero-shot SVS, we use the prompt audio to extract c, t, s, d, as well as the target content $\tilde{c}$ as inputs for the S&D-LM, and obtain $\tilde{s}, \tilde{u}$. Then, since the target's timbre and prompt remain unchanged, according to Equation 7, we concatenate the content $\tilde{c}$, timbre $\tilde{t}$, style information $\tilde{s}$, and phoneme duration $\tilde{d}$ of the target to generate F0 by the pitch predictor, and mel-spectrogram $\tilde{m}$ by the style adaptive decoder. Therefore, the generated target singing voice can effectively transfer the timbre and styles of the prompt audio. In the cross-lingual experiments, the lyrics language of the prompt and the target are different (like English and Chinese), but the rest of the process remains the same. In the speech-to-singing experiments, we use speech data as the prompt audio, allowing the target singing voice to transfer the timbre and styles of the speech data, with the rest of the process remaining consistent.

Training and Inference with Text Prompts As shown in Figure 1 (b), during the training phase of the S&D-LM, we use a text prompt (like alto pop vibrato) to replace the s, d extracted from the prompt audio, and combine the text prompt with c, t, $\tilde{c}$ to generate $\tilde{s}, \tilde{u}$. As shown in Figure 2 (b), during inference, we use a text prompt to replace the s, d extracted from the prompt audio, thus generating the target $\tilde{s}, \tilde{u}$, with the rest of the process remaining consistent with the zero-shot SVS task. Each text prompt encompasses a singing style class, containing a variety of styles. Through these text prompts, we can generate singing voices with reference timbre and independent specific style classes, thus achieving controllable style synthesis.

4 Experiments

4.1 Experimental Setup

In this section, we present the datasets utilized by TransferSinger, delve into the implementation and training details, discuss the evaluation methodologies, and introduce the baseline models.

Dataset Existing open-source singing datasets are relatively sparse. In this endeavor, we collect and annotate a cross-lingual dataset (16 singers, 28h Chinese and English singing) by recruiting professional singers in a professional recording studio. Moreover, we enrich our data by incorporating the M4Singer dataset (Zhang et al., 2022a) (20 singers, 30h Chinese singing), the OpenSinger dataset (Huang et al., 2021) (93 singers, 85h Chinese singing), the AISHELL-3 dataset (Shi et al., 2021) (218 singers, 85h Chinese speech), and a subset of the PopBuTFy database (Liu et al., 2022b) (20 singers, 18h English speech and singing). Then, we manually annotate these singing data with style class labels based on vocal ranges, singing methods, and techniques (like alto pop vibrato). Finally, we randomly designate 40 singers (including singing and speech, Chinese and English) as the unseen test set to evaluate TransferSinger in zero-shot tasks. Please refer to Appendix B for more details.

Implementation Details We set the sample rate to 48000Hz, the window size to 1024, the hop size to 256, and the number of mel bins to 80 to derive mel-spectrograms from raw waveforms. The default size of the codebook for VQ is 512. The S&D-LM model is a decoder-only architecture with 8 Transformer layers and 512 embedding dimensions. Please refer to Appendix A.1 for more details.

Training Details We train our model using four NVIDIA 3090Ti GPUs. The Adam optimizer is used with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The main SVS model takes 300k steps and the S&D-LM model takes 100k steps to train until convergence. Output mel-spectrograms of the style adaptive decoder are transformed into singing voices using a pre-trained HiFi-GAN vocoder (Kong et al., 2020).

Evaluation Details We use both objective and subjective evaluation metrics to validate the performance of TransferSinger. For subjective metrics, we employ the Mean Opinion Score (MOS) to judge **synthesis quality** (including clarity, naturalness, and rich stylistic details) and use the Similarity Mean Opinion Score (SMOS) (Min et al., 2021) to assess **singer similarity** (in terms of timbre and styles) between the synthesized and the prompt audio. Both these metrics are rated from 1 to 5 and reported with 95% confidence intervals. In the ablation study, we employ the Comparative Mean Opinion Score (CMOS) to gauge synthesis quality, along with the Comparative Similarity Mean Opinion Score (CSMOS) to evaluate singer similarity. For objective metrics, we use the Singer Cosine Similarity (Cos) to judge singer similarity, and the F0 Frame Error (FFE) to quantify synthesis quality. Please refer to Appendix C for more details.

Baseline Models We conduct a comprehensive comparative analysis of synthesis quality and singer similarity for TransferSinger with other models. Firstly, we compare our model with the original target ground truth (GT) and the audio generated by HiFi-GAN (GT (vocoder)). Next, we integrate a note encoder into two well-performing speech models that conduct style transfer, training them on speech and singing data to compare their performance, including YourTTS (Casanova et al., 2022) and Mega-TTS (Jiang et al., 2023). Subsequently, we also compare with the best-performing traditional SVS model, RMSSinger (He et al., 2023). In this comparison, we use the prompt singer embedding to synthesize the target singing voice. Lastly, we compare with StyleSinger (Zhang et al., 2023), the first model that conducts style transfer for zero-shot SVS. We use the open-source code of YourTTS and make necessary modifications. As for RMSSinger, Mega-TTS, and StyleSinger, we carefully reproduced their works independently.

4.2 Main Results

Zero-Shot SVS with Style Transfer To assess the performance of TransferSinger and baseline models in the zero-shot SVS with style transfer task, we randomly select samples with unseen singers from the test set as targets and different utterances from the same singers to form prompts. As shown in Table 1, we have the following findings: 1) TransferSinger exhibits outstanding synthesis quality, as indicated by the highest MOS and the lowest FFE. This underscores the model's impressive adaptability in managing zero-shot SVS scenarios. 2) TransferSinger also excels in singer similarity, as denoted by the highest SMOS and Cos. This highlights our model's superior ability to model and transfer different singing styles precisely, thanks to the innovative design of our components. Our style adaptive decoder effectively improves the rich stylistic details of synthesis quality, rendering the singing voices more natural and of superior qual-

Method	MOS ↑	SMOS ↑	Cos ↑	FFE ↓
GT	4.56 ± 0.07	-	-	-
GT (vocoder)	4.32 ± 0.09	4.38 ± 0.06	0.97	0.04
YourTTS (Casanova et al., 2022)	3.64 ± 0.08	3.74 ± 0.07	0.81	0.35
Mega-TTS (Jiang et al., 2023)	3.75 ± 0.07	3.87 ± 0.06	0.83	0.29
RMSSinger (He et al., 2023)	3.86 ± 0.06	3.80 ± 0.08	0.86	0.31
StyleSinger (Zhang et al., 2023)	3.93 ± 0.07	3.99 ± 0.08	0.90	0.27
TransferSinger (ours)	$\mathbf{4.06 \pm 0.08}$	$\mathbf{4.27 \pm 0.08}$	**0.92**	**0.23**

Table 1: Synthesis quality and singer similarity of zero-shot SVS with style transfer. For subjective measurement, we employ MOS and SMOS. In objective evaluation, we utilize Cos and FFE.

Method	MOS ↑	SMOS ↑
YourTTS	3.63 ± 0.07	3.70 ± 0.06
Mega-TTS	3.72 ± 0.09	3.83 ± 0.08
RMSSinger	3.83 ± 0.05	3.78 ± 0.07
StyleSinger	3.91 ± 0.06	3.96 ± 0.09
TransferSinger (ours)	$\mathbf{4.03 \pm 0.08}$	$\mathbf{4.22 \pm 0.05}$

Table 2: Synthesis quality and singer similarity comparisons for controllable style synthesis.

Method	MOS ↑	SMOS ↑
YourTTS	3.58 ± 0.08	3.55 ± 0.09
Mega-TTS	3.65 ± 0.06	3.71 ± 0.07
RMSSinger	3.77 ± 0.10	3.64 ± 0.09
StyleSinger	3.84 ± 0.07	3.82 ± 0.06
TransferSinger(ours)	$\mathbf{3.95 \pm 0.09}$	$\mathbf{4.08 \pm 0.08}$

Table 3: Synthesis quality and singer similarity comparisons for cross-lingual style transfer.

ity. Meanwhile, our style encoder shows an excellent capability for modeling styles across a wide range of categories. Finally, the S&D-LM delivers excellent prediction results for style information and phoneme duration, significantly contributing to synthesis quality and singer similarity.

As shown in Figure 3, TransferSinger not only displays greater details in the mel-spectrogram, but also effectively learns the technique, pronunciation, and rhythm of the prompt audio. In contrast, other models lack details in mel-spectrograms, and their pitch curves remain flat, failing to transfer styles. Upon listening to demos, it is clear that our model effectively transfers the timbre, singing methods, rhythm, techniques, and pronunciation of prompts.

Controllable Style Synthesis We randomly select singing voice samples from the unseen test set and use them as prompts for the baseline models. Then, we use the style labels (like alto pop vibrato) of these prompts as text prompts for TransferSinger to perform controllable style synthesis, and these audio severs as timbre prompts. Moreover, we randomly utilize content information from all songs in the dataset as the target. As shown in Table 2, we use MOS and SMOS to compare TransferSinger with text prompts against other models. TransferSinger with text prompts surpasses other baseline models in synthesis quality and singer similarity. Apart from the advantages of our models in style modeling and transfer, the text prompt encompasses a comprehensible style class, enabling the use of abundant styles to synthesize controllable singing voices. Simultaneously, since the text and timbre prompts can be independent, we can synthesize the controllable target singing voice using the prompt timbre and the specified style class.

Cross-Lingual Style Transfer To test the cross-lingual style transfer performance of various models, we alternately use unseen Chinese and English data as prompts and targets for inference, using MOS and SMOS as evaluation criteria. As shown in Table 3, our TransferSinger outperforms other models regarding synthesis quality and singer similarity. Benefiting from the modeling capability of our style encoder for rich cross-lingual styles, the assistance of the style adaptive decoder in generating singing voices with rich style details, and the powerful prediction capability of the S&D-LM for phoneme duration and style information, our model performs well in a cross-lingual environment.

Speech-to-Singing Style Transfer We conducted experiments on speech-to-singing style transfer and used MOS and SMOS to compare the performance of various models. To be specific, we used unseen speech audio as the prompt audio to transfer timbre and styles to the target singing voice. As shown in Table 4, we found that both synthesis quality and singer similarity of TransferSinger are superior to

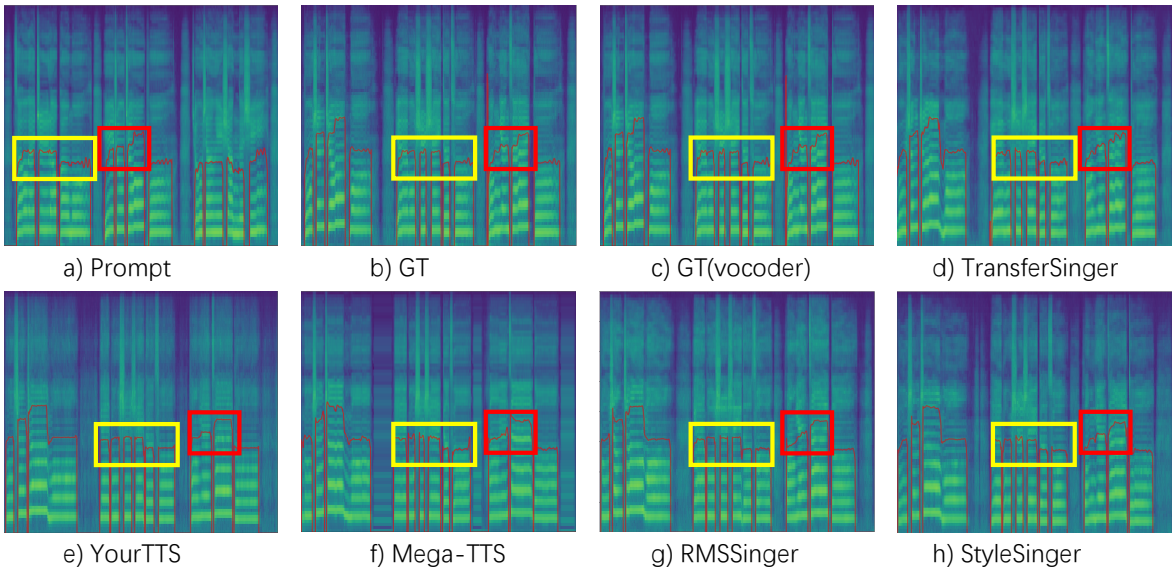


Figure 3: Mel-spectrograms depicting the results of zero-shot SVS with style transfer. TransferSinger effectively captures the rhythm and pronunciation in red boxes, along with the vibrato technique and rhythm in yellow boxes.

Method	**MOS** ↑	**SMOS** ↑
YourTTS	3.53 ± 0.08	3.51 ± 0.07
Mega-TTS	3.60 ± 0.09	3.66 ± 0.08
RMSSinger	3.73 ± 0.07	3.59 ± 0.06
StyleSinger	3.81 ± 0.10	3.80 ± 0.09
TransferSinger (ours)	$\mathbf{3.92 \pm 0.08}$	$\mathbf{4.03 \pm 0.07}$

Table 4: Synthesis quality and singer similarity comparisons for speech-to-singing style transfer.

Setting	**CMOS**	**CSMOS**
TransferSinger	0.00	0.00
w/o SAD	-0.21	-0.19
w/o DM	-0.12	-0.23

Table 5: Synthesis quality and singer similarity comparisons for ablation study. SAD denotes style adaptive decoder and DM means duration model of S&D-LM.

those of the baseline models. This demonstrates the excellent ability of our model in both speech and singing style modeling and transfer.

4.3 Ablation Study

As depicted in Table 5, we undertake ablation studies to showcase the efficacy of various designs within TransferSinger. We use CMOS to test the variation in synthesis quality, and CSMOS to measure the changes in singer similarity. 1) When we eliminate the style adaptive decoder and use an 8-step diffusion decoder (Huang et al., 2022b), both synthesis quality and singer similarity decline, indicating the enhancement our method brings to the diversity of styles in singing voices. 2) When we only predict styles in the S&D-LM and use a simple duration predictor (Ren et al., 2020) to predict phoneme duration, both synthesis quality and singer similarity decrease. This demonstrates the mutual benefits of our method for predicting both phoneme duration and style information.

5 Conclusion

In this paper, we introduce TransferSinger, a model designed to transfer unseen timbre and styles (like singing methods, rhythm, techniques, and pronunciation) from prompts to synthesize high-quality target singing voices. The performance of our model is primarily enhanced through three key components: 1) the style encoder that condenses style information into a compact latent space using a VQ model with ℓ_2 normalization; 2) the Style and Duration Language Model (S&D-LM), which predicts style information and phoneme duration information simultaneously, thus enhancing both; and 3) the style adaptive decoder that employs a novel style adaptive normalization method to generate enhanced details in singing voices. Our experimental results demonstrate that TransferSinger surpasses baseline models in both synthesis quality and singer similarity across various tasks: zero-shot SVS, controllable style synthesis, cross-lingual style transfer, and speech-to-singing style transfer.

6 Limitations

Our method primarily acknowledges two key limitations. First, our multilingual data currently only facilitates cross-lingual style transfer between Chinese and English, primarily due to the challenges in collecting singing voice data. In the future, we plan to gather more diverse language data for conducting multilingual style transfer experiments. Second, our model only allows for global control of singing styles, lacking the ability to finely customize the style techniques used for each phoneme. Looking ahead, our future work aims to control singing styles at the phoneme level for zero-shot SVS tasks.

7 Ethics Statement

TransferSinger, due to its ability to transfer personal timbre and styles for singing voice synthesis, may be used for dubbing in entertainment videos, leading to possible infringement of singers' copyrights. Meanwhile, due to its capacity for transferring cross-lingual speech and singing, our model could potentially result in unfair competition and unemployment for individuals in related singing occupations. Consequently, we will enforce restrictions on our model to mitigate unauthorized usage.

References

Bagus Tris Atmaja and Akira Sasou. 2022. Evaluating self-supervised speech representations for speech emotion recognition. *IEEE Access*, 10:124396–124407.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

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.

Yin-Ping Cho, Yu Tsao, Hsin-Min Wang, and Yi-Wen Liu. 2022. Mandarin singing voice synthesis with denoising diffusion probabilistic wasserstein gan.

Soonbeom Choi and Juhan Nam. 2022. A melody-unsupervision model for singing voice synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7242–7246. IEEE.

Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. 2020. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6184–6188. IEEE.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2016. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*.

Jinzheng He, Jinglin Liu, Zhenhui Ye, Rongjie Huang, Chenye Cui, Huadai Liu, and Zhou Zhao. 2023. Rmssinger: Realistic-music-score based singing voice synthesis. *arXiv preprint arXiv:2305.10686*.

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.

Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2021. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3945–3954.

Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2022a. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech synthesis. *arXiv preprint arXiv:2205.07211*.

Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022b. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605.

Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie Huang, Chunfeng Wang, Xiang Yin, et al. 2023. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *arXiv preprint arXiv:2306.03509*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Sungjae Kim, Yewon Kim, Jewoo Jun, and Injung Kim. 2023. Muse-svs: Multi-singer emotional singing voice synthesizer that controls emotional intensity. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

Neeraj Kumar, Srishti Goel, Ankur Narang, and Brejesh Lall. 2021. Normalization driven zero-shot multi-speaker speech synthesis. In *Interspeech*, pages 1354–1358.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. 2022a. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11020–11028.

Jinglin Liu, Chengxi Li, Yi Ren, Zhiying Zhu, and Zhou Zhao. 2022b. Learning the beauty in songs: Neural singing voice beautifier. *arXiv preprint arXiv:2202.13277*.

Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. 2021. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. In *International Conference on Machine Learning*, pages 7748–7759. PMLR.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. Aishell-3: A multi-speaker mandarin tts corpus and the baselines.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2021. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*.

Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al. 2022a. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems*, 35:6914–6926.

Yongmao Zhang, Jian Cong, Heyang Xue, Lei Xie, Pengcheng Zhu, and Mengxiao Bi. 2022b. Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7237–7241. IEEE.

Yu Zhang, Rongjie Huang, Ruiqi Li, JinZheng He, Yan Xia, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2023. Stylesinger: Style transfer for out-of-domain singing voice synthesis. *arXiv preprint arXiv:2312.10741*.

Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. 2022. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425.

Hyperparameter		TransferSinger
Phoneme Encoder	Phoneme Embedding	320
	Encoder Layers	5
	Encoder Hidden	320
	Kernal Size	9
	Filter Size	1280
Note Encoder	Pitches Embedding	320
	Type Embedding	320
	Duration Hidden	320
Timbre Encoder	Encoder Layers	5
	Hidden Size	320
	Conv1D Kernel	31
Style Encoder	WN Layers	4
	WN Kernel	3
	Conv Layers	5
	Conv Kernel	5
	Hidden Channel	320
	VQ Embedding Size	512
	VQ Embedding Channel	64
Pitch Predictor	Conv Layers	12
	Kernel Size	3
	Residual Channel	192
	Hidden Channel	25
	Time Steps	100
	Max Linear β Schedule	0.06
Style Adapt Decoder	Denoiser Layers	20
	Denoiser Hidden	320
	Time Steps	8
	Noise Schedule Type	VPSDE
S&D-LM	Decoder Layers	8
	Style Embedding Size	514
	Hidden Size	512
	Kernal Size	5
	Attention Heads	8
Total Number of Parameters		328.5M

Table 6: Hyper-parameters of TransferSinger modules.

A Details of Models

A.1 Architecture Details

We list the architecture and hyperparameters of our TransferSinger in Table 6.

A.2 Content Encoder

Our content encoder is composed of a note encoder and a phoneme encoder. The phoneme encoder processes a sequence of phonemes through a phoneme embedding layer and four FFT blocks, culminating in the production of phoneme features. On the other hand, the note encoder is responsible for handling musical score information. It processes note pitches, note types (including rest, slur, grace, etc.), and note duration. Each of these is processed through two embedding layers and a linear projection layer, thereby generating note features.

A.3 Timbre Encoder

Designed to encapsulate the singer's identity, the timbre encoder extracts a global vector t from the prompt audio. The encoder comprises several stacks of convolution layers. To maintain the stability of the timbre information, a one-dimensional timbre vector t is obtained by averaging the output of the timbre encoder over time.

A.4 Pitch Predictor

In our model, the pitch predictor employs a combination of Gaussian diffusion and multinomial diffusion methodologies to generate $F0$ and UV. This process is described mathematically as follows:

$$\begin{aligned} q(x_t|x_{t-1}) &= \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \\ q(y_t|y_{t-1}) &= \mathcal{C}(y_t|(1-\beta_t)y_{t-1} + \beta_t/K), \end{aligned} \tag{8}$$

where $\mathcal{C}$ denotes a categorical distribution with probability parameters, $x_t \sim \{0,1\}^K$, and β_t is the probability of uniformly resampling a category. In the reverse process, we train a neural network to approximate the noise ϵ from the noisy input x_t and $\hat{y_0}$ from the noisy sample y_t at timestep t. The equations of the reverse process are as follows:

$$\begin{aligned} &E_{x_0,\epsilon}[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}||\epsilon - \epsilon_\theta(x_t,t)||], \\ &q(y_{t-1}|y_t,y_0) = \mathcal{C}(y_{t-1}|\theta_{post}(y_t,y_0)), \\ &\theta_{post}(y_t,y_0) = \tilde{\theta}/\sum_{k=1}^{K}\tilde{\theta}_k, \\ &\tilde{\theta} = [\alpha_t y_t + (1-\alpha_t)/K] \odot \\ &[\bar{\alpha}_{t-1}y_0 + (1-\bar{\alpha}_{t-1})/K], \end{aligned} \tag{9}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$. We use $p(y_{t-1}|y_t) = \mathcal{C}(y_{t-1}|\theta_{post}(y_t, \hat{y0}))$ to approximate $q(y_{t-1}|y_t, y_0)$. Our pitch predictor employs a non-causal WaveNet architecture for the denoiser. The optimization of this module is achieved using Gaussian diffusion loss and multinomial diffusion loss.

B Details of Dataset

Currently, no datasets are annotated with style information, and most open-source singing datasets lack note annotations. In this endeavor, we collect and annotate a cross-lingual dataset (16 singers, 28h Chinese singing, 4h English singing) by recruiting professional singers in a professional recording studio. Each singer was compensated at an hourly rate of $600. Singers are informed that the data

Dataset	Total/h	Chinese		English	
		sing	speech	sing	speech
New	28	25	0	3	0
M4Singer	30	30	0	0	0
OpenSinger	85	85	0	0	0
AISHELL-3	85	0	85	0	0
BuTFy	18	0	0	8	10
Total/h	246	140	85	11	10

Table 7: Time distribution of our datasets for Chinese, English, speech, and singing data.

will be used for scientific research. Additionally, we incorporate the M4Singer dataset (Zhang et al., 2022a) (20 singers and 30h Chinese singing) to expand the diversity of singers and styles. Subsequently, we also add the OpenSinger dataset (Huang et al., 2021) (93 singers and 85h Chinese singing), the AISHELL-3 dataset (Shi et al., 2021) (218 singers and 85h Chinese speech), and a subset of the PopBuTFy database (Liu et al., 2022b) (20 singers, 10h English speech, and 8h English singing). We use these datasets under license CC BY-NC-SA 4.0. None of these three datasets has note annotations, so we have hired music experts to manually annotate the note information for these three datasets. Each annotator was compensated at an hourly rate of $20. Participants are informed that the data will be used for scientific research. The time distribution of our datasets for Chinese, English, speech, and singing data are listed in Table 7. With the assistance of music experts, we manually annotate singing data with distinct style class labels. We categorize songs into soprano, tenor, alto, and bass based on vocal ranges. In singing methods, we classify songs as bel canto and pop. Based on certain techniques, like songs that use a lot of falsetto or vibrato, we label them as 'falsetto' or 'vibrato'. These classifications are combined into the final style class labels (like alto pop vibrato), which will be the text prompts.

C Details of Evaluation

C.1 Subjective Evaluation

For each task, we randomly select 20 pairs of sentences from our test set for subjective evaluation. Each pair consists of a prompt audio that provides timbre and styles, and a synthesized singing voice, each of which is listened to by at least 15 professional listeners. In the context of MOS and CMOS evaluations, these listeners are instructed to concentrate on synthesis quality (including clarity, naturalness, and rich stylistic details), irrespective of singer similarity (in terms of timbre and styles). Conversely, during SMOS and CSMOS evaluations, the listeners are directed to assess singer similarity (singer similarity in terms of timbre and styles) to the prompt audio, disregarding any differences in content or synthesis quality (including quality, clarity, naturalness, and rich stylistic details). In both MOS and SMOS evaluations, listeners are requested to grade various singing voice samples on a Likert scale ranging from 1 to 5. For CMOS and CSMOS evaluations, listeners are guided to compare pairs of singing voice samples generated by different systems and express their preferences. The preference scale is as follows: 0 for no difference, 1 for a slight difference, and 2 for a significant difference. It is important to note that all participants are fairly compensated for their time and effort. We compensated participants at a rate of $12 per hour, resulting in a total expenditure of approximately $300 on participant compensation. Participants are informed that the data will be used for scientific research.

C.2 Objective Evaluation

To objectively evaluate the timbre similarity and synthesis quality of the test set, we employ two metrics: Cosine Similarity (Cos) and F0 Frame Error (FFE). Cosine Similarity is used to measure the resemblance in the singer's identity between the synthesized singing voice and the prompt audio. This is done by computing the average cosine similarity between the embeddings extracted from the synthesized voices and the prompt audio, thus providing an objective indication of the performance in singer similarity. Subsequently, we use FFE, which amalgamates metrics for voicing decision error and F0 error. FFE effectively captures essential synthesis quality information.