MM-TTS: Multi-Modal Prompt Based Style Transfer for Expressive Text-to-Speech Synthesis

Wenhao Guan[1], Yishuang Li[2], Tao Li[1], Hukai Huang[1], Feng Wang[1], Jiayan Lin[1], Lingyan Huang[1], Lin Li[2,3*], Qingyang Hong[1*]

[1] School of Informatics, Xiamen University, China
[2] Institute of Artificial Intelligence, Xiamen University, China
[3] School of Electronic Science and Engineering, Xiamen University, China
whguan@stu.xmu.edu.cn,{lilin,qyhong}@xmu.edu.cn

Abstract

The style transfer task in Text-to-Speech (TTS) refers to the process of transferring style information into text content to generate corresponding speech with a specific style. However, most existing style transfer approaches are either based on fixed emotional labels or reference speech clips, which cannot achieve flexible style transfer. Recently, some methods have adopted text descriptions to guide style transfer. In this paper, we propose a more flexible multi-modal and style controllable TTS framework named *MM-TTS*. It can utilize any modality as the prompt in unified multi-modal prompt space, including reference speech, emotional facial images, and text descriptions, to control the style of the generated speech in a system. The challenges of modeling such a multi-modal style controllable TTS mainly lie in two aspects: 1) aligning the multi-modal information into a unified style space to enable the input of arbitrary modality as the style prompt in a single system, and 2) efficiently transferring the unified style representation into the given text content, thereby empowering the ability to generate prompt style-related voice. To address these problems, we propose an aligned multi-modal prompt encoder that embeds different modalities into a unified style space, supporting style transfer for different modalities. Additionally, we present a new adaptive style transfer method named Style Adaptive Convolutions (SAConv) to achieve a better style representation. Furthermore, we design a Rectified Flow based Refiner to solve the problem of oversmoothing Mel-spectrogram and generate audio of higher fidelity. Since there is no public dataset for multi-modal TTS, we construct a dataset named *MEAD-TTS*, which is related to the field of expressive talking head. Our experiments on the MEAD-TTS dataset and out-of-domain datasets demonstrate that MM-TTS can achieve satisfactory results based on multi-modal prompts. The audio samples and constructed dataset are available at https://multimodal-tts.github.io.

Introduction

With the rapid advancement of deep learning, Text-to-Speech (TTS) has witnessed significant progress. The development of typical neural TTS systems presents diversification. To address the challenge of slow inference speed in autoregressive models (Shen et al. 2018; Li et al. 2019), non-autoregressive systems like FastSpeech (Ren et al. 2019) and FastSpeech2 (Ren et al. 2021) have emerged as viable solutions. In order to explore the impact of different generative models on TTS systems, a series of TTS models based on generative models, including Glow-TTS (Kim et al. 2020), Diff-TTS (Jeong et al. 2021), VITS (Kim, Kong, and Son 2021) and ProDiff (Huang et al. 2022b), have been developed. Moreover, to address the mismatch issue in the Mel-spectrogram domain within the two-stage system, several end-to-end TTS models (Kim, Kong, and Son 2021; Tan et al. 2022; Lei et al. 2023) have been proposed. Recently, novel approaches utilizing discrete tokens (Du et al. 2023; Wang et al. 2023) for modeling text and speech sequences have emerged, leading to more natural and high-fidelity speech synthesis. The applications of these technologies have progressively bridged the gap between the generation result of typical TTS system and human-level speech.

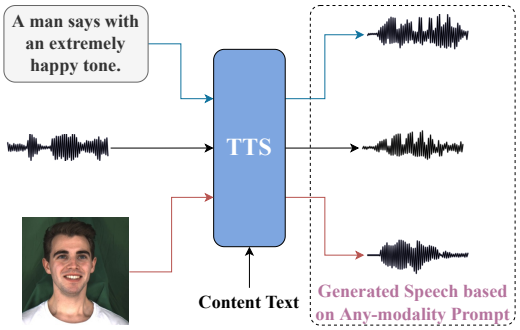
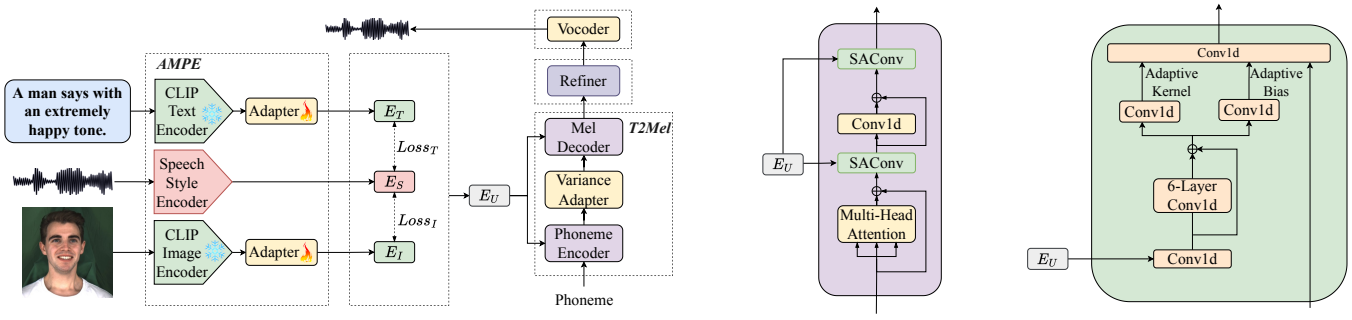


Figure 1: The general style transfer framework for TTS with multi-modal prompts.

The essence of TTS is a cross-modal task that maps text content to corresponding speech. However, simply generating speech corresponding to the text content is not enough. We also hope that the generated speech has rich paralinguistic information such as speaker identity, language, emotion and prosody. A lot of expansion works are proposed to achieve specific purposes. For example, multi-speaker TTS (Casanova et al. 2021; Chen et al. 2020), multilingual TTS (Li et al. 2021a), style transfer (Huang et al. 2022a; Guan et al. 2023a). We divide the style transfer tasks for TTS into four main categories: emotion label based style transfer, reference speech based style transfer, face based

*Corresponding author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Overall Model Archietecture *(b) Archietecture of Phoneme Encoder and Mel Decoder* *(c) SAConv Module*

Figure 2: The model architecture of MM-TTS.

style transfer and text description based style transfer. Emotion label based style transfer (Lee, Rabiee, and Lee 2017; Lorenzo-Trueba et al. 2018) uses several predefined kinds of emotional category labels to represent the style. Reference speech based style transfer (Min et al. 2021; Huang et al. 2022a; Guan et al. 2023a) aims at extracting the style of given speech segment and transferring the style to generate style-related speech for any text content. Face based Style Transfer (Goto et al. 2020; Wang et al. 2022) mainly focuses on extracting the speaker identity of a given face image and then it can be applied in multi-speaker TTS. Text description based style transfer (Guo et al. 2023; Yang et al. 2023) utilizes natural language prompts to generate specific styles of speech. Although these systems can achieve good results, they are not flexible enough. In this paper, we propose a more flexible style transfer framework, which is not constrained by a single modality prompt and can input any modality as the prompt during inference stage.

To achieve more flexible style controllability, we propose a general TTS framework with multi-modal prompts as shown in Figure 1. We can input any modality into this system and subsequently provide any textual input as the content text. The system is capable of generating speech that is stylistically related to the corresponding modality. Specifically, we propose a novel TTS method, *MM-TTS*. Figure 2 illustrates the architecture of MM-TTS for controllable style transfer with multi-modal prompts. To achieve alignment among multi-modal prompt features, we introduce an Aligned Multi-modal Prompt Encoder (AMPE) based on the semantic understanding capability of Contrastive Language-Image Pretraining (CLIP) (Radford et al. 2021) to unify the text, speech and image prompt modality into the unified style space, thus supporting flexible style control via multi-modal inputs. For our text-to-mel model, we propose a Style Adaptive Convolutions (SAConv) module to extract more local details of style information, which also facilitates the style transfer ability. To overcome the over-smoothness problem produced by the text-to-mel model (Ren et al. 2022), we design a novel module called Reflow Refiner, which is based on the Rectified Flow (Liu, Gong, and Liu 2022) to obtain the Mel-spectrogram closer to the real domain. Finally the generated Mel-spectrogram is converted to waveform by a pretrained vocoder.

However, there is no public dataset for the multi-modal TTS, in this paper, we construct a dataset that consists of text, face and speech prompts for expressive and controllable TTS based on the emotional talking head video dataset MEAD (Wang et al. 2020). We call the newly constructed multi-modal expressive TTS dataset as MEAD-TTS.

In summary, the main contributions of this work are as follows:

- We propose a novel method MM-TTS to endow the TTS with multi-modal capabilities. To our knowledge, this is the first multi-modal prompt based style transfer framework for TTS. Abundant experiments are conducted to demonstrate the superiority of our method on objective and subjective evaluations.
- To unify the multi-modal prompt style space, we propose a novel AMPE, allowing flexible style control guided by prompts of different modalities for TTS. Additionally, we introduce a novel module, SAConv, to transfer the style information extracted by the AMPE into any content text to generate style-related speech more effectively.
- A novel Rectified Flow based Refiner is designed to refine the previously generated Mel-spectrograms, thus resolve the over-smoothness problem and ultimately get Mel-spectrograms closer to the real data domain.
- We construct a dataset MEAD-TTS for multi-modal TTS, providing the reference dataset for future works.

Related Work

Reference Speech Based Style Transfer

Reference speech based style transfer is the most popular method for style transfer in TTS, because the speech clips are easy to obtain and contain rich information. Global style token (GST) (Wang et al. 2018) designs a style token layer and a reference encoder to explore the expressiveness of TTS systems unsupervisedly. VAE-Tacotron (Zhang et al. 2019) learns the style representation through VAE (Kingma and Welling 2013). Subsequently, significant efforts have focused on designing robust style encoders for reference speech based style transfer tasks. Some works, based on autoregressive models, have achieved fine-grained style representations through the design of multi-level or multi-scale

style modeling methods (Sun et al. 2020; Li et al. 2021b). However, to overcome the issue of low decoding speed associated with autoregressive models, non-autoregressive architectures have been adopted in some methods. Meta-StyleSpeech (Min et al. 2021) adopts the base architecture upon FastSpeech2 (Ren et al. 2021), applying style adaptive layer norm and meta-learning algorithm to effectively synthesize style-transferred speech. GenerSpeech (Huang et al. 2022a) proposes a multi-level style adaptor and a generalizable content adaptor to efficiently model the style information. IST-TTS (Guan et al. 2023a) designs a novel TTS system that can perform style transfer with interpretability and high fidelity based on controllable variational autoencoder (Shao et al. 2020) and diffusion models. In this work, we propose a more effective SAConv module to transfer extracted style information and thus we obtain the highly style-related speech.

Face Based Style Transfer

We can imagine speakers' voice characteristics from their faces. Due to the consideration of this phenomenon, there are some works that utilize facial features to generate speech that matches the speaker's characteristics. Face2Speech (Goto et al. 2020) uses a face image to control the voice characteristics of the synthesized speech by training a face encoder and a speaker encoder separately and finally designs a loss function to make their embeddings closer. In FR-PSS (Wang et al. 2022), based on Face2Speech, a residual-guided strategy is designed by incorporating a prior speech feature to make the network capture representative face features. The works mentioned above train the face encoder to share a joint embedding space with the speech encoder, independently from the main TTS model. FaceTTS (Lee, Chung, and Chung 2023) designs a multi-speaker TTS model by training a robust cross-modal representation of speaking style, where speaking styles are conditioned on face attributes. But existing works usually focus on extracting the speaker identity of a given face image for multi-speaker TTS, which is limited in style-rich TTS systems. In this paper, we focus on modeling the style attributes of facial images such as gender and emotion.

Text Description Based Style Transfer

Due to the emergence of large language models and text based image generation, using a text description as prompt to guide the generation of texts or images has drawn wide attention recently. In the field of TTS, several recent works (Kim et al. 2021; Guo et al. 2023; Yang et al. 2023; Liu et al. 2023) have emerged to utilize text descriptions to guide speech synthesis. PromptTTS (Guo et al. 2023) designs a style encoder, which maps a style prompt to a semantic space to extract the style representation, to guide the content encoder and the speech decoder. InstructTTS (Yang et al. 2023) introduces a three-stage training strategy to obtain a robust embedding model, which can effectively capture semantic information from the style prompts. It proposes to model acoustic features in discrete latent space and utilizes discrete diffusion model to generate discrete acoustic features. PromptStyle (Liu et al. 2023) proposes a two-stage TTS approach for cross-speaker style transfer with natural language descriptions based on VITS (Kim, Kong, and Son 2021). In this paper, we utilize the strong semantic capability of CLIP to obtain the style representations of text descriptions.

MM-TTS

Overview

MM-TTS is considered a multi-modal prompt based style transfer framework for expressive TTS. In order to achieve highly flexible multi-modal style transfer for TTS, we design several modules to achieve the unified style space and the efficient adaptive style transfer respectively.

As shown in Figure 2, the proposed framework consists of four modules: 1) an aligned multi-modal prompt encoder (AMPE), which includes a CLIP based text encoder, a CLIP based image encoder, a Speech style encoder and two Adapter modules, to align the multi-modal information into a unified style space; 2) a Text-to-Mel model (T2Mel) that maps the given text content to corresponding speech, where we introduce Style Adaptive Convolutions (SAConv) to efficiently transfer the unified style representation to generate the style related speech. We adopt the StyleSpeech proposed in (Min et al. 2021), which is based on FastSpeech2 (Ren et al. 2021), as the T2Mel backbone; 3) a Refiner to refine the Mel-spectrograms and get more realistic Mel-spectrograms. 4) a vocoder to convert the Mel-spectrogram to waveform.

Aligned Multi-Modal Prompt Encoder

To get a unified multi-modal prompt based style space, we introduce an Aligned Multi-modal Prompt Encoder (AMPE) based on CLIP to unify the text, speech and image prompt modality into the unified style space, thus supporting flexible style control via multi-modal inputs.

In practice, in the training phase, we first input prompts of three modalities (text, speech, face) into the AMPE module. Specifically, the fixed CLIP based text and image encoders followed by learned adapter modules are leveraged to extract corresponding text prompt embedding $\boldsymbol{E}_T$ and face image prompt embedding $\boldsymbol{E}_I$, and the speech style encoder is learned to extract the speech prompt embedding $\boldsymbol{E}_S$. The unified style embedding $\boldsymbol{E}_U$ is guided by several MSE loss functions, which is computed as follows:

$$\begin{aligned} L_{AMPE} &= Loss_I + Loss_T \\ &= MSE(\boldsymbol{E}_I, \boldsymbol{E}_S) + MSE(\boldsymbol{E}_T, \boldsymbol{E}_S), \end{aligned} \tag{1}$$

The unified style embedding $\boldsymbol{E}_U$ is different at training phase and inference phase:

$$\begin{aligned} &\boldsymbol{E}_U = \boldsymbol{E}_S, \text{at the training phase} \\ &\boldsymbol{E}_U \in \{\boldsymbol{E}_S, \boldsymbol{E}_I, \boldsymbol{E}_T\}, \text{at the inference phase} \end{aligned} \tag{2}$$

In the inference phase, we can employ arbitrary embedding of different modalities as the $\boldsymbol{E}_U$.

For the speech style encoder in the AMPE module, we design the speech style encoder including four components. The initial three components, namely Spectral Processing,

Temporal Processing, and Multi-head Attention, are analogous to those employed in Meta-StyleSpeech (Min et al. 2021). Additionally, we introduce a fourth component, a Multi-layer GRU (Chung et al. 2014), to capture richer style information. Finally, we get an informative multi-channel vector. The adapter modules in AMPE are simply comprised of two fully connected layers.

Style Adaptive Convolutions

In previous works (Casanova et al. 2021; Guan et al. 2023a), the extracted style information is usually directly fed into the generator through concatenation or summation. Meta-StyleSpeech (Min et al. 2021) proposes Style Adaptive Layer Normalization (SALN) to transfer the statistical properties of reference style features to a given content text, and thus gets style transferred speech. Inspired from Adaptive Convolutions (Chandran et al. 2021) for image style transfer, we propose Style Adaptive Convolutions (SAConv) to transfer the reference style features more precisely.

For the architecture, as illustrated in Figure 2 (c), the SAConv receives the style embedding $\boldsymbol{E}_U$ and predicts the kernel and bias via kernel prediction networks, then the predicted kernel and bias are used for context feature to get prompt style transferred speech. Specifically, the kernel prediction network comprises of an input convolution, a residual convolution module with 6 convolution layers, a kernel convolution and a bias convolution to predict adaptive kernel and bias respectively.

Given a context feature $\boldsymbol{x}$ and the desired style embedding $\boldsymbol{E}_U$, the normalized context feature $\boldsymbol{x}_{norm}$ is derived as follows:

$$\boldsymbol{x}_{norm} = \frac{\boldsymbol{x} - \mu_x}{\sigma_x} \tag{3}$$

where μ_x and σ_x are the mean and standard deviation of the input context feature $\boldsymbol{x}$.

Then, we obtain the predicted kernel $\boldsymbol{conv}_{kernel}$ and bias $\boldsymbol{conv}_{bias}$ of the given style feature $\boldsymbol{E}_U$ by the kernel prediction network. And then the SAConv is computed utilizing predicted kernel $\boldsymbol{conv}_{kernel}$ and bias $\boldsymbol{conv}_{bias}$ as follows:

$$\begin{aligned} &SAConv(\boldsymbol{x}, \boldsymbol{conv}_{kernel}, \boldsymbol{conv}_{bias}) \\ &= \boldsymbol{conv}_{kernel}(\boldsymbol{x}_{norm}) + \boldsymbol{conv}_{bias} \end{aligned} \tag{4}$$

While the gain and bias of SALN are fully connected layers, which only transfer global statistics of style features, our SAConv aims at predicting convolution kernels and biases according to the given style embedding, to transfer the style features more precisely. The detailed model information is in Appendix C.

Rectified Flow Based Refiner

In order to overcome the over-smoothing problem and maintain the effectiveness of our SAConv module to the greatest extent, we propose a novel rectified flow (Liu, Gong, and Liu 2022) based refiner (Reflow Refiner) through two-stage TTS training pipeline like the way in IST-TTS (Guan et al. 2023a), which can yield high-quality results for Mel-spectrogram generation when simulated with a very few number of Euler steps. The Reflow Refiner is an Ordinary Differential Equation (ODE) model that transports distribution π_0 to π_1 by straight line paths as much as possible, where π_0 is the standard gaussian distribution and π_1 is the ground truth distribution. Given empirical observations of $X_0 \sim \pi_0, X_1 \sim \pi_1$, the Reflow Refiner induced from (X_0, X_1) is an ODE conditioned on generated Mel-spectrogram c_{mel} in the first training stage with respect to time $t \in [0, 1]$,

$$dX_t = v(X_t, t, c_{mel})dt, \tag{5}$$

which converts X_0 from π_0 to a X_1 from π_1. The drift force $v : \mathbb{R}^d \to \mathbb{R}^d$ is set to drive the flow to follow the direction $(X_1 - X_0)$ of the linear path pointing from X_0 to X_1 as much as possible, by solving a simple least squares regression problem, the loss $L_{Refiner}$ is as follows:

$$L_{Reflow} = \int_0^1 \mathbb{E}[\| (X_1 - X_0) - v(X_t, t, c_{mel}) \|^2]dt, \tag{6}$$

where $X_t = tX_1 + (1-t)X_0$, X_t is the linear interpolation of X_0 and X_1.

During inference phase, we get v and we solve the ODE starting from $X_0 \sim \pi_0$ to transfer π_0 to π_1 conditioned on c_{mel}. Our MM-TTS utilizes the ODE RK45 sampler for inference.

With the training of the Reflow Refiner, it can make the paths between π_0 and π_1 tending to straight lines. As we expected, perfectly straight paths can be simulated exactly with a single Euler step, which is a one-step model with high quality and fast inference speed. We put more detailed information about the Reflow Refiner module in Appendix C.

Objectives

Our training pipeline has two stages: text-to-mel stage and refiner stage. Combining the above AMPE, SAConv and text-to-mel model, we optimize the model in the first stage by minimizing the loss function as:

$$L_{1s} = L_{Mel} + L_{Var} + L_{AMPE}, \tag{7}$$

where L_{Mel} denotes the MAE loss between the ground-truth and Mel-spectrogram that generated by the Mel Decoder in MM-TTS. L_{Var} denotes the duration, pitch and energy reconstruction loss.

In the second stage for Reflow Refiner training, we optimize the Reflow Refiner model by minimizing the loss function as:

$$L_{2s} = L_{Reflow} \tag{8}$$

MEAD-TTS Dataset

To the best of our knowledge, there is no public expressive dataset for TTS with multi-modal prompts, so we construct a dataset to endow TTS with multi-modal prompt capability based on MEAD (Wang et al. 2020), which is originally a dataset of expressive talking head generation. To get transcriptions of corresponding speech clips, we utilize the whisper tool (Radford et al. 2023). Specifically, we designate the language as English, and using template matching method to match transcription texts given in MEAD. To get face images corresponding to the given speech clip, we randomly

select two video frames to represent the corresponding face images. Moreover, we design a text prompt template to construct text description prompts corresponding to different speech clips. The text prompt template has three main style variables: gender, emotion and emotion level, where the emotion has eight types (neutral, angry, contempt, disgusted, fear, happy, sad and surprised). So the style of this work refers to the summation of the three style variables (gender, emotion, emotion level). We get different text prompt descriptions by extracting the three variables from different data paths, such as "A <gender> says with a <emotion level> <emotion> tone". Finally, we utilize the remarkable generative capability of current LLMs (i.e. GPT-3.5-TURBO) to generate the text descriptions having the same meaning as previous templated prompts. When constructing the paired dataset, we center around a certain speech clip, randomly select a corresponding face image and the corresponding generated text description by LLMs as a data pair.

Through processing on MEAD dataset, the MEAD-TTS dataset consists of 31055 pairs of (speech, face image, text description) data with a total duration of approximately 36 hours for speech modality. In practice, the speech clips are resampled to 16kHz. The more details of the MEAD-TTS dataset is in Appendix A.

Experimental Setup

Datasets

Because the text transcriptions of MEAD-TTS dataset are limited, we first pretrain the model on LJSpeech (Ito and Johnson 2017) to increase the number of text transcription data and strengthen the output diversity of TTS, where the model only receives the speech modality as prompt when training on LJSpeech.

In order to comprehensively evaluate the effectiveness of the proposed method, we conduct experiments in intra-domain and out-of-domain datasets respectively. We utilize the MEAD-TTS dataset for intra-domain evaluation. And the LibriTTS (Zen et al. 2019) transcriptions and speech clips are utilized for out-of-domain evaluation of reference speech based style transfer. For face based style transfer, we use the face images in Oulu-CASIA (Zhao et al. 2011) dataset and transcriptions in LibriTTS for out-of-domain evaluation. For text description based style transfer, we use LibriTTS transcriptions for out-of-domain evaluations. The more details of the used datasets are in Appendix B.

Evaluation Metrics

We evaluate the style transfer quality and similarity by objective metrics and subjective evaluations. The objective metrics include speaker embedding cosine similarity (SECS) and mel cepstral distortion (MCD) metric. The SECS calculates the cosine similarity between the embeddings of two audios extracted from the speaker encoder. It ranges from -1 to 1, and a larger value indicates a stronger similarity. The MCD evaluates the spectral distance between the reference and synthesized Mel-spectrum features. As for subjective evaluations, we conduct 5-scale Mean Opinion Score (MOS) and Similarity Mean Opinion Score (SMOS) test between MM-TTS and the baseline models. The score of MOS test ranges from 1 to 5 with an interval of 0.5, in which 1 means very bad and 5 means excellent. Both MOS and SMOS are reported with 95% confidence interval. We generate 50 speech samples for each model, which are listened by 20 listeners for subjective evaluations. We additionally evaluate the emotion classification accuracy and gender classification accuracy for face prompt and text prompt based style transfer.

Note that in our experiments, the SECS scores are calculated using the speaker encoder in Resemblyzer[1]. For emotion and gender classification, we utilize a pretrained wav2vec 2.0 (Baevski et al. 2020) model as feature extractor, and then add two fully connected layers and one softmax layer for training.

Training Setting and Baseline Model

The proposed MM-TTS was trained for 200K iterations using Adam optimizer (Kingma and Ba 2014) on a single NVIDIA GeForce RTX 2080Ti GPU for both the first text-to-mel stage and second refiner stage training pipeline. Additionally, we utilize a pretrained HiFi-GAN (Kong, Kim, and Bae 2020) as the neural vocoder to convert generated Mel-spectrogram to waveform.

We compare our system with other methods from three aspects: reference speech based style transfer, face based style transfer and text description based style transfer.

As there is no related works for multi-modal TTS, we utilize StyleSpeech proposed in (Min et al. 2021) endowed with multi-modal prompts, which is named as MM-StyleSpeech, as our baseline system. MM-StyleSpeech uses the similar AMPE module as MM-TTS and other modules are the same as original Stylespeech.

For reference speech based style transfer, we conduct experiments on the following systems : 1) GT: This is the ground-truth recording; 2) GT (Mel + Voc): This is the speech synthesized using pretrained HiFi-GAN vocoder for GT Mel-spectrogram; 3) MS-FastSpeech2: This is a multi-speaker FastSpeech2 (Ren et al. 2021) system with x-vector speaker embedding to the encoder output and the decoder input; 4) GenerSpeech (Huang et al. 2022a): This is a style transfer model for TTS with a multi-level style adapter and a generalizable content adapter; 5) IST-TTS (Guan et al. 2023a): This is a style transfer model for TTS with controllable VAE and diffusion models; 6) StyleSpeech (Min et al. 2021): This is a style transfer model for TTS proposed in (Min et al. 2021) with Style Adaptive Layer Normalization; 7) MM-StyleSpeech: This is our proposed baseline model for multi-modal TTS based on StyleSpeech; 8) MM-TTS: This is our proposed model for multi-modal TTS.

Previous face based style transfer methods focus on modelling the speaker identity of face images and previous text description based style transfer methods construct the models based on different self-built datasets, so we only compare with the baseline MM-StyleSpeech.

[1]https://github.com/resemble-ai/Resemblyzer.

Method	Intra-domain				Out-of-domain			
	MOS	SMOS	SECS	MCD	MOS	SMOS	SECS	MCD
GT	4.51 ± 0.04	-	-	-	4.72 ± 0.05	-	-	-
GT (Mel+Voc)	4.49 ± 0.08	4.81 ± 0.06	0.989	1.02	4.69 ± 0.08	4.79 ± 0.07	0.935	1.05
MS-FastSpeech2	3.91 ± 0.05	3.96 ± 0.05	0.924	3.90	3.43 ± 0.04	3.34 ± 0.05	0.682	7.74
GenerSpeech	4.16 ± 0.04	4.41 ± 0.04	0.904	3.76	3.87 ± 0.08	3.89 ± 0.04	**0.732**	7.25
IST-TTS	4.19 ± 0.06	4.52 ± 0.04	0.927	3.72	3.81 ± 0.08	3.83 ± 0.04	0.696	7.31
StyleSpeech	4.09 ± 0.03	4.51 ± 0.03	0.945	3.26	3.69 ± 0.05	3.84 ± 0.03	0.716	6.91
MM-StyleSpeech (ours)	4.11 ± 0.07	4.49 ± 0.06	0.947	3.24	3.72 ± 0.09	3.82 ± 0.07	0.713	6.93
MM-TTS (ours)	$\mathbf{4.36 \pm 0.06}$	$\mathbf{4.61 \pm 0.07}$	**0.956**	**3.17**	$\mathbf{3.95 \pm 0.08}$	$\mathbf{4.03 \pm 0.08}$	0.728	**6.69**

Table 1: The performance comparison of parallel reference speech based style transfer.

Baseline	Intra-domain				Out-of-domain			
	7-point score	Perference			7-point score	Perference		
		C	E	O		C	E	O
MS-FastSpeech2	1.81 ± 0.11	16%	28%	56%	1.62 ± 0.09	7%	20%	73%
GenerSpeech	1.15 ± 0.12	28%	30%	42%	1.29 ± 0.05	33%	21%	46%
IST-TTS	1.12 ± 0.09	30%	25%	45%	1.27 ± 0.06	36%	18%	46%
StyleSpeech	1.27 ± 0.07	27%	26%	47%	1.31 ± 0.10	30%	20%	50%
MM-StyleSpeech	1.01 ± 0.08	33%	27%	40%	1.12 ± 0.07	26%	35%	39%

Table 2: The AXY preference test results for non-parallel reference speech based style transfer. C, E, O denote the preference rate for compared model, equivalent and our model respectively.

Experiment Results and Analysis

For different modality prompts, we conduct subjective and objective evaluations respectively.

Results on Reference Speech Based Style Transfer

For reference speech based style transfer, we classify the experiments into two categories according to the content consistency between the reference and generated speech clips: Parallel Style Transfer (PST) and Non-Parallel Style Transfer (NPST). Table 1 shows the subjective and objective results of PST including MOS, SMOS, SECS and MCD. Given the reference speech, our MM-TTS achieves better results on both audio naturalness and style similarity on MEAD-TTS dataset and out-of-domain datasets. Our method achieves the highest SECS value and the lowest MCD value except the SECS value in out-of-domain dataset. The results indicate that our method is more effective for extracting and transferring the style of reference speech.

For NPST, we select 100 samples from MEAD-TTS and LibriTTS testing sets for intra-domain and out-of-domain evaluation respectively. An AXY test used in (Huang et al. 2022a) is conducted, the range of 7-point score is from -3 to 3, 0 represents "Both are about the same distance". As shown in Table 2, the results indicate that listeners prefer MM-TTS synthesis against the compared models. The proposed SAConv significantly improves the style extraction ability, allowing an arbitrary reference sample to guide the stylistic synthesis of arbitrary content text. We put the visualization results about PST and NPST in Appendix D.1.

Results on Face Based Style Transfer

Table 3 presents the experimental results of face-based style transfer, encompassing MOS and classification accuracy for emotion and gender. In the context of face prompts, our MM-TTS model surpasses the baseline model in terms of audio naturalness and classification performance on the MEAD-TTS dataset as well as out-of-domain datasets. These results highlight the effectiveness of our method in accurately capturing and extracting style attributes from facial images, indicating its good capability in incorporating facial styles into synthesized speech.

Results on Text Description Based Style Transfer

Table 4 displays the experimental results of text description based style transfer. By utilizing the text prompts, our MM-TTS achieves superior scores in terms of audio naturalness and classification accuracy for emotion and gender. The results indicate that our method has good ability in modeling and incorporating the style of natural language descriptions.

Ablation Studies

The MOS and SMOS results of ablation studies are illustrated in Table 5, where "w/o SAConv" denotes substituting the SAConv module with SALN module in StyleSpeech, "w/o Reflow Refiner" denotes removing the second refiner training stage and "w/o Pretrain" denotes only using MEAD-TTS dataset for training without LJSpeech pretraining. It demonstrates that 1) substituting the SAConv module with SALN module results in a significant drop on style similarity, which demonstrates that SAConv can extract more

Method	Intra-domain			Out-of-domain		
	MOS	Acc_{emo}	Acc_{gen}	MOS	Acc_{emo}	Acc_{gen}
GT	4.51 ± 0.04	0.826	1.0	-	-	-
GT (Mel+Voc)	4.49 ± 0.08	0.813	1.0	-	-	-
MM-StyleSpeech	3.98 ± 0.05	0.651	0.996	3.56 ± 0.07	0.249	0.686
MM-TTS	$\mathbf{4.11 \pm 0.06}$	**0.659**	**0.997**	$\mathbf{3.77 \pm 0.08}$	**0.261**	**0.833**

Table 3: The performance comparison of face based style transfer.

Method	Intra-domain			Out-of-domain		
	MOS	Acc_{emo}	Acc_{gen}	MOS	Acc_{emo}	Acc_{gen}
GT	4.51 ± 0.04	0.826	1.0	-	-	-
GT (Mel+Voc)	4.49 ± 0.08	0.813	1.0	-	-	-
MM-StyleSpeech	3.99 ± 0.04	0.635	1.0	3.73 ± 0.08	0.289	0.989
MM-TTS	$\mathbf{4.19 \pm 0.07}$	**0.636**	**1.0**	$\mathbf{3.93 \pm 0.08}$	**0.318**	**0.996**

Table 4: The performance comparison of text description based style transfer.

Method	Intra-domain		Out-of-domain	
	MOS	SMOS	MOS	SMOS
MM-TTS	**4.36**	**4.61**	**3.95**	**4.03**
w/o SAConv	4.22	4.55	3.85	3.81
w/o Reflow Refiner	4.17	4.57	3.79	3.99
w/o Pretrain	4.32	4.59	3.61	3.63

Table 5: The performance of ablation studies in MM-TTS.

Method	Intra-domain		Out-of-domain	
	MOS	SMOS	MOS	SMOS
DDPM (1000 steps)	4.35	4.61	**3.96**	4.03
DDPM (100 steps)	4.31	4.59	3.89	4.02
Reflow (1 step)	4.32	4.56	3.91	4.01
Reflow (MM-TTS)	**4.36**	**4.61**	3.95	**4.03**

Table 6: The performance of utilizing different refiners for the proposed MM-TTS in speech based style transfer task.

effective style representations than SALN; 2) removing the Reflow Refiner leads to a decline in audio naturalness, which indicates that the Reflow Refiner mainly contributes to the fidelity maintaining; 3) without the pretraining on LJSpeech leads to a significant drop on naturalness and similarity especially on out-of-domain evaluations. We put more visualization results about ablation studies in Appendix D.2.

Refiner Evaluation

We conduct MOS and SMOS evaluations for different refiners used in our method. Table 6 shows the experimental results using different refiners in MM-TTS including DDPM (Ho, Jain, and Abbeel 2020) and Reflow. According to the results, the Reflow refiner used in MM-TTS achieves better scores in almost every aspect, and the Reflow (1 Euler step) achieves similar results as Reflow (RK45 sampler). This indicates that our proposed Reflow refiner is effective in both quality and sampling speed than DDPM based refiners.

Limitations

While our approach successfully enables flexible and controllable style transfer using multi-modal prompts, it is important to acknowledge the limitations of our method. Firstly, the dataset we constructed is relatively small in scale, which restricts its applicability to real-world scenarios. This limitation is evident in the performance comparison on out-of-domain datasets, as illustrated in Table 3 and 4. Secondly, the text description prompt template we devised has certain limitations, particularly in comprehending more abstract and complex styles. This restricts the system's ability to effectively capture and transfer such nuanced stylistic attributes. Despite these limitations, we believe that our MM-TTS has taken an important step towards a more universal TTS system with multi-modal capabilities.

Conclusion

In this work, we propose a general multi-modal prompt based style transfer framework for TTS named MM-TTS. Specifically, we develop an aligned multi-modal prompt encoder based on CLIP to unify the different modality prompts into the unified style space, empowering the flexible style control by inputting any modality as the prompt during inference stage. In order to transfer the prompt style information to given text content more effectively, we propose SAConv to pay attention to model more local details of style information. Furthermore, we propose a Rectified Flow based refiner to obtain Mel-spectrograms closer to the real domain. We also construct a multi-modal TTS dataset MEAD-TTS to facilitate future related studies. Subjective and objective experiments demonstrate the superiority of our method. The details of technical appendix is in (Guan et al. 2023b).

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62276220 and 62001405.

References

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

Casanova, E.; Shulby, C.; Gölge, E.; Müller, N. M.; de Oliveira, F. S.; Junior, A. C.; Soares, A. d. S.; Aluisio, S. M.; and Ponti, M. A. 2021. Sc-glowtts: an efficient zero-shot multi-speaker text-to-speech model. *arXiv preprint arXiv:2104.05557*.

Chandran, P.; Zoss, G.; Gotardo, P.; Gross, M.; and Bradley, D. 2021. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7972–7981.

Chen, M.; Tan, X.; Ren, Y.; Xu, J.; Sun, H.; Zhao, S.; and Qin, T. 2020. MultiSpeech: Multi-Speaker Text to Speech with Transformer. In *Proc. Interspeech 2020*, 4024–4028.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Du, C.; Guo, Y.; Shen, F.; Liu, Z.; Liang, Z.; Chen, X.; Wang, S.; Zhang, H.; and Yu, K. 2023. UniCATS: A Unified Context-Aware Text-to-Speech Framework with Contextual VQ-Diffusion and Vocoding. *arXiv preprint arXiv:2306.07547*.

Goto, S.; Onishi, K.; Saito, Y.; Tachibana, K.; and Mori, K. 2020. Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image. In *INTERSPEECH*, 1321–1325.

Guan, W.; Li, T.; Li, Y.; Huang, H.; Hong, Q.; and Li, L. 2023a. Interpretable Style Transfer for Text-to-Speech with ControlVAE and Diffusion Bridge. In *Proc. INTERSPEECH 2023*, 4304–4308.

Guan, W.; Li, Y.; Li, T.; Huang, H.; Wang, F.; Lin, J.; Huang, L.; Li, L.; and Hong, Q. 2023b. MM-TTS: Multi-Modal Prompt Based Style Transfer for Expressive Text-to-Speech Synthesis. *arXiv preprint arXiv:2312.10687*.

Guo, Z.; Leng, Y.; Wu, Y.; Zhao, S.; and Tan, X. 2023. PromptTTS: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Huang, R.; Ren, Y.; Liu, J.; Cui, C.; and Zhao, Z. 2022a. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. *Advances in Neural Information Processing Systems*, 35: 10970–10983.

Huang, R.; Zhao, Z.; Liu, H.; Liu, J.; Cui, C.; and Ren, Y. 2022b. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2595–2605.

Ito, K.; and Johnson, L. 2017. The lj speech dataset. 2017. *URL https://keithito. com/LJ-Speech-Dataset*.

Jeong, M.; Kim, H.; Cheon, S. J.; Choi, B. J.; and Kim, N. S. 2021. Diff-tts: A denoising diffusion model for text-to-speech. *proceedings of INTERSPEECH*.

Kim, J.; Kim, S.; Kong, J.; and Yoon, S. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33: 8067–8077.

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

Kim, M.; Cheon, S. J.; Choi, B. J.; Kim, J. J.; and Kim, N. S. 2021. Expressive Text-to-Speech Using Style Tag. In *Proc. Interspeech 2021*, 4663–4667.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Lee, J.; Chung, J. S.; and Chung, S.-W. 2023. Imaginary Voice: Face-Styled Diffusion Model for Text-to-Speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Lee, Y.; Rabiee, A.; and Lee, S.-Y. 2017. Emotional end-to-end neural speech synthesizer. *arXiv preprint arXiv:1711.05447*.

Lei, Y.; Yang, S.; Wang, X.; Xie, Q.; Yao, J.; Xie, L.; and Su, D. 2023. UniSyn: An End-to-End Unified Model for Text-to-Speech and Singing Voice Synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13025–13033.

Li, N.; Liu, S.; Liu, Y.; Zhao, S.; and Liu, M. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6706–6713.

Li, S.; Ouyang, B.; Li, L.; and Hong, Q. 2021a. Lightts: Lightweight multi-speaker multi-lingual text-to-speech. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8383–8387. IEEE.

Li, X.; Song, C.; Li, J.; Wu, Z.; Jia, J.; and Meng, H. 2021b. Towards Multi-Scale Style Control for Expressive Speech Synthesis. In *Proc. Interspeech 2021*, 4673–4677.

Liu, G.; Zhang, Y.; Lei, Y.; Chen, Y.; Wang, R.; Xie, L.; and Li, Z. 2023. PromptStyle: Controllable Style Transfer

for Text-to-Speech with Natural Language Descriptions. In *Proc. INTERSPEECH 2023*, 4888–4892.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Lorenzo-Trueba, J.; Henter, G. E.; Takaki, S.; Yamagishi, J.; Morino, Y.; and Ochiai, Y. 2018. Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis. *Speech Communication*, 99: 135–143.

Min, D.; Lee, D. B.; Yang, E.; and Hwang, S. J. 2021. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. In *International Conference on Machine Learning*, 7748–7759. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518. PMLR.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2021. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019. FastSpeech: fast, robust and controllable text to speech. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 3171–3180.

Ren, Y.; Tan, X.; Qin, T.; Zhao, Z.; and Liu, T.-Y. 2022. Revisiting Over-Smoothness in Text to Speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8197–8213.

Shao, H.; Yao, S.; Sun, D.; Zhang, A.; Liu, S.; Liu, D.; Wang, J.; and Abdelzaher, T. 2020. Controlvae: Controllable variational autoencoder. In *International Conference on Machine Learning*, 8655–8664. PMLR.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4779–4783. IEEE.

Sun, G.; Zhang, Y.; Weiss, R. J.; Cao, Y.; Zen, H.; Rosenberg, A.; Ramabhadran, B.; and Wu, Y. 2020. Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6699–6703. IEEE.

Tan, X.; Chen, J.; Liu, H.; Cong, J.; Zhang, C.; Liu, Y.; Wang, X.; Leng, Y.; Yi, Y.; He, L.; et al. 2022. Naturalspeech: End-to-end text to speech synthesis with human-level quality. *arXiv preprint arXiv:2205.04421*.

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

Wang, J.; Wang, Z.; Hu, X.; Li, X.; Fang, Q.; and Liu, L. 2022. Residual-guided personalized speech synthesis based on face image. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4743–4747. IEEE.

Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; and Loy, C. C. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, 700–717. Springer.

Wang, Y.; Stanton, D.; Zhang, Y.; Ryan, R.-S.; Battenberg, E.; Shor, J.; Xiao, Y.; Jia, Y.; Ren, F.; and Saurous, R. A. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, 5180–5189. PMLR.

Yang, D.; Liu, S.; Huang, R.; Lei, G.; Weng, C.; Meng, H.; and Yu, D. 2023. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *arXiv preprint arXiv:2301.13662*.

Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R. J.; Jia, Y.; Chen, Z.; and Wu, Y. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech 2019*, 1526–1530.

Zhang, Y.-J.; Pan, S.; He, L.; and Ling, Z.-H. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6945–6949. IEEE.

Zhao, G.; Huang, X.; Taini, M.; Li, S. Z.; and PietikäInen, M. 2011. Facial expression recognition from near-infrared videos. *Image and vision computing*, 29(9): 607–619.