FaceSpeak: Expressive and High-Quality Speech Synthesis from Human Portraits of Different Styles

Tian-Hao Zhang[1,*], **Jiawei Zhang**[1], **Jun Wang**[2], **Xinyuan Qian**[1,†], **Xu-Cheng Yin**[1]

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
[2]Tencent AI Lab, Shenzhen, China
tianhaozhang@xs.ustb.edu.cn, jiaweizhang@xs.ustb.edu.cn

Abstract

Humans can perceive speakers' characteristics (e.g., identity, gender, personality and emotion) by their appearance, which are generally aligned to their voice style. Recently, vision-driven Text-to-speech (TTS) scholars grounded their investigations on real-person faces, thereby restricting effective speech synthesis from applying to vast potential usage scenarios with diverse characters and image styles. To solve this issue, we introduce a novel FaceSpeak approach. It extracts salient identity characteristics and emotional representations from a wide variety of image styles. Meanwhile, it mitigates the extraneous information (e.g., background, clothing, and hair color, etc.), resulting in synthesized speech closely aligned with a character's persona. Furthermore, to overcome the scarcity of multi-modal TTS data, we have devised an innovative dataset, namely Expressive Multi-Modal TTS (EM^{2}TTS), which is diligently curated and annotated to facilitate research in this domain. The experimental results demonstrate our proposed FaceSpeak can generate portrait-aligned voice with satisfactory naturalness and quality.

Demos — https://facespeak.github.io

Introduction

Human voices contain munificent information in aspects such as age (Grzybowska and Kacprzak 2016; Singh et al. 2016), gender (Li et al. 2019), emotional nuances (Wang and Tashev 2017; Zhang, Wu, and Schuller 2019), physical fitness (Verde et al. 2021), and speaker identity (Deaton 2010; Ravanelli and Bengio 2018). These vocal characteristics are intrinsically related to an individual's physical and psychological makeup (Xu et al. 2024; Hardcastle, Laver, and Gibbon 2012), offering a unique profile of the speaker. For example, the emotional content conveyed through speech is often mirrored in facial expressions. This inherent correlation between voice and image sparked research exploration in various fields, including emotion recognition (Zhou et al. 2021; Lei and Cao 2023; Zhang et al. 2021), speaker verification (Qian, Chen, and Wang 2021; Nawaz et al. 2021), face-speech retrieval (Li et al. 2023a), and speech separation (Gao and Grauman 2021; Lee et al. 2021).

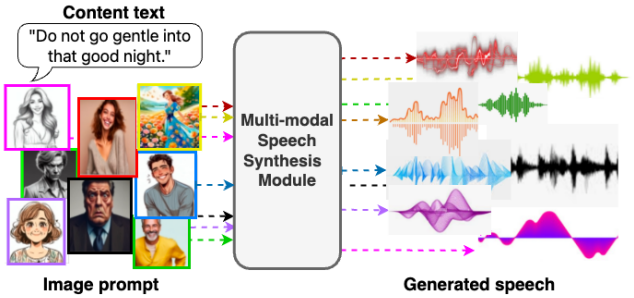


Figure 1: Our proposed multi-modal speech synthesis framework, namely FaceSpeak, which performs expressive and high-quality speech synthesis, given *image prompt* of different styles and the *content text* (Note: image-speech data from various characters are encoded with distinct color codecs).

In recent years, growing research interest has focused on the controllability of synthesized speeches. To this end, one solution is to introduce an auxiliary reference input to model the style features of speech. For example, PromptTTS (Guo et al. 2023) and InstructTTS (Yang et al. 2023) leverage textual descriptions to control the speech synthesis style. In contrast, the input text description still needs human crafting efforts and expertise, while some individuals may struggle to accurately express their intended synthesis goals. Other works employ vision as a reference. For example, visualTTS (Lu et al. 2022) generates temporal synchronized speech sequences for visual dubbing, and MM-TTS (Guan et al. 2024) transfers multi-modal prompts (i.e., text, human face, pre-recorded speech) into a unified style representation to control the generation process. Despite both works using images (real human faces) as the reference, thus cannot adapt to nonphotorealistic portraits which are widespread in digital assistants, video games, and virtual reality scenarios.

Although previous attempts have been made to generate speech based on visual cues, they have faced notable limitations. One such limitation is the predominant reliance on real-face datasets, which lack diversity in image styles necessary for comprehensive speech synthesis guidance. This significantly restricts the potential applications of synthesizing speech from portrait images. Additionally, prior methods often employ entangled embeddings of facial images to

[*]Work done during internship at Tencent AI Lab.
[†]Corresponding author.

guide speech synthesis, potentially introducing extraneous information that hampers performance. Moreover, relying on entangled visual features can further constrain the flexibility of the synthesis system, as the synthesized speech in this case can only be controlled by a single image. However, decoupling identity and emotion features can enable the control of speech synthesis by using different images providing identity and emotion information separately, greatly increasing the diversity and flexibility of multi-modal TTS.

In this paper, we tackle the aforementioned issues through a novel multi-modal speech synthesis process. As shown in Fig. 1, given the text of the content and the images in different styles (e.g., fantasy art and cartoon), our aim is to generate high-quality vivid human speech that is aligned with the characteristics indicated by vision. Our key contributions are summarized as follows.

1. We introduce EM^{2}TTS, a pioneering multi-style, multi-modal TTS dataset. It is designed and re-annotated through a collaborative multi-agent framework which leverages chatGPT[1] for crafting intermediate textual descriptions, PhotoMaker (Li et al. 2023b) for generating human portraits from text, and DALL-E to establish multi-modal coherence. It provides large-scale and diverse style images that enable the training model to generate high-quality and image-coherent speech, thus facilitates the state-of-the-art (SOTA) multi-modal TTS development.
2. We propose a novel speech synthesis method given human portrait prompt, namely FaceSpeak, to generate speech that is aligned with the characteristics indicated by the visual input. To our best knowledge, this is the first multi-modal expressive speech synthesis work that allows input any-style images. In particular, we disentangle the identity and expression features from facial images, ensuring that the synthesized audio aligns with the speaker's characteristics, while mitigating the impact of irrelevant factors in the images and enhancing the flexibility and diversity of synthesis systems.
3. Extensive experiments demonstrate that our proposed FaceSpeak can synthesize image-aligned, high-quality, diverse, and expressive human speech. Its superior performance is also validated through the numerous subjective and objective evaluations.

Related Work

Existing TTS works utilizing prompts leverage a multitude of modalities, including reference speech, textual descriptions, and human faces.

Speech prompt: Traditional TTS system extracts features from a reference speech to obtain the desired voice with unique vocal characteristics. For example, Meta-StyleSpeech (Min et al. 2021), which is built on FastSpeech2 (Ren et al. 2020), fine-tunes the gain and bias of textual input based on stylistic elements extracted from a speech reference, facilitating effective style-transferred speech synthesis. YourTTS (Casanova et al. 2022), based on VITS, proposes modifications for zero-shot multi-speaker and multilingual training, resulting in good speaker similarity and speech quality. GenerSpeech (Huang et al. 2022) proposes a multi-level style adapter and a generalizable content adapter to efficiently model style information. Mega-TTS (Jiang et al. 2023) employs various techniques (e.g., VQ-GAN, codec-LM) to extract different speech attributes (e.g., content, timbre, prosody, and phase), leading to successful speech disentanglement. Despite the impressive results, they are still limited by the availability of pre-recorded reference speech with a clean background. Moreover, the synthesized speech is often limited by the intrinsic attributes of the reference speech.

Text prompt: In contrast to traditional TTS systems that require users to have acoustic knowledge to understand style elements such as prosody and pitch, the use of text prompts is more user-friendly as text descriptions offer a more intuitive and natural means of expressing speech style. For example, PromptTTS (Guo et al. 2023) utilizes the BERT model as a style encoder and a transformer-based content encoder to extract the corresponding representations from the text prompt to achieve voice control. InstructTTS (Yang et al. 2023) proposes a novel three-stage training procedure to obtain a robust sentence embedding model that can effectively capture semantic information from style prompts and control the speaking style in the generated speech. Sally (Ji et al. 2024) employs an autoregressive codec-LM as a style encoder and a non-autoregressive codec-LM as a decoder to generate acoustic tokens across varying granularities, capitalizing on the robust coding capabilities of language models. Promptspeaker (Zhang et al. 2023) integrates the Glow model to create an invertible mapping between semantic and speaker representations, thereby enabling text-driven TTS. Despite promising results, these works still rely on human effort to provide detailed text descriptions, which may be unavailable at scenarios requiring rapid content creation.

Image prompt: Image prompt-based TTS allows a more comprehensive and expressive synthesis process, as the visual context can provide additional information and nuances that enhance the overall quality and authenticity of the generated speech. For example, VisualTTS (Lu et al. 2022) pioneers the use of silent pre-recorded videos as conditioned inputs to not only generate human-like speech but also achieve precise lip-speech synchronization. Similarly, Imaginary Voice (Lee, Chung et al. 2023) introduces a face-styled diffusion TTS model within a unified framework which designs a speaker feature binding loss to enforce similarity between the generated and real speech segments in the speaker embedding space. To be noted, MMTTS (Guan et al. 2024) proposes an aligned multi-modal prompt encoder that embeds three different modalities into a unified style space. Despite allowing any modality input, it limits speech generation to real faces, overlooking the potential influence of images with diverse styles.

Summary: Considering the aforementioned limitations, in this paper, we aim to generate expressive and high-quality human speech associated with characters from input images of various styles. Henceforth, we introduce a portrait-speech pairing dataset comprising multi-style images. Using this

[1]https://chat.openai.com/

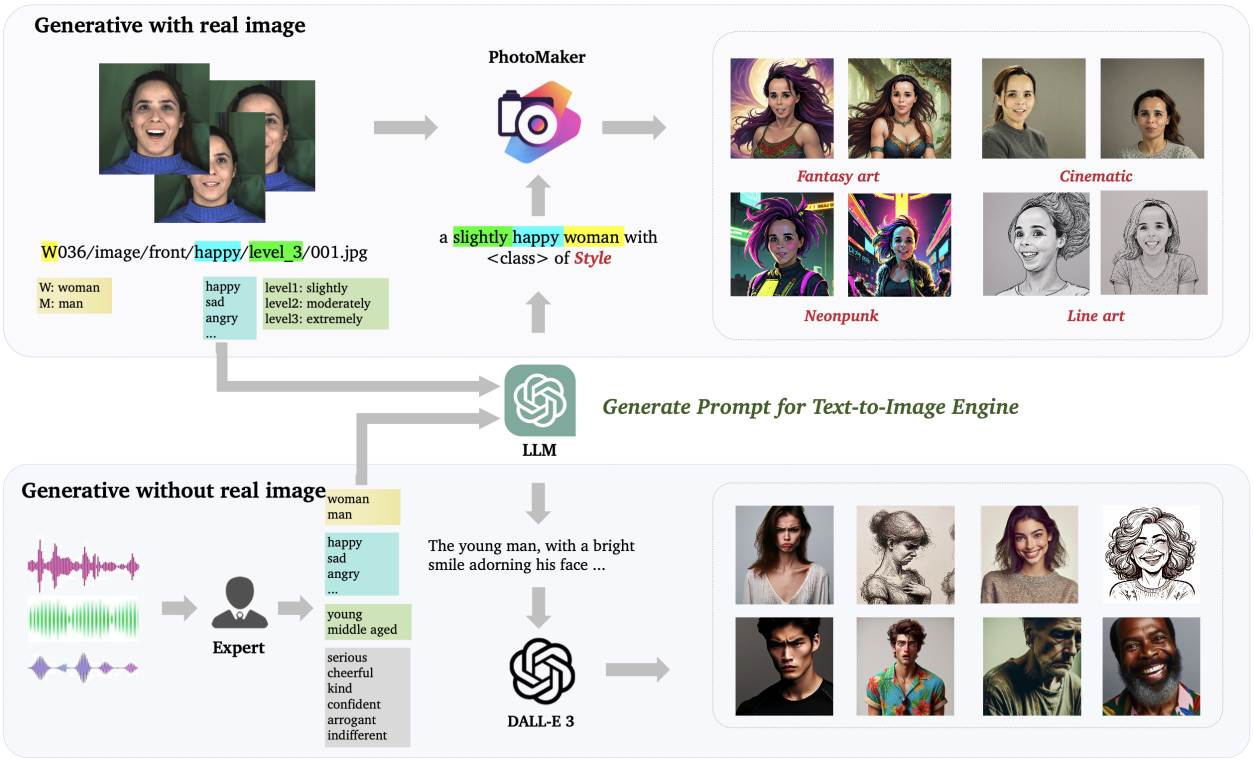


Figure 2: Our image generation pipeline of 1) EM^{2}TTS-MEAD subset (top): we specify the desired output style and transfer the real human image to images of different styles using PhotoMaker. 2) EM^{2}TTS-ESD-EmovDB subset (bottom): we use a *human expert* to label the character factors for chatGPT to create the descriptive text, which is utilized by DALL-E-3 to produce images that are highly aligned with the specified parameters.

dataset, we develop FaceSpeak, enabling the model to be generalized across various image styles. In particular, our FaceSpeak framework achieves enhanced accuracy and flexibility in controlling visual features for synthesized speech, which is achieved through the deliberate decoupling of identity and emotion information within the visual features of the portrait.

Proposed EM^{2}TTS Dataset

The widely-used TTS datasets, such as LibriTTS (Zen et al. 2019) and VCTK (Yamagishi et al. 2019), predominantly exhibit a single-modal nature, comprising solely audio recordings without corresponding textual or visual labels. Existing multi-modal TTS datasets, including ESD (Zhou et al. 2022), EmovDB[2], and Expresso (Nguyen et al. 2023), lack visual labels for character profiles. While IEMOCAP (Busso et al. 2008), MEAD (Wang et al. 2020), CMU-MOSEI (Zadeh et al. 2018) and RAVDESS (Livingstone and Russo 2018) datasets provide labels with limited aspects such as emotion and facial expressions, no existing TTS datasets provide face data with diverse image styles.

To address the above limitations, we re-design and annotate an expansive multi-modal TTS dataset, termed EM^{2}TTS. It is enriched with diverse text descriptions and a wide range of facial imagery styles. Due to the distinct data characteristics and varying levels of annotation completion, we have delineated the dataset into two distinct emotional subsets. Please see *Appendix* for more details.

EM^{2}TTS-MEAD

For the MEAD dataset (Wang et al. 2020) with real human face recordings, we initially applied a random selection strategy to extract video frames. Then, following steps are designed to proceed the data: 1) *Automatic text generation*: we generate corresponding text labels that encompass gender, emotion and its intensity, utilizing the raw data from the MEAD dataset. In particular, our methodology draws from the MMTTS framework (Guan et al. 2024), which correlates emotion intensity levels with specific degree words, as shown in Figure 2. 2) *Image style transfer*: we leverage the innovative image style conversion model, PhotoMaker (Li et al. 2023b), which excels in producing a variety of portrait styles representing the same individual, guided by character images and textual prompts. For each speaker exhibiting varying emotion intensities, we have created four styles of images (i.e., fantasy art, cinematic, neonpunk, and line art), enriching the visual diversity of the dataset.

Nevertheless, a significant challenge arises in discerning the correlation between a speaker's appearance and their timbre, particularly with *hard samples*. These cases involve

[2]https://www.openslr.org/115

a mismatch between the physical attributes and the vocal characteristics of the speaker (e.g., a physically strong person's voice may sound similar to that of a slim woman). Consequently, models trained exclusively on EM^{2}TTS-MEAD face difficulties in aligning with intuitive expectations of human perception.

EM^{2}TTS-ESD-EmovDB

The unimodal emotional speech datasets ESD and EmovDB lack accompanying speaker images, thus performing style transfer based on real human face is unfeasible. Therefore, we design the next steps to process them: 1) *Manual annotation*: we explore a human expert to label age, gender, and characteristics by listening to each speech data; 2) *Text expansion*: we use the Large Language Model (LLM) model e.g., ChatGPT to expand the label words into texts with varying contents but similar meanings; 3) *Text-driven image generation*: the enriched texts were then fed into DALL-E-3, a text-to-image model capable of generating a multitude of images in distinct styles. We assigned these generated images to the corresponding speeches, resulting in style-free images that emphasize the character's emotions, surpassing the limitations of the recorded images.

Proposed Method

Let us denote $\mathbf{I}_i$, $\mathbf{x}_i$ and $\mathbf{t}_i$ as the visual image of arbitrary style, the corresponding voice signal, and the co-speech content of character i, respectively. As depicted in Fig. 3, our proposed FaceSpeak algorithm aims to generate a speech waveform $\mathbf{s}_i$ corresponding to the imagined voices of characters given the input text $\mathbf{t}_i$ and images of different styles (either real $\mathbf{I}_i^R$ or generated $\mathbf{I}_i^G$).

The proposed FaceSpeak consists of two sub-modules: 1) Multi-style image feature decomposition module that receives different styles of portrait images for feature extraction and decouples emotion and speaker information. With this module, we can get the disentangled identity and emotion embeddings extracted from the portrait visual features; 2) Expressive TTS module receives the identity and emotion embeddings as control vectors for generating high-quality speech that matches the portrait images. Detailed descriptions are given below.

Multi-Style Image Feature Disentanglement

To enhance the coherence of identity and emotion between the input image and the synthesized speech, it is crucial to mitigate the influence of extraneous visual elements (e.g., background, clothing, distracting objects) in the extracted visual features. The FaRL model (Zheng et al. 2022), leveraging the multi-modal pre-trained CLIP model on large-scale datasets of face images and correlated text, ensures the extraction of predominantly face-related visual features with robust generalization capabilities. Thus, we apply it on real images and corresponding multi-style images to extract high-dimensional intermediate visual representations:

$$\mathbf{e}_i = \mathrm{FaRL}(\mathbf{I_i}) \tag{1}$$

where $\mathbf{e}_i \in \mathbf{R}^{512}$ contains both the emotion and speaker information of the portrait.

Our subsequent objective is to decouple the emotion and the speaker information within $\mathbf{e}_i$. Let us define the Identity Adapter Module (IAM) and the Expression Adapter Module (EAM) to learn the mapping from $\mathbf{e}_i$ to the identity embedding $\boldsymbol{\alpha}_i$ and emotion embedding $\boldsymbol{\beta}_i$, respectively:

$$\begin{aligned} \boldsymbol{\alpha}_i &= \mathrm{IAM}(\mathbf{e}_i) = \mathrm{FC}\left(\mathrm{GeLU}(\mathrm{FC}(\mathbf{e}_i))\right) \\ \boldsymbol{\beta}_i &= \mathrm{EAM}(\mathbf{e}_i) = \mathrm{FC}\left(\mathrm{GeLU}(\mathrm{FC}(\mathbf{e}_i))\right) \end{aligned} \tag{2}$$

Subsequently, we used the emotion classification model following $\boldsymbol{\beta}_i$ to bias the $\boldsymbol{\beta}_i$ features toward emotion characteristics. Furthermore, we incorporated the emotion classification model after introducing Gradient Reverse Layer (GRL) following $\boldsymbol{\alpha}_i$, which aims to minimize the sentiment information retained in $\boldsymbol{\alpha}_i$. We utilize the Cross-Entropy loss to constrain the two emotion classification models, formulated as follows:

$$\begin{aligned} \mathcal{L}_{emo} &= \mathrm{CrossEntropy}(\mathrm{CLS}(\boldsymbol{\beta}_i), \mathrm{L}_e) \\ \mathcal{L}_{grl} &= \mathrm{CrossEntropy}(\mathrm{GRL}(\mathrm{CLS}(\boldsymbol{\alpha}_i)), \mathrm{L}_e) \end{aligned} \tag{3}$$

where CLS represents the classification layer, and L_e denotes the emotion categorization label of the input image $\mathbf{I_i}$. GRL inverts the sign of the incoming gradient during the back-propagation phase. By this strategic reversal, IAM learns to remove or minimize features that are correlated with emotion, emphasizing the identity aspects of the input.

To enhance the decoupling of identity embedding $\boldsymbol{\alpha}_i$ and emotion embedding $\boldsymbol{\beta}_i$, we propose a feature decoupling approach that leverages Mutual information (MI) minimization. This strategy effectively reduces the correlation between identity and emotion representations, enabling a more robust and accurate analysis of each aspect.

Mutual information based decoupling: MI is a fundamental concept in information theory that quantifies the statistical dependence between two random variables V_1 and V_2, calculated as:

$$I(V_1; V_2) = \sum_{v_1 \in V_1} \sum_{v_2 \in V_2} p(v_1, v_2) \log \left(\frac{p(v_1, v_2)}{p(v_1)p(v_2)} \right) \tag{4}$$

where $p(v_1, v_2)$ is the joint probability distribution between v_1 and v_2, while $p(v_1)$ and $p(v_2)$ are their marginals. However, it is still a challenge to obtain a differentiable and scalable MI estimation. In this work, we use vCLUB (Cheng et al. 2020), an extension of CLUB, to estimate an upper bound on MI, which allows efficient estimation and optimization of MI when only sample data are available and a probability distribution is not directly available. Given the sample pairs of identity embedding and emotion embedding $\{(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)\}_{i=1}^N$, where N denotes the number of samples, the MI can be computed as:

$$\mathcal{L}_{mi} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[\log q_\theta(\boldsymbol{\beta}_i | \boldsymbol{\alpha}_i) - \log q_\theta(\boldsymbol{\beta}_j | \boldsymbol{\alpha}_i) \right] \tag{5}$$

where $q_\theta(\boldsymbol{\beta}_i | \boldsymbol{\alpha}_i)$ is a variational approximation which can make vCLUB holds a MI upper bound or become a reliable MI estimator. At each iteration during the training stage, we first obtain a batch of samples $\{(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)\}$ from IAM and EAM, then update the variational approximation $q_\theta(\boldsymbol{\beta}_i | \boldsymbol{\alpha}_i)$ by maximizing the log-likelihood $\mathcal{L}_\theta =$

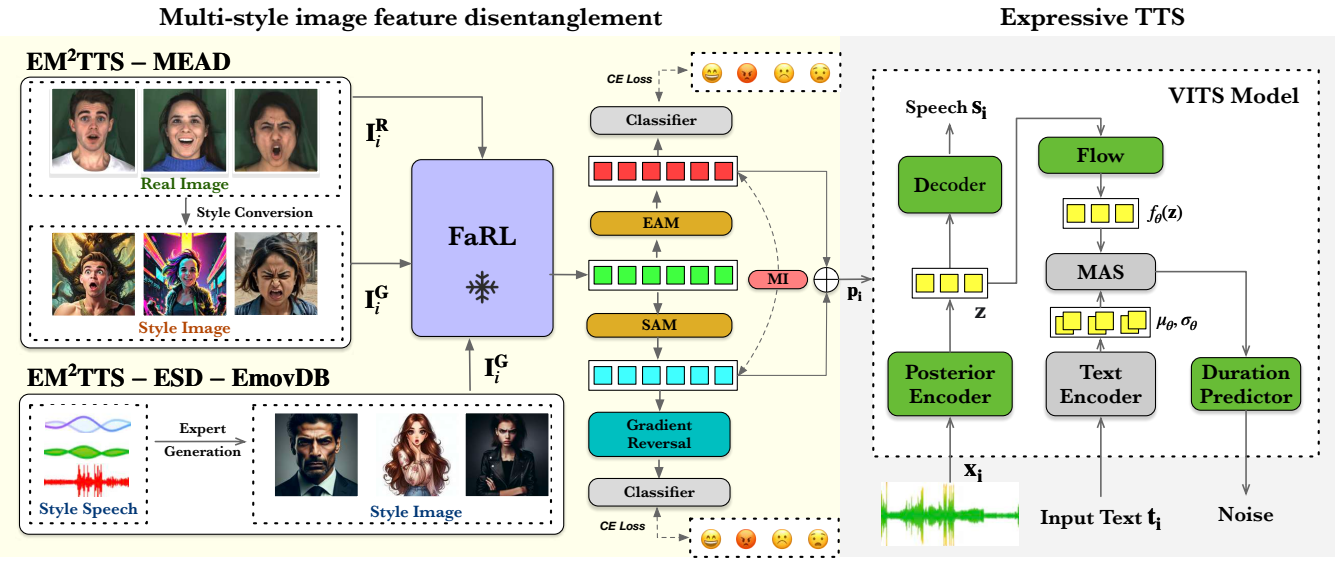


Figure 3: Block diagram of our proposed FaceSpeak which generates speech given the input text $\mathbf{t}_i$ and images of different styles (either real $\mathbf{I}_i^R$ or generated $\mathbf{I}_i^G$). It consists of two sub-modules: multi-style image feature disentanglement (yellow region) and expressive TTS (gray region).

$\frac{1}{N}\sum_{i=1}^{N}\log q_\theta(\boldsymbol{\beta}_i|\boldsymbol{\alpha}_i)$. The updated $q_\theta(\boldsymbol{\beta}_i|\boldsymbol{\alpha}_i)$ can be used to calculate the vCLUB estimator. Finally, we sum the decoupled identity embedding and the emotion embedding obtained as the ultimate control embedding p_i for speech synthesis.

Expressive TTS

We use VITS2 (Kong et al. 2023), one of the SOTA TTS models, as our speech synthesis backbone. As shown in Figure 3, The VITS2 model consists of a Posterior Encoder and a Text Encoder that generate posterior distribution and prior distribution based on the input speech and text, respectively; a transformer-based Flow module to refine the latent representation produced by the posterior encoder; a Monotonic Alignment Search (MAS) module to estimate an alignment between input text and target speech; a Duration Predictor module to predict the duration of each phoneme; a Decoder to reconstructs the speech from the latent representation generated by the posterior encoder. In particular, we injected the control embedding from the portrait images into the Posterior Encoder, Decoder, Flow module, and Duration Predictor to generate the speech corresponding to the portrait images, which are highlighted in green in Figure 3. In the training stage, identity embedding $\boldsymbol{\alpha}_i$ and emotion embedding $\boldsymbol{\beta}_i$ are decoupled from the same image and the final loss can be expressed as:

$$\mathcal{L} = \mathcal{L}_{vits} + \lambda_1\mathcal{L}_{mi} + \lambda_2\mathcal{L}_{emo} + \lambda_3\mathcal{L}_{grl} \tag{6}$$

where λ_1, λ_2 and λ_3 are the hyper-parameters to balance the individual losses. During inference, $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ can come from the same image or be provided by different images separately, and the inference process of Expressive TTS is consistent with work (Kong et al. 2023).

Experiments

Dataset and Experimental Setup

In the evaluation of our method, we conduct separate experiments on intra-domain and out-of-domain data, respectively. We use EM^{2}TTS-MEAD as the intra-domain experimental data, while for the out-of-domain evaluation, we perform different settings for real portrait scenes and multi-style virtual portrait scenes, respectively. For the real portrait scenes, we follow the MMTTS's setup which uses the face images in Oulu-CASIA dataset and transcriptions in LibriTTS, while for the multi-style virtual portrait scenario, we use a different image generation API from the training set to generate new test data based on EM^{2}TTS-ESD. Specifically, we compare our method with the following system:1) GT: Ground Truth. 2) VITS2: A multispeaker TTS baseline. 3) MMTTS: A style transfer TTS system using prompt including image. 4) MM-StyleSpeech: Same as MMTTS using StyleSpeech as backbone. The proposed FaceSpeak is trained for 150K iterations using the Adam optimizer on NVIDIA GeForce RTX 3090 GPUs. Detailed training parameters and network configuration can be found in the *Appendix*.

Synthetic Quality on Real Portraits

We first evaluate the quality of the speech synthesized by FaceSpeak given a real portrait as the prompt. We conduct a Mean Opinion Score (MOS) (Sisman and Yamagishi 2021)

Method	Intra-domain			Out-of-domain		
	NMOS ↑	ISMOS ↑	ESMOS ↑	NMOS ↑	ISMOS ↑	ESMOS ↑
GT	4.42 ± 0.02	-	4.52 ± 0.03	-	-	-
VITS2	3.55 ± 0.06	3.68 ± 0.07	3.38 ± 0.13	3.42 ± 0.05	3.56 ± 0.10	3.31 ± 0.09
MM-StyleSpeech	3.58 ± 0.08	3.64 ± 0.04	3.89 ± 0.11	3.23 ± 0.08	3.61 ± 0.07	3.78 ± 0.08
MM-TTS	3.94 ± 0.05	3.82 ± 0.08	4.08 ± 0.08	3.41 ± 0.06	3.68 ± 0.04	3.91 ± 0.05
FaceSpeak	4.13 ± 0.04	3.97 ± 0.07	4.36 ± 0.05	4.28 ± 0.05	3.77 ± 0.09	3.98 ± 0.07

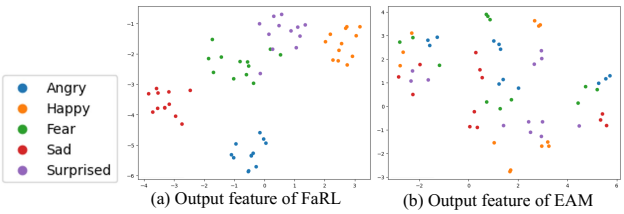
Table 1: MOS results with 95% confidence interval (N-: naturalness; IS-: identity similarity; ES-: emotion similarity; -: information not applicable.)

Method	Intra-domain				Out-of-domain			
	7-point score ↑	Preference (%)			7-point score ↑	Preference (%)		
		B	E	O		B	E	O
MM-StyleSpeech	1.25 ± 0.08	28	22	50	1.97 ± 0.12	10	15	75
MM-TTS	1.03 ± 0.06	33	14	53	1.42 ± 0.09	16	24	60

Table 2: AXY preference test results. B, E and O respect the preference rate for baseline model, equivalent and our model, respectively.

with 95 % confidence intervals to assess speech quality. Naturalness-MOS (NMOS) and Similarity-MOS (SMOS) evaluate the speech naturalness and image-speech similarity. Specifically, we generate 50 speech samples for each model, which are rated by 20 volunteers on a scale of 1 to 5, with higher scores indicating better results (↑). We also perform an AXY preference test (Skerry-Ryan and Battenberg 2018) to verify the style transfer effect based on image prompt. In the test, "A" is the reference speech that is stylistically consistent with the image prompt, "X" and "Y" are the speech generated by the compared model or our proposed FaceSpeak. The participants decide whether the speech style of "X" or "Y" is closer to that of image prompt where the scale range of [-3,3] indicating "X" is closer to "Y" is closer. Table 1 displays the subjective results including NMOS, ISMOS and ESMOS. Our FaceSpeak achieves better results on both speech naturalness and style similarity, showing the effectiveness of DSP module on speaker representation extraction by using IAM and EAM. As shown in Table 2, the results of AXY test indicate that listeners prefer FaceSpeak synthesis against the compared models. The generated data and the method significantly improves the style extraction ability, allowing an arbitrary reference sample to guide the stylistic synthesis of arbitrary content text.

As shown in the Table 3, we further objectively measure the quality of speech through MCD (Kubichek 1993), emotion and gender classification accuracy, and speaker similarity. MCD measures the spectral distance between the reference and synthesized speech and FaceSpeak achieved a MCD result of 3.32. We measure speaker similarity (SS) between two speech samples in Resemblyzer [3], our result is 0.95. A hubert-based pre-trained model [4] is used for gender classification (Acc_{gen}) and emotion2vec is used to predict the emotion category (Acc_{emo}) of speech. On the intra-domain data, we obtained Acc_{gen} for 99.40% and Acc_{emo} for 60.92%. For out-of-domain data, the results for Acc_{gen} and Acc_{emo} are 92.42% and 31.32%, respectively.



(a) Output feature of FaRL (b) Output feature of EAM

Figure 4: Visualization of emotion embeddings (colors index emotions).

Synthetic Quality on Multi-Style Virtual Portraits

In evaluating the quality of FaceSpeak's speech synthesis based on multi-style virtual portraits, the models we compare are whether or not trained with the virtual portrait images of our proposed multi-style dataset EM^{2}TTS, respectively. Since models trained with EM^{2}TTS will have "seen" multi-style portraits in the domain, our evaluation will only be performed on out-of-domain multi-style portraits. As illustrated in Table 4, the model trained with the virtual portrait images performs better in all evaluation metrics, confirming the effectiveness of our methods.

Results of Decoupled Identity and Emotion Information

TSNE results of decoupled features: Fig. 4 (a) visualizes the embeddings extracted from FaRL in emotion, which are randomly distributed. By applying the EAM module, as shown in Fig. 4 (b), the learned embeddings with the same

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

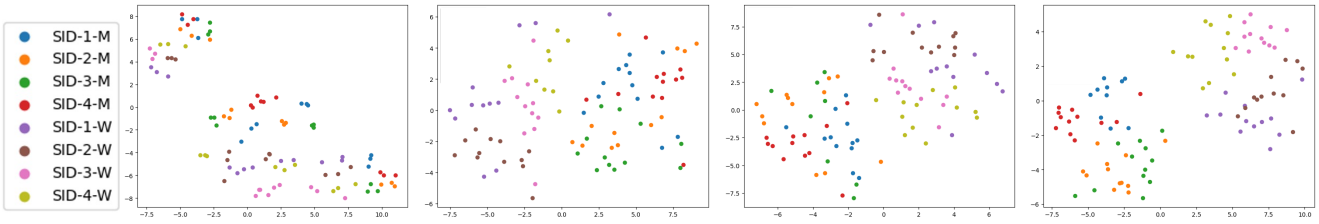
[4]https://github.com/m3hrdadfi/soxan.git

Method	Intro-domain			Out-of-domain		
	MCD ↓	ACC_{emo} ↑	ACC_{gen} ↑	SS ↑	ACC_{emo} ↑	ACC_{gen} ↑
GT	-	84.54	100.00	-	-	-
FaceSpeak	3.32	60.92	99.40	0.95	31.32	92.42

Table 3: Subjective results on real portraits controlled speech synthesis.

Method	NMOS ↑	ISMOS ↑	ESMOS ↑	ACC_{emo} ↑	ACC_{gen} ↑
w/o EM^{2}TTS	4.31 ± 0.05	3.88 ± 0.09	4.02 ± 0.06	18.34	84.24
w/ EM^{2}TTS	4.38 ± 0.06	4.06 ± 0.08	4.47 ± 0.04	26.56	92.22

Table 4: Objective and Subjective results on out-of-domain multi-style portraits controlled.



(a) Output feature of FaRL (b) w/o vCLUB and GRL (c) w/ GRL and w/o vCLUB (d) w/ GRL and vCLUB

Figure 5: Visualization of identity embeddings (colors index identities; M: male; F: female).

emotion are more clustered, while the others are more discriminative. Fig. 5 (a) shows the embeddings extracted from FaRL in identity, which are also randomly distributed. By applying the IAM module, from Fig. 5 (b) to (d), it is observed that embeddings from different speakers are more distinguished when using both the GRL and vCLUB strategies. These intuitively verify the effectiveness of our decoupled identity and emotion characteristics.

Speech synthesis controlled by combined portraits: By decoupling identity and emotion information in an image, we can use different image combinations to control the synthesized speech. As shown in Figure 6, we define $\mathbf{X}$ as the image that provides identity embedding and $\mathbf{Y}$ as the image that provides emotion embedding, ensuring that $\mathbf{X}$ and $\mathbf{Y}$ have different genders and emotions. We let listeners discriminate the synthesized speech by deciding whether it is matched to the $\mathbf{X}$-image or $\mathbf{Y}$-image in terms of identity and emotion, respectively. As a result, 98.6% of the speech are determined to correctly match the $\mathbf{X}$ image on identity, while 92.1% of the speech are determined to correctly match the $\mathbf{Y}$ image on emotion, which proves that our proposed FaceSpeak speech synthesis system can reliably control the synthesis by combining different images, greatly improving the diversity and flexibility.

Figure 6: The accuracy in matching the synthesized speech to the emotion and identity styles when they are controlled separately by different images.

Conclusion

In this paper, we introduce FaceSpeak, a pioneering approach for multi-modal speech synthesis which extracts key identity and emotional cues from diverse character images to drive the TTS module to synthesize the corresponding speech. To tackle the problem of data scarcity, we introduced an innovative EM^{2}TTS dataset, which is meticulously curated and annotated to support and advance research in this emerging field. Additionally, novely methods are proposed for decoupling emotional and speaker-specific features, enhancing both the adaptability and fidelity of our system. Experimental results confirm that FaceSpeak can generate high-quality, natural-sounding speech that authentically align with the visual attributes of the character.

Looking ahead, we aspire to broaden the diversity of speaker categories within the Facespeak system, integrating a wider array of emotions, roles, and other dynamic attributes. By leveraging larger, more comprehensive datasets, we aim to advance the system's development, enhancing its adaptability and versatility.

Acknowledgments

This work is supported by CCF-Tencent Rhino-Bird Open Research Fund, National Natural Science Foundation of China under Grant No. 62306029, Beijing Natural Science Foundation under Grants L233032, Shenzhen Research Institute of Big Data under Grant No. K00120240007.

References

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; et al. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359.

Casanova, E.; Weber, J.; Shulby, C. D.; Junior, A. C.; et al. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *Proc. of Int. Conf. on Machine Learning*, 2709–2720. PMLR.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; et al. 2020. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*.

Deaton, A. 2010. Understanding the mechanisms of economic development. *Journal of Economic Perspectives*, 24(3): 3–16.

Gao, R.; and Grauman, K. 2021. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, 15490–15500. IEEE.

Grzybowska, J.; and Kacprzak, S. 2016. Speaker age classification and regression using i-vectors. In *Proc. of Interspeech*, 1402–1406.

Guan, W.; Li, Y.; Li, T.; Huang, H.; Wang, F.; Lin, J.; Huang, L.; Li, L.; and Hong, Q. 2024. MM-TTS: Multi-modal Prompt based Style Transfer for Expressive Text-to-Speech Synthesis.

Guo, Z.; Leng, Y.; Wu, Y.; Zhao, S.; and Tan, X. 2023. PromptTTS: Controllable text-to-speech with text descriptions. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, 1–5. IEEE.

Hardcastle, W. J.; Laver, J.; and Gibbon, F. E. 2012. The handbook of phonetic sciences.

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

Ji, S.; Zuo, J.; Fang, M.; Jiang, Z.; Chen, F.; et al. 2024. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, 10301–10305. IEEE.

Jiang, Z.; Ren, Y.; Ye, Z.; Liu, J.; Zhang, C.; Yang, Q.; Ji, S.; et al. 2023. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *arXiv preprint arXiv:2306.03509*.

Kong, J.; Park, J.; Kim, B.; Kim, J.; et al. 2023. VITS2: Improving Quality and Efficiency of Single-Stage Text-to-Speech with Adversarial Learning and Architecture Design. *arXiv preprint arXiv:2307.16430*.

Kubichek, R. 1993. Mel-cepstral distance measure for objective speech quality assessment. *In Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1: 125–128.

Lee, J.; Chung, J. S.; et al. 2023. Imaginary voice: Face-styled diffusion model for text-to-speech. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, 1–5. IEEE.

Lee, J.; Chung, S.-W.; Kim, S.; Kang, H.-G.; and Sohn, K. 2021. Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, 1336–1345.

Lei, Y.; and Cao, H. 2023. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing*.

Li, S.; Dabre, R.; Lu, X.; Shen, P.; Kawahara, T.; and Kawai, H. 2019. Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation. In *Proc. of Interspeech*, 4400–4404.

Li, X.; Wen, Y.; Yang, M.; Wang, J.; Singh, R.; and Raj, B. 2023a. Rethinking Voice-Face Correlation: A Geometry View. In *Proc. of ACM Int. Conf. on Multimedia*, 2458–2467.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; et al. 2023b. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, 13(5): e0196391.

Lu, J.; Sisman, B.; Liu, R.; Zhang, M.; and Li, H. 2022. VisualTTS: TTS with Accurate Lip-Speech Synchronization for Automatic Voice Over. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*. IEEE.

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

Nawaz, S.; Saeed, M. S.; Morerio, P.; Mahmood, A.; et al. 2021. Cross-modal speaker verification and recognition: A multilingual perspective. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, 1682–1691.

Nguyen, T. A.; Hsu, W.-N.; d'Avirro, A.; Shi, B.; Gat, I.; Fazel-Zarani, M.; et al. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

Qian, Y.; Chen, Z.; and Wang, S. 2021. Audio-visual deep neural network for robust person verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1079–1092.

Ravanelli, M.; and Bengio, Y. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, 1021–1028. IEEE.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Singh, R.; Keshet, J.; Gencaga, D.; and Raj, B. 2016. The relationship of voice onset time and voice offset time to physical age. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, 5390–5394. IEEE.

Sisman, B.; and Yamagishi, J. 2021. An overview of voice conversion and its challenges: From statistical modeling to deep learning. In *IEEE/ACM Trans. on Audio, Speech and Language Processing*, 132–157.

Skerry-Ryan, R.; and Battenberg, E. 2018. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. *Proceedings of the 35th International Conference on Machine Learning(ICML)*, 80: 3–16.

Verde, L.; Giuseppe De Pietro, A. G.; Alrashoud, M.; et al. 2021. Exploring the Use of Artificial Intelligence Techniques to Detect the Presence of Coronavirus Covid-19 Through Speech and Voice Analysis. *IEEE Access*.

Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; et al. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Proc. of European Conf. on Computer Vision*, 700–717. Springer.

Wang, Z.-Q.; and Tashev, I. 2017. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, 5150–5154. IEEE.

Xu, C.; Liu, Y.; Xing, J.; Wang, W.; Sun, M.; Dan, J.; et al. 2024. FaceChain-ImagineID: Freely Crafting High-Fidelity Diverse Talking Faces from Disentangled Audio. *arXiv preprint arXiv:2403.01901*.

Yamagishi, J.; Veaux, C.; MacDonald, K.; et al. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).

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

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria; et al. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246.

Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R. J.; Jia, Y.; et al. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Zhang, S.; Ding, Y.; Wei, Z.; and Guan, C. 2021. Continuous emotion recognition with audio-visual leader-follower attentive fusion. In *Proc. of Int. Conf. on Computer Vision*, 3567–3574.

Zhang, Y.; Liu, G.; Lei, Y.; Chen, Y.; Yin, H.; Xie, L.; and Li, Z. 2023. Promptspeaker: Speaker Generation Based on Text Descriptions. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–7. IEEE.

Zhang, Z.; Wu, B.; and Schuller, B. 2019. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, 6705–6709.

Zheng, Y.; Yang, H.; Zhang, T.; Bao, J.; Chen, D.; et al. 2022. General facial representation learning in a visual-linguistic manner. In *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, 18697–18709.

Zhou, H.; Du, J.; Zhang, Y.; Wang, Q.; Liu, Q.-F.; and Lee, C.-H. 2021. Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 29: 2617–2629.

Zhou, K.; Sisman, B.; Liu, R.; and Li, H. 2022. Emotional voice conversion: Theory, databases and ESD. *Speech Communication*, 137: 1–18.