FleSpeech: Flexibly Controllable Speech Generation with Various Prompts

Anonymous ACL submission

Abstract

Controllable speech generation methods typically rely on single or fixed prompts, hindering creativity and flexibility. These limitations make it difficult to meet specific user needs in certain scenarios, such as adjusting the style while preserving a selected speaker's timbre, or choosing a style and generating a voice that matches a character's visual appearance. To overcome these challenges, we propose *FleSpeech*, a novel multi-stage speech generation framework that allows for more flexible manipulation of speech attributes by integrating various forms of control. FleSpeech employs a multimodal prompt encoder that processes and unifies different text, audio, and visual prompts into a cohesive representation. This approach enhances the adaptability of speech synthesis and supports creative and precise control over the generated speech. Additionally, we develop a data collection pipeline for multimodal datasets to facilitate further research and applications in this field. Comprehensive subjective and objective experiments demonstrate the effectiveness of FleSpeech. Audio samples are available at https://anyone499.github.io/FleSpeech/

1 Introduction

Speech synthesis plays a pivotal role in content creation and human-computer interaction. With the advancement of powerful generative models, such as large language models (Wang et al., 2023; Betker, 2023; Lajszczak et al., 2024; Anastassiou et al., 2024; Kim et al., 2024) and diffusion models (Vyas et al., 2023; Eskimez et al., 2024; Chen et al., 2024a), speech synthesis has experienced rapid progress in recent years (Xie et al., 2024). Beyond a focus on realism, there is a growing emphasis on *flexible and controllable* speech synthesis (Guan et al., 2024), such as the ability to manipulate the style of generated speech based on textual descriptions (Liu et al., 2023; Ji et al., 2024a; Leng et al., 2024; Zhu et al., 2024).

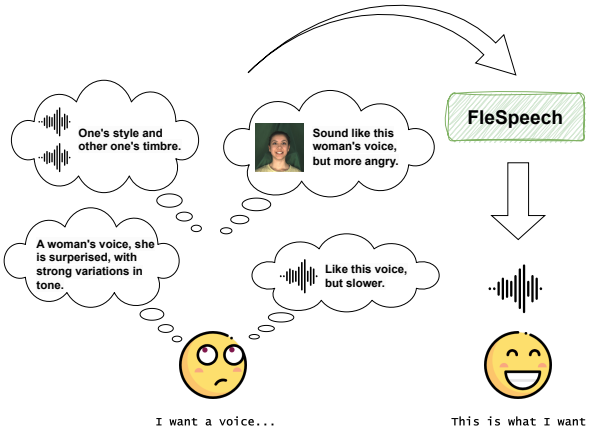


Figure 1: FleSpeech can flexibly generate speech that matches the given prompts.

Despite the variety of available speech generation control methods, each approach has its inherent limitations. For instance, while speech synthesis based on natural language descriptions offers flexibility, language often struggles to precisely capture all desired attributes, particularly when it comes to describing a speaker's timbre, as textual representations are inherently limited. In contrast, the reference audio-based method can clearly define all attributes but relies on existing audio, which lacks creativity and flexibility. These constraints make it difficult to address specific user needs in certain scenarios, such as adjusting style while preserving a selected speaker timbre or choosing a style and generating a voice that aligns with a character's visual appearance.

To overcome these constraints and move beyond controllable speech synthesis techniques based on a single or a few control methods, we propose a more flexible controllable speech generation method, *FleSpeech*, which supports multiple forms of control and allows for the combination of different

control strategies, thereby meeting the flexible control requirements across various scenarios as illustrated in Fig. 1. To this end, we first introduce a multi-stage speech generation framework, with each stage modeling the style and timbre of speech. With this framework, we can provide different prompts at different stages, enabling flexibly controllable speech generation. Second, we propose a multimodal prompt encoder to embed multimodal prompts into a unified representation. Finally, considering the scarcity of multimodal data, we built a data collection pipeline to facilitate research in this area. We will release this data collection pipeline upon the acceptance of this paper.

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

- We propose FleSpeech, a multi-stage speech generation framework that supports multiple prompt inputs to flexibly control different properties of speech. Experiments across different tasks demonstrate both the objective and subjective superiority of this method.
- We propose a unified multimodal prompt encoder, which allows us to input any combination of text, audio, and visual modal prompts and operate them in a unified embedding space.
- We built a pipeline to facilitate data collection for subsequent multimodal speech generation work.

2 Related Work

2.1 Controllable Speech Synthesis

The employment of category labels, such as speaker identity (Chen et al., 2020; Gibiansky et al., 2017) and emotion (Lee et al., 2017; Lorenzo-Trueba et al., 2018), serves as a prevalent technique for controlling specific speech attributes. To address the limited control capabilities of labels, Skerry-Ryan et al. (Skerry-Ryan et al., 2018) introduced a style transfer method based on reference acoustic representation. Subsequently, this reference audio-based approach has gained substantial popularity, particularly in the context of emotion transfer (Li et al., 2022; Lei et al., 2022) and zero-shot TTS (Wang et al., 2023; Kim et al., 2024; Du et al., 2024).

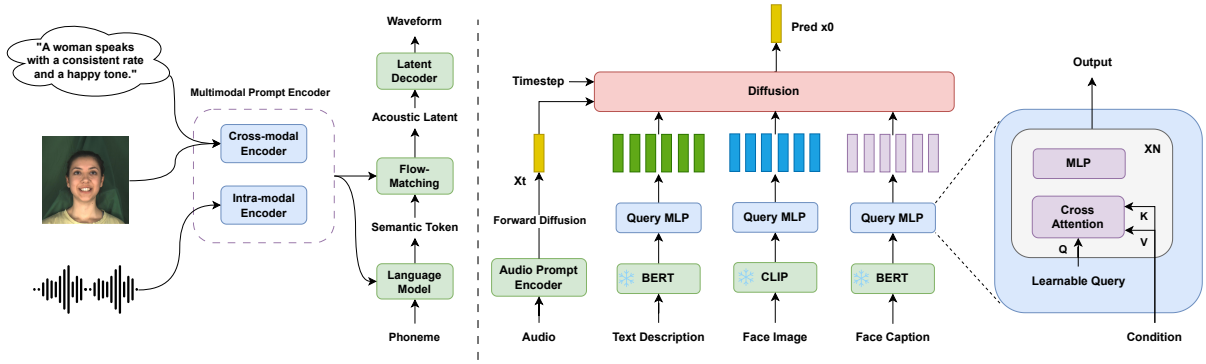
To achieve more flexible control, InstructTTS (Yang et al., 2024) and PromptTTS (Guo et al., 2023) are pioneering text description-based speech synthesis, employing natural language to specify the attributes to be controlled. Subsequent efforts (Lyth and King, 2024; Yamauchi et al., 2024; Ji et al., 2024b; Leng et al., 2024; Jin et al., 2024) are focused on exploring the use of automated methods to capture more diverse natural language descriptions, thereby enabling control over an expanded range of attributes.

Additionally, a speaker's facial image can also serve as a form of control information for speech synthesis (Goto et al., 2020; Lee et al., 2023; Wang et al., 2022; Lee et al., 2024). Specifically, AnyoneNet (Wang et al., 2022) employs face embeddings, projecting them into the same embedding space as reference audio embeddings. This approach aims to generate voices that align with the character's visual appearance, thus facilitating the production of speaker videos that incorporate speech, derived from a single facial image and accompanying text.

Most recently, research has begun to explore control methods beyond single-modality-based methods. MM-TTS (Guan et al., 2024) pioneers a unified framework that accommodates multimodal prompts from text, audio, or facial modalities. Further advancing this field, StyleFusion TTS (Chen et al., 2024b) introduces a multi-prompt framework that leverages both style descriptions and an audio prompt to simultaneously control audio style and timbre. Unlike StyleFusion TTS, which necessitates simultaneous input of both prompts during inference, our proposed FleSpeech accommodates inputs from any number of arbitrary modalities. This flexibility significantly enhances the adaptability and controllability of speech synthesis.

2.2 Speech Attribute Editing

Editing speech attributes typically involves modifications to timbre or speaking styles. The former, known as Voice Conversion (VC), specifically aims to transform the timbre to match that of another target speaker while retaining the linguistic information. A typical method employs pre-trained models to extract speaker timbre representations and speech content features, which are then merged to reconstruct the converted speech (Qian et al., 2019; Wang et al., 2021; Ning et al., 2023). However, this approach often struggles to generalize to unseen speakers due to model capacity constraints when handling large-scale speech data. To address this challenge, language model-based voice conver-



(a) Overall Model Architecture. (b) Architecture of Multimodal Prompt Encoder.

Figure 2: The model architecture of FleSpeech.

sion methods have begun to emerge (Wang et al., 2024a,b).

Instead of changing timbre, style editing focuses on modifying the speech style while preserving linguistic content and timbre. VoxEditor (Sheng et al., 2024) introduces a voice attribute editing model that facilitates the modification of speech style attributes using a given source audio and textual description. Similarly, AudioBox (Vyas et al., 2023) presents a flow-matching-based framework that enables the restyling of any audio sample through text descriptions. Extending beyond just editing timbre or style, our proposed FleSpeech allows for the simultaneous editing of both speaker timbre and style.

3 Method

3.1 Overview

FleSpeech is designed to flexibly control the synthesis of speech either through any single-form prompt or a combination of different prompt formats. For instance, it can control style using a text description while managing timbre with reference audio. To facilitate this, as illustrated in Fig. 2a, FleSpeech comprises a language model module for semantic token prediction and a Flow Matching-based module for acoustic feature prediction. To handle different forms of prompts, a multimodal prompt encoder (MPE) is proposed. Specifically, MPE is designed to handle prompts in any format, i.e., text, audio, or image, to obtain a unified representation. This unified representation serves as a condition in either the language model or the flow matching module, facilitating targeted control.

Here, we first introduce the language model and flow matching, both of which play crucial roles in speech generation and are classified as components of the multimodal prompt-based speech generator. Subsequently, we describe MPE, which is used to control the generator.

3.2 Multimodal Prompt-based Speech Generator

Langauge model for semantic generation Inspired by the outstanding performance of language models in speech synthesis tasks (Wang et al., 2023), we tokenize speech into semantic tokens and then employ a decoder-only transformer-based language model to predict these tokens. Specifically, the input text is first converted into a phoneme sequence. The language model then takes this phoneme embedding sequence, concatenated with the global condition embedding obtained via MPE, to predict semantic tokens in an autoregressive manner. Details about the model parameters are provided in Appendix A.

As for speech tokenization, inspired by VecTok (Zhu et al., 2023), our tokenizer employs WavLM (Chen et al., 2022a), pre-trained on 94k hour dataset[1], to extract speech features. We then use the K-means clustering method to discretize these features into 300 tokens, primarily associated with linguistic information.

Flow matching for acoustic feature generation The absence of acoustic details in semantic tokens results in a gap with the corresponding audio. To bridge this gap, a diffusion transformer based flow-matching-based module, similar to Stable Diffusion 3 (Esser et al., 2024), is used to generate acoustic features from semantic tokens, supplemented by the conditional embedding created by the MPE. Details about this module can be found in Appendix A.

Compared to pre-designed acoustic features such as the Mel-spectrum, Glow-WaveGAN (Cong et al.,

[1]https://huggingface.co/microsoft/wavlm-large

2021) demonstrates that the acoustic latent representation learned by a variational autoencoder performs better in acoustic feature prediction and vocoder-based speech synthesis processes. Therefore, instead of using the Mel-spectrum as the acoustic feature to be predicted by the flow matching module as in CosyVoice (Du et al., 2024), we adopt WaveGAN implemented in Glow-WaveGAN to extract the latent representation as the acoustic feature via the encoder. The decoder is then used as a vocoder to generate the final audio.

3.3 Multimodal Prompt Encoder

The objective of the MPE is to obtain a unified condition embedding based on prompts from multiple modalities. Given that the reference audio contains the most comprehensive information and is always available during the speech generation training process, the core idea behind MPE is to map the representations of textual and visual prompts to the space of reference audio embeddings. To achieve this, following the approach of IP-Adapter (Ye et al., 2023), a query-based encoder structure is employed, which uses some learnable query tokens to extract speech-related information from the representations of different prompts. Additionally, due to the many-to-one relationship between reference audio and other prompt modalities, such as multiple voices that correspond to the textual style description "a male speaking loudly and very fast", a diffusion-based method is adopted to model this diversity. Specifically, as shown in the Fig. 2b, the embeddings from different prompt modalities are input into the query-based encoder separately. These embeddings are then concatenated with the noisy audio embedding x_t and fed into the diffusion process. The diffusion model subsequently predicts the ground truth audio embedding x_0 through denoising.

The reference audio prompt embedding, serving as the anchor for prompt embeddings from different modalities, captures all time-invariant information, such as style and timbre. Consequently, the embedding created by the reference audio encoder can be directly used as the conditional embedding in speech generation. Similar to Meta-StyleSpeech (Min et al., 2021), the reference audio encoder consists of six attention blocks, and the output of the last block is average-pooled to obtain a global audio embedding.

The textual prompt embedding can be derived from either the description of the speaking style or facial visual information. In this case, the description text is embedded using a pre-trained BERT (Devlin et al., 2019)[2], which is to capture the semantic information of the descriptions.

The visual prompt embedding, specifically referring to the embedding of face information, is inspired by ID-Animator (He et al., 2024) and aims to capture both static and dynamic information naturally present in face videos. Static information encompasses the facial features of the speaker in a specific frame, such as gender, age, hair colour, and body type, and is closely related to the acoustic features of the speaker. In contrast, dynamic information reflects the speaker's state and behaviour, such as laughing or chatting. This dynamic information complements the static facial features and helps capture nuances that go beyond the capabilities of static images.

MPE is designed to accept inputs from any modality during both training and inference. Embeddings from non-input modalities are masked prior to the diffusion process. Furthermore, given that different speech attributes are modelled at various stages, the parameters of MPE corresponding to token prediction and acoustic feature generation are not shared.

3.4 Training Strategy

To address the scarcity of multimodal data, we propose a three-stage training strategy. We use two types of data: 50,000 hours of large-scale low-expressivity speech data from LibriHeavy and 616 hours of high-expressivity speech data collected from the open-source dataset.

In the first stage, the model is trained on a combination of two datasets to achieve basic speech synthesis capabilities with the large-scale corpus ensuring stability. In the second stage, the model is fine-tuned on high-expressive data to achieve domain alignment. In the third stage, we freeze the generation model backbone and start training the multimodal encoder to enable the model to support modal inputs other than speech prompts. Notably, during this stage, the multimodal prompt encoder is updated with the generation loss in addition to the diffusion loss. The details of the training objective can be found in Appendix B.

[2]https://huggingface.co/google-bert/bert-large-uncased

4 Multimodal Dataset

Due to the scarcity of multimodal controllable speech synthesis data, we propose a method for constructing such a database. Compared to existing data, the collected data is not only larger in scale but also includes facial modality with richer facial annotation information. Details about the collected data and comparisons with other multimodal speech synthesis datasets can be found in Appendix C.

The collection of the talking head video dataset is based on the CelebV-HQ (Zhu et al., 2022), GRID (Cooke et al., 2006), LRS2 (Chung et al., 2017), and MEAD (Wang et al., 2020) datasets, which primarily feature talking faces with one person speaking most of the time. After web crawling, the videos are segmented according to the timestamps provided in the dataset. To ensure speech quality, we first apply the S-DCCRN (Lv et al., 2022) model to denoise the crawled videos, retaining only those with a Signal-to-Noise Ratio (SNR) test score greater than 0.6 and a DNS-MOS (Reddy et al., 2022) greater than 2.6. Finally, we use Whisper (Radford et al., 2023) [3] to get the speech transcription and filter out sentences with fewer than three words. Additionally, the face descriptions are also created, and the details are introduced in section 4.1.

The collection of the speech dataset is based on a large-scale, high-quality TTS dataset, TextrolSpeech (Ji et al., 2024a), which concludes emotional content and attribute labels such as gender and emotion. Based on this, we re-caption the speaking style according to the distribution of our entire dataset. This re-caption method is detailed in section 4.2

4.1 Face Description

Following ID-Animator (He et al., 2024), we use both static and dynamic face descriptions. First, we crop all face videos based on timestamp and face range coordinates, selecting a random frame as the face image prompt. This image is processed ShareGPT4V (Chen et al., 2025) [4] to generate a static description focused on facial attributes (e.g., gender, age, fatness). To capture the speaker timbre, influenced by facial expressions, we extract video clips and use Video-LLava (Lin et al., 2023) to generate dynamic descriptions focused on facial changes and movements during speech. Finally, we combine both descriptions using a large language model (LLM) [5] to ensure cohesive and high-quality outputs with relevant details and human-like expression.

4.2 Speaking Style Description

To obtain text descriptions of speaking style, we extract gender and emotion labels from the TextrolSpeech and MEAD datasets. For other talking head video datasets, we use a face gender classification model (Serengil and Ozpinar, 2021) [6] to extract gender labels. Acoustic attributes, including pitch, speech rate, and Root Mean Square(RMS) of energy are extracted using the signal processing method. Silent frames are filtered by checking for zero pitch values. In addition, we calculate the mean and variance of pitch to measure the pitch and its fluctuation, and the average RMS to measure the volume.

After feature extraction, we analyze their distribution and apply Mean and One Standard Deviation Splitting to divide each attribute into three intervals: "low," "normal," and "high" intervals. We then use a LLM to generate multiple synonymous words or phrases for each attribute category. Using different prompts, we combine these into single sentences to create various speech style descriptions with the same method. This stage enables the simultaneous generation of multiple speech style descriptions with similar meanings. This method has been shown to provide rich and diverse contextual clues to enhance the effectiveness of zero-shot control.

5 Experiment Setup

5.1 Test Dataset

To comprehensively evaluate the performance and generalization of the proposed model, two groups of datasets are used for testing. One test set is reserved from the collected multimodal data, which includes 20 voice prompts from TextrolSpeech and 20 facial prompts from the talking head video dataset. The other test set is an out-of-domain dataset from the HDTF dataset (Zhang et al., 2021), consisting of image and audio prompts that undergo the same data processing procedures as the training set. Additionally, we selected 16 emotional audio and image prompts from the MEAD dataset to

[3] https://huggingface.co/openai/whisper-large-v3
[4] https://huggingface.co/Lin-Chen/ShareGPT4V-7B
[5] We use ChatGPT (gpt-3.5-turbo) as the LLM.
[6] https://github.com/serengil/deepface

evaluate emotion accuracy. The synthesized transcripts were derived from a random selection of 100 sentences from the multimodal dataset.

5.2 Evaluation Metrics

Objective metrics includes Word Error Rate (WER), Speaker Similarity (SPK-Sim), UTMOS (Saeki et al., 2022) [7], Emotion Accuracy, Gender Accuracy, and other speech attribute accuracy. Details about these objective metrics can be found in Appendix D.1.

Subjective metrics include the Mean Opinion Score (MOS) to evaluate speech naturalness (N-MOS) and similarity (Sim-MOS). Higher N-MOS means better naturalness while higher Sim-MOS indicates better similarity with the specific target. Details about the subjective metrics can be found in Appendix D.2

6 Experimental Results

We evaluated FleSpeech using both single-type prompts and various combinations of prompt types. Additionally, the extended capabilities of FleSpeech, including speech editing and voice conversion, were also assessed. The introduction to the various comparison methods, including MM-TTS (Guan et al., 2024), Salle (Ji et al., 2024a), NaturalSpeech2 (Shen et al., 2024), and PromptTTS2 (Leng et al., 2024) can be found in the Appendix E.

6.1 Single-Prompt Controllable TTS

To evaluate FleSpeech's single-prompt control capabilities, we compared it with other models using text, face image, or audio as the prompt. We also conducted an ablation study to show the effectiveness of FleSpeech's design.

6.1.1 Comparsion with Other Methods

Speech generation with text prompt was conducted using a set of text prompts with various emotional and prosodic attributes. As shown in the *Text* section of Table 1, FleSpeech achieved excellent results in terms of different style attributes and emotional accuracy. Subjective testing results indicate that the speech generated by FleSpeech closely follows the text prompts and exhibits high naturalness.

Speech generation with audio prompt is presented in the *Audio* section of Table 1. Compared to MM-TTS, FleSpeech demonstrates significantly better speaker similarity, primarily due to the model capacity of the large-scale speech synthesis system. Furthermore, FleSpeech outperforms NaturalSpeech2 in terms of emotion accuracy, gender accuracy, and speaker similarity, highlighting that its multi-stage framework is more effective at capturing various attributes, such as style and tone from the audio prompts. With the cascading structure of LM and flow matching, FleSpeech has significantly improved naturalness and audio quality.

Speech generation with face prompt presented in the *Face* section of Table 1 showcases that FleSpeech achieved optimal performance across most metrics except for speaker similarity. This is primarily due to the absence of an explicit objective relationship between speaker timbre and facial features. Instead, the matching is more subjective in nature. Subjective results indicate that the speech generated by FleSpeech has a higher correlation with the facial images, suggesting its ability to capture key information from the face and synthesize matching speaker timbre.

6.1.2 Ablation Study

To evaluate the effectiveness of face captions, an ablation study was conducted, which can be found in the *Face* section of Table 1. We first removed the dynamic attributes of the face description (w/o Face dyn-cap), which resulted in a sharp decline in emotional similarity, indicating a reduced ability of the model to capture emotional information from the face. Moreover, when we eliminated both the static and dynamic attributes of the face description (w/o Face cap), the model relied solely on Clip representations for speaker-timbre-related information. The experimental results show a comprehensive decline in terms of all metrics, demonstrating the effectiveness of combining Clip and facial descriptions. Finally, we replaced Clip with FaceNet (w/ FaceNet emb), a facial recognition model capable of extracting embeddings that represent unique attributes among different individuals for face-driven speech synthesis. The experimental results indicated that FaceNet's ability to capture facial information is insufficient for synthesizing speech corresponding to the face prompt.

We further visualized the speaker embedding similarity matrix between different generated sentences. As shown in Fig. 3, compared to the results with w/ FaceNet emb, Clip (i.e., w/o Face cap) exhibits higher speaker consistency, indicating the

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

Table 1: Experimental results on speech generation based on a single prompt. $^{\diamond}$ means the results are obtained from the authors. † means the reproduced results.

Prompt	Model	Accuracy(%)↑						WER(%)↓	SPK-Sim↑	UTMOS↑	N-MOS↑	Sim-MOS↑
		Emotion	Gender	Speed	Pitch	Fluctuation	Volum					
Text	MM-TTS$^{\diamond}$	58.3	-	-	-	-	-	13.2	-	1.311	3.25 ± 0.08	3.32 ± 0.03
	SaLLE†	22.4	55.2	58.3	53.5	56.8	61.7	27.2	-	1.764	3.02 ± 0.11	3.17 ± 0.09
	PromptTTS2†	63.5	82.6	94.6	90.6	83.2	**95.2**	8.7	-	1.778	3.91 ± 0.08	3.61 ± 0.07
	FleSpeech	**66.7**	**89.3**	**95.1**	**93.3**	**95.5**	92.9	**7.5**	-	**2.351**	$\mathbf{3.95 \pm 0.09}$	$\mathbf{4.05 \pm 0.07}$
Audio	MM-TTS$^{\diamond}$	58.8	79.3	-	-	-	-	12.8	0.553	1.430	3.56 ± 0.12	3.38 ± 0.10
	NaturalSpeech2†	64.4	88.1	-	-	-	-	7.6	0.663	2.602	3.84 ± 0.04	3.52 ± 0.04
	FleSpeech	**66.8**	**89.9**	-	-	-	-	**5.8**	**0.725**	**2.835**	$\mathbf{3.94 \pm 0.04}$	$\mathbf{3.75 \pm 0.06}$
Face	MM-TTS$^{\diamond}$	56.6	70.6	-	-	-	-	17.2	0.572	2.155	3.01 ± 0.04	3.08 ± 0.09
	PromptTTS2†	63.2	72.7	-	-	-	-	11.1	**0.643**	**2.643**	3.73 ± 0.08	3.88 ± 0.05
	FleSpeech	**64.5**	**87.3**	-	-	-	-	**7.0**	0.629	2.457	$\mathbf{3.91 \pm 0.08}$	$\mathbf{3.96 \pm 0.07}$
	w/o Face dyn-cap	64.0	87.1	-	-	-	-	7.1	0.629	2.393	3.82 ± 0.06	3.91 ± 0.03
	w/o Face cap	63.0	83.7	-	-	-	-	7.2	0.631	2.442	3.72 ± 0.06	3.83 ± 0.06
	w/ FaceNet emb	58.5	63.8	-	-	-	-	8.2	0.560	2.524	3.58 ± 0.04	3.25 ± 0.08

Table 2: Experimental results on speech generation based on multiple prompts.

Model	Text2Token	Token2Latent	Accuracy(%)↑						WER(%)↓	SPK-Sim↑	UTMOS↑
			Emotion	Gender	Speed	Pitch	Fluctuation	Volum			
FleSpeech	Text	Audio	66.1	85.4	95.8	92.0	95.3	94.0	7.0	0.706	2.557
FleSpeech	Text	Face	64.9	86.3	95.2	93.7	94.9	96.4	7.2	0.610	2.598
FleSpeech	Audio	Audio	62.7	85.8	-	-	-	-	5.9	0.702	3.008
FleSpeech	Audio	Face	63.3	86.1	-	-	-	-	6.1	0.603	2.760
w/o Face cap	Audio	Face	63.1	81.3	-	-	-	-	6.5	0.610	2.667

effectiveness of the Clip encoder in extracting implicit representations. By gradually adding static or dynamic face captions, the colors outside the diagonal gradually deepen, indicating a stronger binding between facial images and speaker timbre. FleSpeech demonstrates the highest speaker consistency, highlighting the effectiveness of combining Clip with dynamic and static captions.

6.1.3 Overall Analysis

In addition to individual tasks, we conducted an overall analysis of the different experimental results. The comparative results in various sections of table 1 indicate that the audio modality achieves the highest accuracy in terms of emotion and gender, followed by text. This suggests that audio provides the most fine-grained information, and through the text prompt encoder, the model can effectively extract relevant speech attributes from textual descriptions. Image prompts, on the other hand, are generally less discernible, leading to a decrease in accuracy. Moreover, the WER and UTMOS of speech generated from text prompts show a significant decline, which may be attributed to the one-to-many problem, especially in the text modality, where a larger sample space results in poorer stability. Finally, despite being trained on a small-scale dataset, we observed that MM-TTS using face prompt outperforms the audio prompt in terms of SPK-Simi and UTMOS. This reflects the generalization advantage of the face prompt, considering the complex acoustic environments present in the audio prompt.

6.2 Multi-Prompt Controllable TTS

To evaluate the unique flexible control capability of FleSpeech, we assessed its performance using multiple prompts. Specifically, we provide different prompts at various stages to control different speech aspects. We examined four combinations of prompts. To validate the effectiveness of each stage, we included emotional or neutral prompts in the first stage and only neutral prompts in the second stage. As shown in Table 2, compared to using a single prompt for control, FleSpeech effectively controls style and emotional attributes while reproducing the timbre of the target speaker despite some performance loss.

Additionally, we removed the facial caption(w/o Face cap) in the combination of audio and prompts. We observed a significant decrease in gender accuracy, which indicates that fine-grained information provided by the audio prompt affects speaker timbre modeling in the second stage. The experimental results demonstrate that incorporating the face caption can alleviate the impact of audio prompts, leading to higher consistency with the face prompt.

Furthermore, by comparing the results of different tasks, we found that the WER and UTMOS

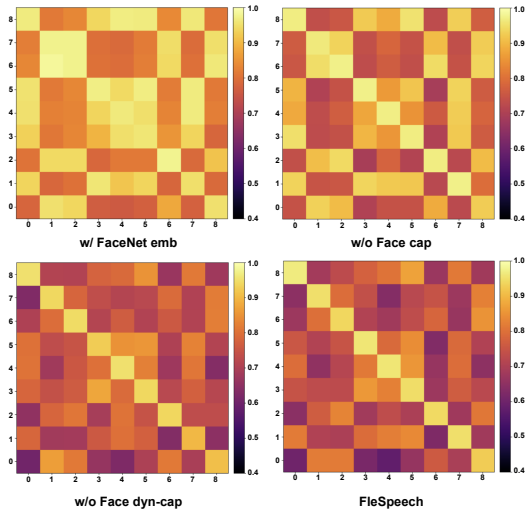


Figure 3: Cosine similarity matrix of speaker embeddings between face-prompt-based synthesized speech and ground-truth speech. The horizontal axis represents different synthesized speech, while the vertical axis represents ground-truth speech. The diagonal indicates that the image prompt and ground-truth speech are from the same speaker. Lighter colors indicate higher similarity.

are highest for the model using two audio prompts, while models using text as the first-stage prompt have the lowest values. This further indicates a negative correlation between the diversity and stability of the speech attribute space. Moreover, models using text as the first-stage prompt generally achieve higher SPK-Sim compared to those using audio modality. This suggests that more fine-grained information in the first stage can influence the speaker timbre modeling in the second stage.

6.3 Extensibility

In addition to speech synthesis, we conducted additional experiments on other tasks to evaluate the scalability of FleSpeech.

6.3.1 Speaking Style Editing

Speaking style editing refers to modifying speech attributes without altering the content or speaker timbre. To edit the attribute of a given utterance based on the text description, the transcription of this utterance obtained via Whisper and the text description can be used as the input for the language model. Then this utterance can work as the audio prompt for the second stage. We compared our method with Audiobox (Vyas et al., 2023), a unified audio generation model based on flow matching that can redesign the provided audio examples using natural language instructions. As shown in Table 3, FleSpeech achieves satisfactory results. Regarding emotional expression, FleSpeech scores lower, primarily because Audiobox incorporates non-verbal sounds, such as laughter, which enhance emotional perception.

Table 3: Experimental results in speaking style editing.

Model	Accuracy(%)↑					WER(%)↓	SPK-Sim↑
	Emotion	Speed	Pitch	Fluctuation	Volum		
Audiobox	**66.3**	83.3	**98.3**	83.3	83.3	8.4	0.712
FleSpeech	63.6	**91.6**	**98.3**	**91.6**	**91.6**	**7.2**	**0.745**

6.3.2 Voice Conversion

FleSpeech allows for the speaker timbre editing by facial caption when given a facial image and its corresponding caption. For instance, it can explicitly specify attributes such as the speaker's age, race, and fatness, which have been previously proved to be associated with speaker timbre (Stathopoulos et al., 2011; Souza and Santos, 2018; Yang et al., 2022). We evaluate the effectiveness of these edits through accuracy testing and subjective preference assessments. The MOS indicates the degree of match, with higher scores reflecting better alignment. Preference indicates perceived accuracy, where participants choose which audio, before or after editing, better matches the edited facial caption. The test results are shown in Table 4, where FleSpeech achieves an editing accuracy exceeding 70%, demonstrating its capability to effectively edit speaker-timbre-related attributes to match facial features. The subjective scores further corroborate this conclusion. Additionally, the accuracy for age is higher than for BMI, suggesting that age is more perceptible in facial images.

Table 4: Experimental results in voice conversion.

Characteristic	Acc(%)↑	MOS↑	Preference(%)↑
BMI	72.6	3.75 ± 0.04	62.4
Age	81.0	3.87 ± 0.08	74.1
Race	75.3	3.83 ± 0.06	66.5

7 Conclusion

In this work, we propose a flexible and controllable speech generation framework called FleSpeech. Specifically, we implement a two-stage speech generation framework composed of a language model and a flow matching module, allowing for flexible control by providing different prompts at various stages. Additionally, we introduce a multimodal prompt encoder that can accept prompts from different modalities and embed them into a unified style space, enabling more adaptable prompting. Comprehensive subjective and objective experiments demonstrate the effectiveness of FleSpeech.

Limitation

Although our approach successfully achieves flexible control over speech attributes, it is important to acknowledge its limitations. First, the information extracted from face images is limited. Many unexplored aspects, such as accent, are related to speaking style and restrict the matching accuracy between face and speech. Second, the relatively small scale of our collected dataset limits the control over additional attributes, such as background sound. Despite these limitations, our FleSpeech has taken an important step toward a more flexible and controllable speech generation system.

References

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

James Betker. 2023. Better speech synthesis through scaling. *CoRR*, abs/2305.07243.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2025. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, and Tao Qin. 2020. Multispeech: Multi-speaker text to speech with transformer. In *INTERSPEECH*, pages 4024–4028.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022a. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024a. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *CoRR*, abs/2410.06885.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. 2022b. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151. IEEE.

Zhiyong Chen, Xinnuo Li, Zhiqi Ai, and Shugong Xu. 2024b. Stylefusion TTS: multimodal style-control and enhanced feature fusion for zero-shot text-to-speech synthesis. In *PRCV (11)*, volume 15041 of *Lecture Notes in Computer Science*, pages 263–277.

Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3444–3453.

Jian Cong, Shan Yang, Lei Xie, and Dan Su. 2021. Glow-wavegan: Learning speech representations from gan-based variational auto-encoder for high fidelity flow-based speech synthesis. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 2182–2186.

Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. 2024. E2 TTS: embarrassingly easy fully non-autoregressive zero-shot TTS. *CoRR*, abs/2406.18009.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Andrew Gibiansky, Sercan Ömer Arik, Gregory Frederick Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice

2: Multi-speaker neural text-to-speech. In *NIPS*, pages 2962–2970.

Shunsuke Goto, Kotaro Onishi, Yuki Saito, Kentaro Tachibana, and Koichiro Mori. 2020. Face2speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image. In *INTERSPEECH*, pages 1321–1325.

Wenhao Guan, Yishuang Li, Tao Li, Hukai Huang, Feng Wang, Jiayan Lin, Lingyan Huang, Lin Li, and Qingyang Hong. 2024. MM-TTS: multi-modal prompt based style transfer for expressive text-to-speech synthesis. In *AAAI*, pages 18117–18125.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, Man Zhou, and Jie Zhang. 2024. Id-animator: Zero-shot identity-preserving human video generation. *CoRR*, abs/2404.15275.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024a. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP*, pages 10301–10305.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024b. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305. IEEE.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. 2024. Speechcraft: A fine-grained expressive speech dataset with natural language description. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1255–1264.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: A 50, 000 hours ASR corpus with punctuation casing and context. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 10991–10995.

Jaehyeon Kim, Keon Lee, Seungjun Chung, and Jaewoong Cho. 2024. Clam-tts: Improving neural codec language model for zero-shot text-to-speech. In *ICLR*.

Mateusz Lajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, Alexis Moinet, Sri Karlapati, Ewa Muszynska, Haohan Guo, Bartosz Putrycz, Soledad López Gambino, Kayeon Yoo, Elena Sokolova, and Thomas Drugman. 2024. BASE TTS: lessons from building a billion-parameter text-to-speech model on 100k hours of data. *CoRR*, abs/2402.08093.

Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. 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)*, pages 1–5. IEEE.

Minyoung Lee, Eunil Park, and Sungeun Hong. 2024. Fvtts : Face based voice synthesis for text-to-speech. In *Interspeech 2024*, pages 4953–4957.

Younggun Lee, Azam Rabiee, and Soo-Young Lee. 2017. Emotional end-to-end neural speech synthesizer. *CoRR*, abs/1711.05447.

Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie. 2022. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:853–864.

Yichong Leng, Zhifang Guo, Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, Lei He, Xiangyang Li, Sheng Zhao, Tao Qin, and Jiang Bian. 2024. Prompttts 2: Describing and generating voices with text prompt. In *ICLR*.

Tao Li, Xinsheng Wang, Qicong Xie, Zhichao Wang, and Lei Xie. 2022. Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1448–1460.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Guanghou Liu, Yongmao Zhang, Yi Lei, Yunlin Chen, Rui Wang, Lei Xie, and Zhifei Li. 2023. Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions. In *INTERSPEECH*, pages 4888–4892.

Jaime Lorenzo-Trueba, Gustav Eje Henter, Shinji Takaki, Junichi Yamagishi, Yosuke Morino, and Yuta Ochiai. 2018. Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis. *Speech Commun.*, 99:135–143.

Shubo Lv, Yihui Fu, Mengtao Xing, Jiayao Sun, Lei Xie, Jun Huang, Yannan Wang, and Tao Yu. 2022. S-dccrn: Super wide band dccrn with learnable complex feature for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7767–7771. IEEE.

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.

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.

Ziqian Ning, Qicong Xie, Pengcheng Zhu, Zhichao Wang, Liumeng Xue, Jixun Yao, Lei Xie, and Mengxiao Bi. 2023. Expressive-vc: Highly expressive voice conversion with attention fusion of bottleneck and perturbation features. In *ICASSP*, pages 1–5.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 5210–5219.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2022. Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Sefik Ilkin Serengil and Alper Ozpinar. 2021. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2024. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Zhengyan Sheng, Yang Ai, Li-Juan Liu, Jia Pan, and Zhen-Hua Ling. 2024. Voice attribute editing with text prompt. *CoRR*, abs/2404.08857.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. In *ICLR*.

Lourdes Bernadete Rocha de Souza and Marquiony Marques dos Santos. 2018. Body mass index and acoustic voice parameters: is there a relationship? *Brazilian journal of otorhinolaryngology*, 84(4):410–415.

Elaine T Stathopoulos, Jessica E Huber, and Joan E Sussman. 2011. Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age.

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

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. 2023. Audiobox: Unified audio generation with natural language prompts. *CoRR*, abs/2312.15821.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111.

Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. 2021. VQMIVC: vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. In *Interspeech*, pages 1344–1348.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*, volume 12366 of *Lecture Notes in Computer Science*, pages 700–717.

Xinsheng Wang, Qicong Xie, Jihua Zhu, Lei Xie, and Odette Scharenborg. 2022. Anyonenet: Synchronized speech and talking head generation for arbitrary persons. *IEEE Transactions on Multimedia*, 25:6717–6728.

Zhichao Wang, Yuanzhe Chen, Xinsheng Wang, Zhuo Chen, Lei Xie, Yuping Wang, and Yuxuan Wang. 2024a. Streamvoice: Streamable context-aware language modeling for real-time zero-shot voice conversion. *arXiv preprint arXiv:2401.11053*.

Zhichao Wang, Yuanzhe Chen, Xinsheng Wang, Lei Xie, and Yuping Wang. 2024b. Streamvoice+: Evolving

into end-to-end streaming zero-shot voice conversion. *IEEE Signal Processing Letters*.

Tianxin Xie, Yan Rong, Pengfei Zhang, and Li Liu. 2024. Towards controllable speech synthesis in the era of large language models: A survey.

Kazuki Yamauchi, Yusuke Ijima, and Yuki Saito. 2024. Stylecap: Automatic speaking-style captioning from speech based on speech and language self-supervised learning models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11261–11265. IEEE.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. 2024. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zhihan Yang, Zhiyong Wu, and Jia Jia. 2022. Speaker characteristics guided speech synthesis. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *CoRR*, abs/2308.06721.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3661–3670.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. 2022. Celebv-hq: A large-scale video facial attributes dataset. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII*, volume 13667 of *Lecture Notes in Computer Science*, pages 650–667.

Xinfa Zhu, Yuanjun Lv, Yi Lei, Tao Li, Wendi He, Hongbin Zhou, Heng Lu, and Lei Xie. 2023. Vec-tok speech: speech vectorization and tokenization for neural speech generation. *CoRR*, abs/2310.07246.

Xinfa Zhu, Wenjie Tian, Xinsheng Wang, Lei He, Yujia Xiao, Xi Wang, Xu Tan, Sheng Zhao, and Lei Xie. 2024. Unistyle: Unified style modeling for speaking style captioning and stylistic speech synthesis. In *ACM Multimedia*, pages 7513–7522.

A Model Configurations

The language model for semantic prediction adopts the LLaMA architecture with 16 layers and 16 attention heads. The hidden size and intermediate size are 1024 and 4096, respectively. The flow matching for the acoustic feature prediction is based on the DiT architecture with 12 layers, 12 attention heads, and a hidden dimension of 768. For MPE, the number of queries in QueryMLP is set to 16, with 6 layers, 6 attention heads, and an intermediate size of 256. The reference audio encoder consists of 6 attention blocks with a hidden size of 512. During both the training and inference stages, the length of the audio prompt is fixed at 6 seconds.

The diffusion model consists of a diffusion process and a denoising process. For the diffusion process, given the audio embedding x_0, the forward diffusion process transforms it into Gaussian noise under the noise schedule β as follows:

$$\mathrm{d}x_t = -\frac{1}{2}\beta_t x_t \mathrm{t} + \sqrt{\beta_t}\mathrm{d}\omega_t, t \in [0,1] \quad (1)$$

For the denoising process, the denoising process aims to transform the noisy representation x_t to the audio embedding x_0 by the following formulation (Song et al., 2021).

$$\mathrm{d}x_t = -\frac{1}{2}\left(x_t + \nabla \log p_t\left(x_t\right)\right)\beta_t \,\mathrm{d}t, \quad t \in [0,1] \quad (2)$$

The diffusion module is trained to estimate the gradients of log-density of noisy data ($\bigtriangledown log p_t(z_t)$) by predicting the origin audio embedding x_0, conditioned on the embeddings from different prompt modalities, noised audio embedding, and diffusion step t that indicates the degree of noise in the diffusion model.

Both language model and flowing matching module are trained on 8 NVIDIA TESLA V100 GPUs (32GB each) with a batch size of 2 per GPU and a gradient accumulation step of 50. The two models are first trained 600k steps on the LibriHeavy (Kang et al., 2024) dataset which is a 50,000 hours ASR corpus, followed by an additional 300k steps on a collected multimodal dataset. We optimize the models using the AdamW optimizer, warming up the initial learning rate from 1×10^{-7} over the first 5k updates to a peak of 3×10^{-4}, and subsequently applying cosine decay.

B Training Objective

In the first stage, the language model performs the next token prediction task and is optimized using the cross-entropy loss. Meanwhile, flow matching reconstructs the hidden layer features and is optimized with L_2 loss. In the second stage, the MPE is optimized using the L_1 loss calculated between the output embedding $Pred\ x_0$ derived from different prompt modalities and ground-truth embedding

Table 5: Comparison between public datasets for controllable speech generation. Rec means recording, You means youtube, Pod means podcast

Dataseet	Duration	Clips	Modality	Audio Source	Description Form
FSNR0	26h	19k	Speech	Internal dataset	Style tag
TextrolSpeech	330h	236k	Speech	Recording, Emotional dataset	LLM template
MEAD-TTS	36h	31k	Speech, Facial image	Recording	LLM template, Face image
Collected data	616h	449k	Speech, Facial image	Rec, You, Pod, Emotional dataset	LLM template, Face image, Face caption

x_0 obtained from the audio modality, in addition to the loss function of the first stage.

To achieve flexible control, the MPE applies masking to the received prompts of different modalities. Specifically, the audio modality prompt, serving as the target for the MPE, remains consistently present. For data containing both text and audio modality prompts, the MPE maps the text prompt embeddings to the audio embeddings without any masking. In the case of data that includes prompts from all three modalities, there is a one-third probability of masking the text style description, a one-third probability of masking the facial image and facial description, and a one-third probability of not applying any masking. This strategy enables the model to accept various combinations of prompts as input during the inference stage.

C Details of Collected Data

As shown in Table 5, previous work has attempted to construct public datasets for controllable speech generation, but these datasets either have limited size or lack multimodal prompts. In view of this, we constructed a multimodal dataset collection pipeline. Through this pipeline, we collected a multimodal TTS dataset consisting of a 285.9-hour talking head video dataset and a 330-hour speech dataset, totaling approximately 615.9 hours.

D Evalution Metrics

D.1 Objective Metrics

WER is a commonly used metric to assess the intelligibility of generated speech. It is typically calculated by comparing the transcribed text obtained from an Automatic Speech Recognition (ASR) system with the reference text. A lower WER indicates higher intelligibility of the speech. Here, the WER is calculated based on the Whisper (Radford et al., 2023) [8] model.

SPK-Sim is used to evaluate the similarity between the generated audio and the reference audio in terms of speaker characteristics. A higher SPK-Sim value indicates greater similarity between the synthesized speech and the reference audio in terms of the speaker's identity. Here, we use WavLM-large (Chen et al., 2022b)fine-tuned on the speaker verification task, to obtain speaker embeddings. These embeddings are then used to calculate the cosine similarity between the speech samples of each test utterance and the reference clips.

Emotion Accuracy is used to measure the model's ability to control emotions. A higher emotion accuracy indicates a stronger ability of the model to control emotions. Here, emotion2vec+seed (Ma et al., 2023) [9] is adopted to predict the emotion of the synthesized audio and compare it with the given emotion type.

Gender Accuracy is used to measure the model's ability to control gender. A higher gender accuracy means a better gender control ability. Here, an internal ECAPA-TDNN (Desplanques et al., 2020) model fine-tuned on the gender classification task is adopted.

For the accuracy of other speech attributes, we utilize the previously mentioned pipeline for style label annotation to extract attribute values and compare their relative magnitudes across different labels. For example, the speech rate associated with the "fast speaking rate" label exceeds that of the "slow speaking rate" label. For face attribute evaluation, we extract speaker embeddings from MPE and use a face classifier [10] to predict Body Mass Index (BMI). Additionally, we apply the DeepFace (Serengil and Ozpinar, 2021) model to determine gender, race, and age. We then train an MLP-based predictor to infer facial attributes from the speaker embeddings, comparing the predicted attributes against the provided facial descriptions to compute the accuracy.

[8] https://huggingface.co/openai/whisper-large-v3

[9] https://huggingface.co/emotion2vec/emotion2vec_plus_seed

[10] https://github.com/lsimmons2/bmi-project

D.2 Subjective Metrics

In the subjective evaluation, each sample was rated on a scale from 1 to 5, with increments of 0.5 based on its similarity to the reference utterance, where a score of 1 indicates "very bad" and a score of 5 signifies "excellent." Both Normalized Mean Opinion Score (N-MOS) and Similarity Mean Opinion Score (Sim-MOS) are reported with a 95% confidence interval. We selected 50 speech samples for each test, which were listened to by 20 listeners for subjective evaluations.

To clarify, Sim-MOS here varied across different tasks, focusing on aspects such as speech style matching with a text prompt, speaker similarity with an audio prompt, and voice-face matching with a facial prompt.

E Comparison models

To evaluate the performance of FleSpeech, we implemented the following system.

- **MM-TTS** (Guan et al., 2024): A FastSpeech2-based multimodal controllable speech synthesis framework that integrates multimodal inputs into a unified representation space. It supports text descriptions, face images, or speech as prompts. Note that text descriptions in this model are limited to describing the speaker's emotions.
- **Salle** (Ji et al., 2024a): A VALL-E-based text-prompt-driven controllable speech synthesis framework, where text descriptions are concatenated with synthesized phonemes as style prompts.
- **NaturalSpeech2** (Shen et al., 2024): A TTS system with latent diffusion models to enable zero-shot speech synthesis.
- **PromptTTS2** (Leng et al., 2024): A NaturalSpeech2-based speech synthesis framework capable of generating speech that aligns with text style descriptions. We extend its function to support the face prompt just as described in the PromptTTS2 appendix. The CLIP model extracts embedding from the face image, which is then fed into the TTS model.
- **FleSpeech** (proposed): Our proposed framework, which adopts a multi-stage training framework and follows a multi-stage training strategy.

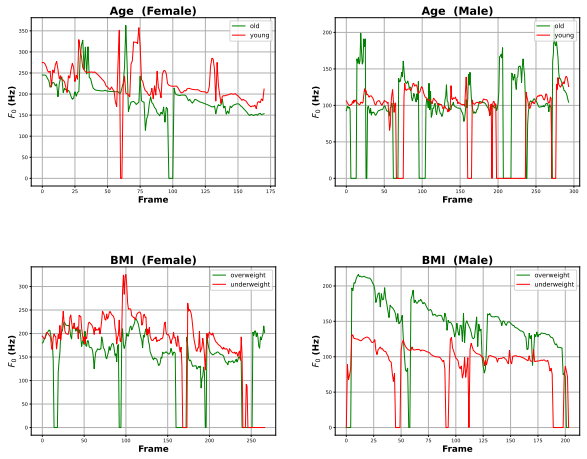


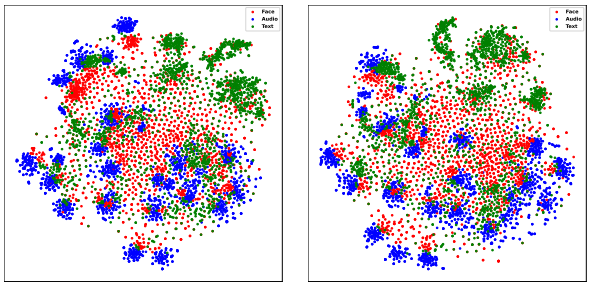
Figure 4: Fundamental Frequency (F0) curve of the speech at different ages and BMI levels groups by gender

F Visualizing the Relationship between Facial Attributes and Voice

To further validate that FleSpeech can establish associations between facial attributes and voice characteristics, we extracted the fundamental frequency (F0) from synthesized speech prompted by different BMI and age groups. As shown in the upper two panels of Fig. 4, the F0 of older women decreases as age increases. In contrast, for elderly males, despite vocal cord atrophy, the F0 tends to increase. The lower two panels of Fig. 4 reveal that being overweight tends to cause articulation difficulties, leading to a decrease in F0 for females, whereas males experience an increase in F0. These findings align with conclusions from prior research, indicating that the proposed FleSpeech can capture variations in speech characteristics across different ages and BMI levels.

G Visualization of MPE Embedding Space

We design MPE to encode prompts from different modalities into a unified space. To validate this, we utilized the MPE to extract embeddings corresponding to each single-modality prompt. Considering that the MPEs in the language model and flow matching do not share parameters, we conducted analyses on both. The test set comprised 2000 randomly selected sentences containing prompts from all three modalities, including 20 speakers with 200 sentences each. The MPE outputs are projected to 2D by t-SNE (Van der Maaten and Hinton, 2008). Each color represents a modality.



(a) MPE in language model (b) MPE in flow matching

Figure 5: TSNE visualization of MPE output embedding clustering.

As illustrated in Fig. 5, both MPEs exhibited similar trends: embeddings mapped by the MPE from different modalities reside within the same embedding space and are not partitioned into multiple subspaces where partitioning into subspaces would imply that each modality is encoded separately, failing to capture the intermodal relationships. Furthermore, the embeddings from audio prompts demonstrated stronger clustering, indicating that audio prompts are more directional than text and facial prompts. In contrast, text and facial prompts exhibit a one-to-many relationship with voice attributes, showing more significant variability.