UniSpeaker: A Unified Approach for Multimodality-driven Speaker Generation

Anonymous ACL submission

Abstract

While recent advances in speaker cloning based on reference speech have significantly improved the authenticity of synthetic speech, speaker generation driven by multimodal cues such as visual appearance, textual descriptions, and other biometric signals remains in its early stages. To pioneer truly multimodal-controllable speaker generation, we propose UniSpeaker, the first framework supporting unified voice control from arbitrary modality combinations. Specifically, self-distillation is firstly applied to a large-scale speech generation model for speaker disentanglement. To overcome data sparsity and one-to-many mapping challenges, a novel KV-Former based unified voice aggregator is introduced, where multiple modalities are projected into a shared latent space through soft contrastive learning, ensuring accurate alignment with user-specified vocal characteristics. Additionally, to advance the field, the first Multimodal Voice Control (MVC) benchmark is established to evaluate voice suitability, diversity, and quality. When tested across five MVC tasks, UniSpeaker is shown to surpass existing modality-specific models. Speech samples and the MVC benchmark are available at https://UniSpeaker.github.io.

1 Introduction

In recent years, the field of speech generation has seen remarkable progress (Wang et al., 2023a; Du et al., 2024), enabling the generated speech to closely resemble the actual recordings. However, traditional zero-shot speech synthesis still faces limitations, particularly in scenarios like virtual character voiceovers, where ideal reference speech may be unavailable(Guo et al., 2023). Therefore, the focus of voice control in generative models need to shift from speaker cloning to speaker generation. Unlike cloning voices from reference samples, leveraging more accessible modalities such as text prompts or face images offers great potential for generating desired voice characteristics. Thus, several studies (Shimizu et al., 2024; Zhang et al., 2023; Lu et al., 2021; Sheng et al., 2023) have explored speaker generation based on other modalities like text or face images, aligning these representations with speaker embeddings to control voice features. Moreover, VoxEditor (Sheng et al., 2024) introduces relative voice descriptions, enabling finer control over voice attributes .

Task	Model	Voice Description Modality			
		Speech	Text	Face	Attribute
TTS	PromptSpeaker	✓	✓		
	Promptttts++	✓	✓		
	Imaginary Voice			✓	
	Synthe-sees	✓		✓	
	3D-face	✓		✓	
SSC	PromptVC	✓	✓		
	HYbridVC	✓	✓		
	FaceVC	✓		✓	
	FVMVC	✓		✓	
	HearFace	✓		✓	
	VoxEditor	✓			✓
TTS&SSC	UniSpeaker	✓	✓	✓	✓

Table 1: Comparison between UniSpeaker and previous studies on multimodality-driven voice control tasks. TTS stands for text-to-speech synthesis. SSC stands for speech-to-speech conversion, preserving both the content and prosody of source speech.

Despite above advances, existing methods often handle voice description modalities in isolation, typically involving only one extra modality to align with the reference speech, as shown in Table 1.This leads to two shortcomings:

Data Sparsity Face-voice alignment data often requires intensive processing from sources like recorded talks. Textual descriptions of timbre rely heavily on human subjective perception (Wallmark, 2019) and thus require manual annotation. As a result, multimodal-aligned data is significantly scarcer than pure speech data, leading to a sparse timbre space and restricted voice diversity.

One-to-Many Mapping Due to cross-modal information gaps, it is challenging to fully reconstruct true voice characteristics from other modalities alone. Consequently, a single face or text description may correspond to multiple plausible timbres during inference, requiring further voice attributes editing (Sheng et al., 2024) to meet the desired requirements. However, current speaker generation models struggle with one-to-many ambiguity in absolute voice descriptions and lack effective coordination among diverse modalities like speech, text, face, and attributes.

To overcome these challenges, we propose UniSpeaker, which can first utilize absolute voice descriptions to generate a coarse speaker timbre, and then enables iterative refinement through relative attribute editing to better align with user expectations. Furthermore, UniSpeaker introduces a coordinated fusion of multimodal inputs, facilitating consistent and controllable speaker generation.

Specifically, we propose a Unified Multimodal Voice Aggregator (MFA) that aligns multimodal inputs into a unified voice space. The core of MVA is the KV-Former, which selectively integrates relevant details from each modality, capturing mutual information between modalities to form a coherent representation. The output is then fed to a generative model for voice control and speaker embedding alignment. To handle the correlation between voice characteristics of different speakers, soft contrastive learning (SoftCL) is proposed during training, which loosens strict one-to-one constraints and leverages intra-modal discriminative information. Inspired by ImageBind (Girdhar et al., 2023), our speech-anchoring mechanism enables cross-modal alignment without requiring parallel data, addressing data scarcity and ensuring diverse voice characteristics. In addition, while large-scale speech generation models excel in voice control, scalable multimodal integration remains underexplored. To address this, we apply self-distillation techniques (Anastassiou et al., 2024) on CosyVoice (Du et al., 2024) to improve voice disentanglement and ensure the versatility across diverse multimodal inputs.

Due to the lack of publicly accessible benchmarks, a multimodal voice control (MVC) benchmark was developed, evaluating speech on voice suitability, voice diversity, and speech quality across 5 tasks for diverse multimodal voice characterics. Using this benchmark, UniSpeaker was assessed and showed superior performance over previous modality-specific models across all tasks.

2 Relate Work

2.1 Multimodality-driven speech generation

Modeling diverse voice characteristics is a key challenge in speech synthesis. Recent works like PromptTTS2 (Leng et al., 2024), Audiobox (Vyas et al., 2023), InstructSpeech(Huang et al., 2024) and others (Guan et al., 2024; Yang et al., 2024), have explored text-based control of speech style or emotion, while only a few studies have specifically focused on text-driven voice control (Shimizu et al., 2024; Zhang et al., 2023). Text prompt-based style control TTS methods typically convert speech attributes like pitch, energy, duration, and emotion into natural prompts via LLMs. Since these style prompts primarily reflect prosody and capture minimal speaker individuality, achieving the desired voice control remains challenging.

In multimodal voice control, researchers have attempted to align different voice description modalities with speaker embeddings using models such as memory networks (Sheng et al., 2023), mixture density networks (Shimizu et al., 2024), and latent diffusion (Yao et al., 2024), as well as loss functions like MSE loss (Lu et al., 2021), cosine similarity loss (Zhang et al., 2023), and perceptual loss (Weng et al., 2023). However, these methods rely on parallel datasets and are difficult to extend to new modalities. In terms of performance, previous face-based methods (Lee et al., 2023) often struggle with mismatched voice characteristics, such as generating a youthful voice for an elderly face. Additionally, VoxEditor (Sheng et al., 2024) is limited to performing voice attribute editing on existing source speech, restricting voice diversity.

2.2 Large speech generation models

Recent advances in speech generation have improved naturalness and robustness, with a focus on voice diversity through novel models, modeling objectives and larger datasets. When integrating multimodal voice descriptions, it is crucial to preserve the performance of pre-trained speech generation models in terms of naturalness, robustness, and prosody. Large-scale models (Wang et al., 2023a) typically use a neural codec to convert speech waveforms into discrete acoustic tokens, which are then processed by an autoregressive language model. However, these discrete token sequences entangle content, speaker, and prosody, making it challenging to align multimodal voice characteristics without disrupting speech content and prosody.

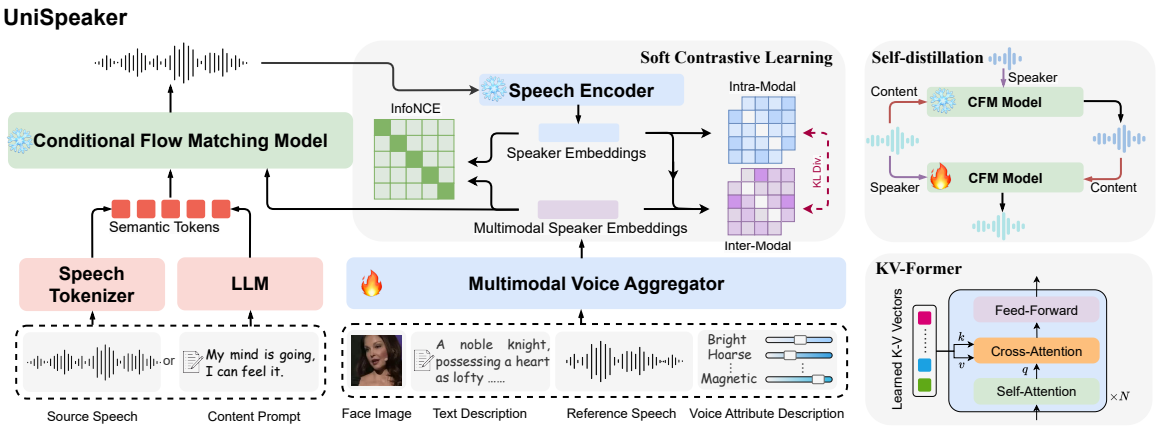

Figure 1: The overview of UniSpeaker. Multimodal speaker embeddings control the voice characteristics of generated speech and are derived from various voice description modalities, with semantic tokens generated from speech or text for speech-to-speech and text-to-speech respectively.

CosyVoice (Du et al., 2024) addresses this by using semantic tokens (Radford et al., 2023) with a conditional flow matching model (CFM). These tokens capture content and prosody, leaving speaker information easier to disentangle. This makes CosyVoice an ideal backbone for UniSpeaker.

3 Methods

In this section, we first review the basic speech generation backbone, then introduce how multimodal voice descriptions are integrated into a pre-trained large speech generation model.

3.1 Basic Speech Generation Framework

The basic speech synthesis system employs two core components frozen from Cosyvoice, respectively for text-to-token generation and token-to-speech synthesis.

Large Language Model (LLM) Given speech-text pairs $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}$, the LLM processes the sequence $\{\mathbf{s}, \mathbf{Y}, \mathbf{C}\}$, where $\mathbf{s}$ represents the speaker embedding extracted from the speech$\mathbf{x}$, $\mathbf{Y}$ denotes the text embedding of the transcript $\mathbf{y}$, and $\mathbf{C}$ comprises the semantic tokens derived from $\mathbf{x}$. The LLM is trained autoregressively to predict $\mathbf{C}$.

Conditional Flow Matching (CFM) To construct a probability density path from a prior distribution to $p_0(\mathbf{X})$ to distribution of the Mel-spectrograms $q(\mathbf{X})$, the model is trained using optimal-transport conditional flow matching (OT-CFM) (Tong et al., 2023) as follows,

$$\mathcal{L}_{\text{OT-CFM}} = \mathbb{E}_{t,p_0(\mathbf{X}_0),q(\mathbf{X}_1)} \|\omega_t(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)|\mathbf{X}_1) - \nu_t(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)|\theta_{CFM})\|_1, \tag{1}$$

with the flow trajectory and vector field defined as:

$$\begin{aligned} \phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1) &= (1-t)\mathbf{X}_0 + t\mathbf{X}_1, \\ \omega_t(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)|\mathbf{X}_1) &= \mathbf{X}_1 - \mathbf{X}_0. \end{aligned} \tag{2}$$

The speaker embeddings $\mathbf{s}$, speech tokens $\mathbf{C}$, and masked Mel-spectrogram $\tilde{\mathbf{X}}_1$ are fed into the CFM to match the vector field with learnable θ_{CFM},

$$\nu_t\left(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)\middle|\theta_{CFM}\right) = \text{NN}\left(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1), t; \mathbf{s}, \mathbf{C}, \tilde{\mathbf{X}}_1\right). \tag{3}$$

3.2 Multimodal Voice Description Integration

We incorporate multiple modalities into the CFM model, allowing various inputs to control the voice characteristics of generated speech. As shown in Figure 2 , each modality is first processed by a pre-trained, modality-specific encoder to obtain the corresponding representation. Each kind of representation is then transformed into a latent vector via adaptive average pooling or a multi-layer perceptron. Those vectors across modalities are mapped into a unified voice space through a shared MVA, producing the corresponding speaker embeddings. These speaker embeddings are then fed into the CFM for speech generation.

Modality-Specific Encoders UniSpeaker integrates three modality-specific encoders for multimodal processing. Facial images are encoded

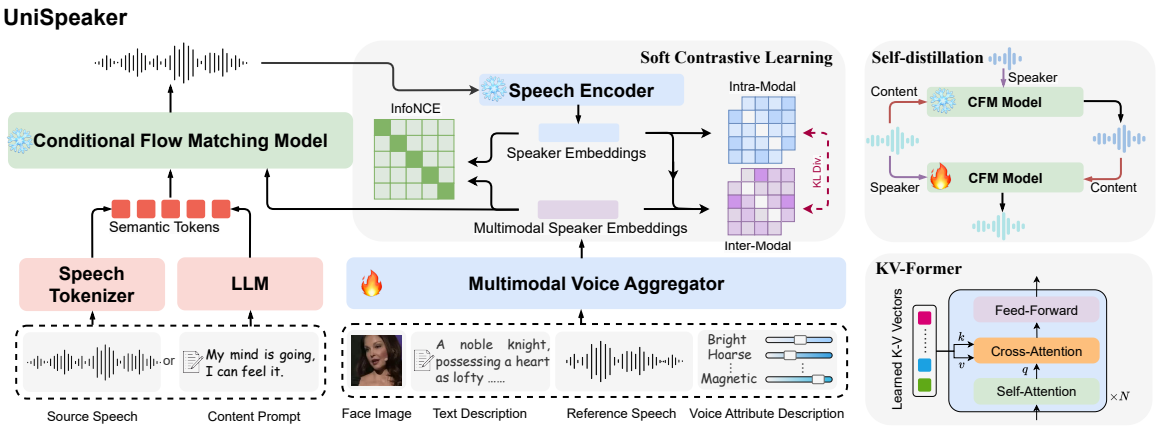

Figure 2: The overview of UniSpeaker. Multimodal speaker embeddings control the voice characteristics of generated speech and are derived from various voice description modalities, with semantic tokens generated from speech or text for speech-to-speech and text-to-speech respectively.

through MTCNN (Zhang et al., 2016) for detection and FaceNet (Schroff et al., 2015) for generating global representations $\mathbf{s}_f$. Textual descriptions are processed by T5 (Raffel et al., 2020) to yield variable-length embeddings $\mathbf{s}_t$. For reference speech, a pre-trained speaker verification network (Wang et al., 2023b) extract speaker embeddings $\mathbf{s}_r$. Following VoxEditor (Sheng et al., 2024), the system enables voice attribute control through learned interpolation. Given two speech samples (A, B) and text prompt t , their embeddings $\{\mathbf{s}_r^A, \mathbf{s}_r^B, \mathbf{s}_t\}$ are processed to predict density difference $\alpha \in [0, 1]$. The target speaker embedding $\mathbf{s}_r$ is obtained via linear interpolation: $\mathbf{s}_r = (1 - \alpha) \cdot \mathbf{s}_r^B + \alpha \cdot \mathbf{s}_t$, enabling continuous voice modulation through α.

Multimodal Voice Aggregator To establish a unified voice space, the multimodal representations $\{\mathbf{s}_f, \mathbf{s}_r, \mathbf{s}_t\}$ are aimed to align with speaker embeddings. Previous methods relied on limited datasets that matched only two modalities for alignment, resulting in a sparse distribution in the voice space and weak generalization capabilities.

KV-Former is proposed as a novel multimodal voice aggregator, distinct from existing module like Q-Former(Li et al., 2023b). This architecture integrates learnable key-value vectors into a simplified transformer, as shown in Figure 2. The multimodal representations act as queries and perform multi-head cross-attention with the learnable key-value vectors to retrieve the most informative representation in the voice subspace. The formulation of this process is as follows,

$$\mathbf{q} = \mathbf{W}^q \mathbf{s}_m, \mathbf{k} = \mathbf{W}^k \mathbf{m}, \mathbf{v} = \mathbf{W}^v \mathbf{m}, \quad (4)$$

$$\mathbf{a}_m = \mathrm{Softmax}\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d}}\right)\mathbf{v}, \quad (5)$$

where $\mathbf{W}$ are the projection matrices in attention, $\mathbf{s}_m \in \{\mathbf{s}_f, \mathbf{s}_r, \mathbf{s}_t\}$ represents various state vectors, $\mathbf{m}$ are learnable key-value vectors, d corresponds to the dimension of $\mathbf{f}$, and $\mathbf{a}_m$ is the output of cross attention. In this process, the learnable key-value vectors create an information bottleneck, facilitating interaction with the three modalities to build a shared voice space.

To establish a stable alignment across modalities while preserving natural voice characteristics, MVA adopts a speech-anchoring mechanism, which enables emergent alignment in voice space through shared key-value vectors and joint training, overcoming parallel data scarcity and maintaining voice diversity. It also supports easy expansion to new modalities via modality-specific encoders. During training, a stochastic mixed-modality strategy is employed, where reference speech is used as input with 50% probability.

To integrate multimodal inputs for voice control without losing the general abilities of CFM, the output of MVA is fed to the CFM and adapt the model without changing its weights. The MVA is trained to optimize $\mathcal{L}_{\text{OT-CFM}}$ with Equation (3) adapted for speaker embeddings as,

$$\begin{aligned}\nu_t\big(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)\big|\theta_{MVA}\big) \\ = \mathrm{NN}\big(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1), t; \mathbf{v}_m, \mathbf{C}\big),\end{aligned} \quad (6)$$

where $\mathbf{v}_m \in \{\mathbf{v}_f, \mathbf{v}_r, \mathbf{v}_t\}$ and $\mathbf{v}_m \in \{\mathbf{v}_f, \mathbf{v}_r, \mathbf{v}_t\}$ are MVA outputs for the corresponding inputs $\mathbf{s}_f$, $\mathbf{s}_r$, and $\mathbf{s}_t$. In this manner, CFM can incorporate

multiple modalities for voice control while maintaining its natural speech generation capability.

Soft Contrastive Learning Optimizing MVA solely through OT-CFM leads to slow convergence and potential voice discordance. To address this, Soft Contrastive Learning (SoftCL) is first introduced for speech-anchored multimodal alignment, inspired by studies (Wang et al., 2024). As shown in Figure 2, our approach ensures both inter-modal alignment and consistency within the similarity distribution. The basic inter-modal alignment is achieved by InfoNCE (Radford et al., 2021) as $\mathcal{L}_{\text{INTER}}$, attracting paired multimodal and speaker embeddings while repelling unpaired ones.

To bring cross-modal similarity closer to the distribution within each modality, intra-modal similarity is serve as soft labels to guide the inter-modal similarity generation. Given a batch of N multimodal-voice speaker embeddings pairs $\{(\mathbf{v}_m^i, \mathbf{s}_r^i)\}_{i=1}^N$, the intra-model similarity vector $p_i(\mathbf{s}_r, \mathbf{s}_r) = \{p_{ij}(\mathbf{s}_r, \mathbf{s}_r)\}_{j=1}^N$ can be obtained by:

$$p_{ij}(\mathbf{s}_r, \mathbf{s}_r) = \frac{\exp\left(\text{sim}\left(\mathbf{s}_r^i, \mathbf{s}_r^j\right)/\tau\right)}{\sum_{j=1}^N \exp\left(\text{sim}\left(\mathbf{s}_r^i, \mathbf{s}_r^j\right)/\tau\right)}, \quad (7)$$

where τ is a learnable temperature coefficient, and $\text{sim}()$ computes dot product similarity.

To address the issue where positive samples dominate with high confidence while negative sample relationships are overlooked, we disentangle the negatives in the distribution to boost the relation alignment. Given the self-similarity vector $p_i(\mathbf{s}_r, \mathbf{s}_r) \in \mathbb{R}^{1\times N}$, the neg-disentangled distribution $p_i^*(\mathbf{s}_r, \mathbf{s}_r) \in \mathbb{R}^{1\times(N-1)}$ is as follows,

$$p_{ij}^* = \frac{\exp\left(p_{ij}\right)}{\sum_{k=1,k\neq i}^N \exp\left(p_{ik}\right)}. \quad (8)$$

We also apply the above negative disentanglement to $p_i(\mathbf{s}_r, \mathbf{v}_m)$, yielding $p_i^*(\mathbf{s}_r, \mathbf{v}_m)$. Then, the intra-modality alignment supervision can be achieved with negative disentanglement as follows,

$$\mathcal{L}_{\text{INTRA}} = \frac{1}{N}\sum_{i=1}^N \text{KL}\left(p_i^*(\mathbf{s}_r, \mathbf{s}_r) \| p_i^*(\mathbf{s}_r, \mathbf{v}_m)\right), \quad (9)$$

where KL is the Kullback-Leibler Divergence.

Generally, UniSpeaker is trained to optimize the following loss function,

$$\mathcal{L} = \mathcal{L}_{\text{OT-CFM}} + \lambda_1 \mathcal{L}_{\text{INTRA}} + \lambda_2 \mathcal{L}_{\text{INTER}}, \quad (10)$$

Self-distillation Due to the cross-modal gap, our preliminary experiments reveal that CFM tend to extracts speaker information from semantic tokens while overlooking facial cues. To address this, we propose a simple yet effective self-distillation approach that requires only input conditioning modification. The process begins by generating converted speech using source speech semantic tokens, a Mel-spectrogram prompt, and randomly sampled speaker embeddings. Then, given the semantic tokens $\bar{\text{C}}$ of converted speech and speaker embeddings $\mathbf{s}$ of source speech, the CFM is fine-tuned to predict the source speech. The masked Mel-spectrogram prompt was removed to enhance the dominance of speaker embeddings in voice characteristic control, transforming Equation (3) as below,

$$\begin{aligned}\nu_t\left(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1)\middle|\theta_{FM}\right) \\ = \text{NN}\left(\phi_t^{OT}(\mathbf{X}_0, \mathbf{X}_1), t; \mathbf{s}, \bar{\text{C}}\right).\end{aligned} \quad (11)$$

In this way, the voice characteristics of the generated speech is controlled by the speaker embeddings input to the CFM, enabling direct multimodal integration in CFM without LLM modifications.

4 Dataset and Benchmark

Four modality-specific datasets were used to train the UniSpeaker, including LRS3-TED (Afouras et al., 2018), LibriTTS-P (Kawamura et al., 2024), VCTK-R (Sheng et al., 2024), and speaker identity description dataset [1] collected from the internet, totaling about 1000 hours of audio data.

The MVC Benchmark was established to evaluate multimodal voice control across five tasks. For face-related evaluation, 600 face images were randomly sampled from the LRS3-TED test set. Textual descriptions were generated by rewriting 600 sentences from the validation set using GPT-3.5-TURBO while preserving original meaning. For voice attribute editing, 200 sentences from VCTK were edited across all attributes for evaluation.

The benchmark assesses generated speech through three dimensions as below:

Voice suitability evaluates whether the voice characteristics of the generated speech align with the input description through three specific metrics: Speaker Similarity with Target (SST), Speaker Similarity Consistency (SSC), and MOS-Match.

[1] We requested the same data from the authors of CosyVoice-Instruct.

Task	Methods	Voice Suitability			Voice Diversity	Speech Quality	
		SST(%) ↑	SSC(%) ↑	MOS-Match ↑	SSD(%) ↓	WER(%) ↓	MOS-Nat ↑
FaceTTS	Imaginary Voice(Lee et al., 2023)	10.08	38.46	2.39 ± 0.09	32.17	8.23	2.45 ± 0.08
	Face-StyleSpeech(Kang et al., 2023)	11.02	37.09	2.78 ± 0.12	30.78	7.09	3.29 ± 0.10
	SYNTHE-SEES(Park et al., 2024)	10.97	38.81	2.92 ± 0.11	31.09	9.14	3.39 ± 0.09
	UniSpeaker(Ours)	**12.48**	**40.75**	$\mathbf{3.18 \pm 0.10}$	**14.09**	**4.01**	$\mathbf{3.82 \pm 0.08}$
FaceVC	FaceVC(Lu et al., 2021)	8.97	50.91	2.21 ± 0.11	30.19	10.90	2.79 ± 0.10
	SP-FaceVC(Weng et al., 2023)	9.52	52.29	2.39 ± 0.09	29.86	14.92	3.04 ± 0.10
	FVMVC(Sheng et al., 2023)	9.49	51.33	2.69 ± 0.07	22.60	11.94	3.31 ± 0.08
	UniSpeaker(Ours)	**11.68**	**55.13**	$\mathbf{3.09 \pm 0.10}$	**15.91**	**4.98**	$\mathbf{3.80 \pm 0.09}$
TextTTS	PromptSpeaker(Zhang et al., 2023)	17.39	-	3.64 ± 0.13	29.84	14.70	3.37 ± 0.10
	Promptttts++(Shimizu et al., 2024)	16.87	-	3.63 ± 0.12	35.42	15.08	3.41 ± 0.11
	CosyVoice-Instruct (Du et al., 2024)	14.51	-	3.71 ± 0.13	34.62	7.03	$\mathbf{3.91 \pm 0.09}$
	UniSpeaker (Ours)	**23.09**	-	$\mathbf{3.85 \pm 0.11}$	**21.10**	**6.46**	3.87 ± 0.13
TextVC	PromptVC(Yao et al., 2024)	16.59	-	3.47 ± 0.07	36.98	7.08	3.64 ± 0.10
	UniSpeaker(Ours)	**24.45**	-	$\mathbf{3.81 \pm 0.09}$	**24.04**	**6.29**	$\mathbf{3.77 \pm 0.11}$
AVE	VoxEditor(Sheng et al., 2024)	41.48	-	$\mathbf{3.78 \pm 0.09}$	49.92	8.01	3.57 ± 0.10
	UniSpeaker(Ours)	**49.04**	-	$\mathbf{3.79 \pm 0.10}$	**34.92**	**4.09**	$\mathbf{3.92 \pm 0.09}$

Table 2: Objective and subjective evaluation results of comparison systems. The definitions of all metrics can be found in Section 4. "-" denotes the results are not available.

Speaker similarity is calculated using cosine similarity between embeddings extracted via a pre-trained verification model[2]. The SST measures embedding similarity between generated and reference speech, while SSC evaluates voice consistency when different face images of the same speaker are used as input (Sheng et al., 2023). MOS-Match is derived from subjective listening tests, providing a mean opinion score to quantify how well the generated speech aligns with the input description.

Voice diversity assesses the capability to generate diverse voice characteristics from descriptions of various speakers, instead of producing voices that are very similar to each other. To evaluate this diversity, a metric called Speaker Similarity Diversity (SSD) is used. SSD measures similarity in speaker characteristics between speech generated from different descriptions (Sheng et al., 2023).

Speech quality assesses the robustness and naturalness of the generated speech with metrics: word error rate (WER) and MOS-Nat. An automatic speech recognition model[3] was used to transcribe the speech for WER. MOS-Nat provides mean opinion scores to evaluate speech naturalness.

5 Experiments

5.1 Experiment Settings

We trained the UniSpeaker using 4 NVIDIA TESLA V100 32G GPUs for 30K steps. The models were optimized using the AdamW optimizer with a learning rate of 1e-5 and a 10K warmup steps. The weights λ_1 and λ_2 in Equation (10) were set to 0.05. In the MVA architecture, the KV size is set to 128, the attention dimension is 786, and there are a total of 8 layers. The speech tokenizer and LLM were the same as those used in CosyVoice. For TTS, the LLM accepted only text inputs without speaker embeddings.

UniSpeaker was compared with 11 task-specific expert models in five tasks. We used the official code or pre-trained checkpoints of Imaginary Voice (Lee et al., 2023), FaceVC (Lu et al., 2021), SP-FaceVC (Weng et al., 2023), FVMVC (Sheng et al., 2023), and CosyVoice-Instruct (Du et al., 2024). The others were reproduced according to their original papers and evaluated using the same dataset.

5.2 Evaluation Results

Experimental results comparing UniSpeaker with current SOTA baselines are presented in Table 2, including both objective and subjective metrics.

Comprehensively, our method consistently outperforms baselines across five tasks, achieving superior suitability, diversity and speech quality. Notably, the joint multimodal training does not compromise performance on any individual modality, achieving an optimal balance. These results indicate that our approach goes beyond simple task concatenation, which achieves coordinated alignment between multiple modalities and speech, leading to overall performance enhancement.

[2] https://github.com/modelscope/3D-Speaker

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

Voice Suitability Our findings revealed that: 1) Across five tasks, UniSpeaker outperformed previous approaches, while maintaining competitive performance on AVE MOS-Match against VoxEditor's more complex architecture. 2) For face-based voice control, previous methods were just able to control the gender of the voice characteristics but exhibited obvious voice inconsistencies in subjective aspects such as age. In contrast, UniSpeaker achieved substantial improvements in both voice-age matching and overall subjective perception. 3) Additionally, ABX test in Figure 6 of Appendix showed cases where generated voices match face images even more closely than actual speaker, we are pleased to invite readers to listen to the samples on the demo page. 4) For text control, UniSpeaker resolves ambiguity issues arising from the text concatenation method used in CosyVoice-instruct and achieves consistent semantic-to-voice mapping, where similar semantics generate similar voice characteristics.

Voice Diversity Clearly, UniSpeaker achieve greater voice diversity across all five tasks. Furthermore, we visualized the speaker embeddings of the generated speech from both SYNTHE-SEES and UniSpeaker systems using t-SNE (Chan et al., 2019), as shown in Figure 3. It reveals that the voice space generated by our method is significantly richer, whereas the voice space of the baseline is relatively sparse. The result confirms UniSpeaker's ability to generate distinct voice characteristics for different faces where baselines fail.

Speech Quality By freezing the CFM during training, UniSpeaker preserve the general abilities of the backbone. Consequently, UniSpeaker surpasses previous methods in overall speech quality, only the MOS-Nat slightly lags behind CosyVoice-Instruct. This lag is due to the CFM occasionally learning noise patterns from the dataset. Conversely, CosyVoice-Instruct only integrate multimodal voice descriptions in the LLM, resulting in minimal impact on speech quality.

5.3 Ablation Study

Ablation studies about proposed modules and the training strategies (see in Table 3) show that:

1) KV-Former based MVA proves beneficial for voice control with a shared multimodal voice space. It utilizes multimodal data for joint modeling through shared key-value vectors, resulting in a uniform distribution of the voice space. This promotes alignment between different modalities

Figure 3: The voice space visualized by t-SNE. Compared to the baseline, UniSpeaker achieves richer voice diversity through face image input.

Task	Methods	SST(%) ↑	SSD(%) ↓	SSC(%) ↑
FaceTTS	UniSpeaker	12.48	14.09	40.75
	w.o. MVA	11.40	15.07	40.61
	w.o. SoftCL	11.57	15.94	38.28
FaceVC	UniSpeaker	11.68	15.91	55.13
	w.o. MVA	10.70	19.07	54.61
	w.o. SoftCL	11.08	19.24	51.55
TextTTS	UniSpeaker	23.09	21.10	-
	w.o. MVA	21.07	21.18	-
	w.o. SoftCL	22.57	34.51	-
TextVC	UniSpeaker	24.45	24.04	-
	w.o. MVA	21.50	24.26	-
	w.o. SoftCL	22.06	35.07	-
TTS	UniSpeaker	44.30	10.03	33.32
	w.o. self-distillation	38.49	9.80	29.68
VC	UniSpeaker	39.37	10.34	50.64
	w.o. self-distillation	31.07	10.16	43.62

Table 3: The ablation study of UniSpeaker, measured by SST, SSD and SSC.

and enhances the performance in both voice diversity and suitability. 2) Removing SoftCL results in a decline, specifically creating a significant mismatch between the generated voice and the input voice descriptions. Typically, when dealing with edge cases within the training set, such as elderly or young faces, removing SoftCL often leads to generate voices that sound like those of young adults. 3) Eliminating self-distillation also has notable effects. Experimental results indicated that self-distillation significantly enhanced voice control, particularly in SST. Due to the limited data scale, there was a slight reduction in diversity.

Data Scaling Analysis The influence of multimodal data scale on the shared voice space was investigated. For face-driven voice control, we trained UniSpeaker using various datasets: solely LRS3, and additional datasets of varying sizes. As shown in Figure 4, increasing multimodal data improves FaceVC and FaceTTS performance, confirming the advantages of joint multimodal modeling. Notably, SSC metrics were observed to be less sensitive to additional multimodal data, consistent

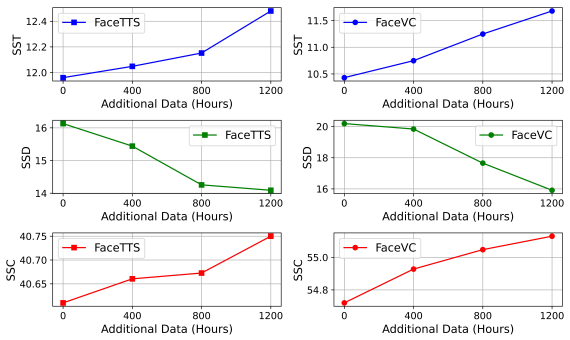

Figure 4: Evaluation of joint voice modeling across multimodal data scales. Here, the horizontal axis"0" indicates that only the LRS3 dataset was used.

Modality	SST (%) ↑	MOS-Match ↑
Face	10.71	2.79
Text	19.26	3.66
Face&Text	22.13	3.97

Table 4: Comparison of Performance Between Unimodal and Monomodal Speaker Generation.

with their dependence on intra-modal relationships.

Module Parameter Analysis Further ablation studies were conducted to examine the effects of key-value vector dimensions and MVA layer counts. Performance remained stable within a certain range, confirming effectiveness (established in prior experiments) while demonstrating new evidence of parameter robustness.

5.4 Visual Analysis

We randomly selected 8 unseen speakers and sampled 100 different face images from each for FaceTTS. The t-SNE visualization of speaker embeddings from generated speech is presented in Figure 5 (a), shows that the voice remained consistent across various facial images with different angles and backgrounds towards the same speaker. This indicates that UniSpeaker demonstrates strong robustness to noisy information in facial images. In addition, the identity descriptions are rewritten by the LLM in 60 different styles, ensuring that the semantics remained consistent despite variations in phrasing. Figure 5 (b) presents a visualization of the speech generated by TextTTS using these identity descriptions. The results demonstrate that for identity descriptions with the same semantics, the generated voices remain consistent.

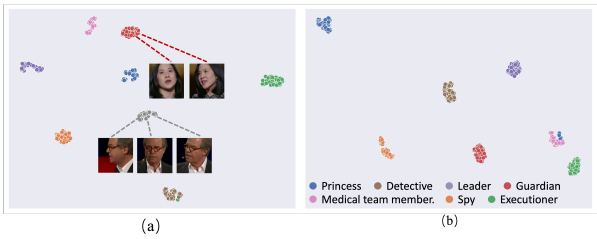

Figure 5: The visual analysis of UniSpeaker . In Figure (a), same colored points represent speech generated from different facial images of the same speaker. In Figure (b), same colored points represent speech generated from different identity descriptions of the same speaker, with annotations as text abbreviations.

5.5 Further Discussion

UniSpeaker maps multiple modalities to a unified voice space, enabling a more comprehensive characterization of voice. While both facial images and text descriptions exhibit a one-to-many relationship with voice features, they complement each other. We aim to explore whether UniSpeaker can better utilize multimodal inputs to achieve more accurate generation than single-input methods. In our experiment, 50 character images from various movies were paired with identity descriptions generated by an LLM. A multimodal TTS task was performed to compare unimodal and monomodal approaches. For the former, the multimodal speaker embeddings are combined via simple interpolation with equal weighting. Results in Table 4 demonstrate that integrating multiple modalities using MVA offers clear advantages over single-modality approaches. Importantly, UniSpeaker not only accepts multimodal inputs but actively coordinates them, leveraging mutual information from both to generate voice characteristics that better align with user needs.

6 Conclusion

In this paper, we propose UniSpeaker, a speech generation model that leverages multimodal voice description for voice control. With a unified voice aggregator and designed training strategies, UniSpeaker outperforms previous modality-specific models across five tasks, generating voices that better match the input descriptions. In the future, we will explore how to more effectively utilize multiple voice descriptions of different modalities for one speaker simultaneously and apply our method on other more modalities for voice control.

7 Limitations

In this section, we highlight the limitations of the proposed method and suggest possible direction in future work.

1. **Data Scale**: The dataset used for training is not large enough, which may limit the model's ability to generalize across a wide range of scenarios. Future work could focus on expanding the dataset size through automatic data collection techniques, which would improve the robustness and diversity.

2. **Multi-Style Input**: While the current method primarily focuses on realistic images, handling multi-style images (such as artistic or abstract representations) remains a limitation. Future research could explore how to adapt the model to work effectively with a broader range of image styles, beyond realistic inputs.

References

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. 2024. Seed-tts: A family of high-quality versatile speech generation models. *CoRR*, abs/2406.02430.

David M. Chan, Roshan Rao, Forrest Huang, and John F. Canny. 2019. GPU accelerated t-distributed stochastic neighbor embedding. *J. Parallel Distributed Comput.*, 131:1–13.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *CoRR*, abs/2407.05407.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15180–15190. IEEE.

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 *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18117–18125. AAAI Press.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Promptts: Controllable text-to-speech with text descriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Rongjie Huang, Ruofan Hu, Yongqi Wang, Zehan Wang, Xize Cheng, Ziyue Jiang, Zhenhui Ye, Dongchao Yang, Luping Liu, Peng Gao, et al. 2024. Instructspeech: Following speech editing instructions via large language models. In *Forty-first International Conference on Machine Learning*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Minki Kang, Wooseok Han, and Eunho Yang. 2023. Face-stylespeech: Improved face-to-voice latent mapping for natural zero-shot speech synthesis from a face image. *CoRR*, abs/2311.05844.

Masaya Kawamura, Ryuichi Yamamoto, Yuma Shirahata, Takuya Hasumi, and Kentaro Tachibana. 2024. Libritts-p: A corpus with speaking style and speaker identity prompts for text-to-speech and style captioning. *CoRR*, abs/2406.07969.

Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. 2023. Imaginary voice: Face-styled diffusion model for text-to-speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

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. Promptts 2: Describing and generating voices with text prompt. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jingyi Li, Weiping Tu, and Li Xiao. 2023a. Freevc: Towards high-quality text-free one-shot voice conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Hsiao-Han Lu, Shao-En Weng, Ya-Fan Yen, Hong-Han Shuai, and Wen-Huang Cheng. 2021. Face-based voice conversion: Learning the voice behind a face. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 496–505. ACM.

Jae Hyun Park, Joon-Gyu Maeng, Taejun Bak, and Young-Sun Joo. 2024. SYNTHE-SEES: face based text-to-speech for virtual speaker. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 10321–10325. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society.

Zhengyan Sheng, Yang Ai, Yan-Nian Chen, and Zhen-Hua Ling. 2023. Face-driven zero-shot voice conversion with memory-based face-voice alignment. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 8443–8452. ACM.

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

Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana. 2024. Promptttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 12672–12676. IEEE.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. 2023. Conditional flow matching: Simulation-free dynamic optimal transport. *CoRR*, abs/2302.00482.

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.

Zachary Wallmark. 2019. A corpus analysis of timbre semantics in orchestration treatises. *Psychology of Music*, 47(4):585–605.

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. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023b. CAM++: A fast and efficient network for speaker verification using context-aware masking. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 5301–5305. ISCA.

Qian Wang, Jia-Chen Gu, and Zhen-Hua Ling. 2024. Multiscale matching driven by cross-modal similarity consistency for audio-text retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 11581–11585. IEEE.

Shao-En Weng, Hong-Han Shuai, and Wen-Huang Cheng. 2023. Zero-shot face-based voice conversion: Bottleneck-free speech disentanglement in the real-world scenario. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence*, pages 13718–13726. AAAI Press.

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 Trans. Audio Speech Lang. Process.*, 32:2913–2925.

Jixun Yao, Yuguang Yang, Yi Lei, Ziqian Ning, Yanni Hu, Yu Pan, Jingjing Yin, Hongbin Zhou, Heng Lu, and Lei Xie. 2024. Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 10571–10575. IEEE.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503.

Yongmao Zhang, Guanghou Liu, Yi Lei, Yunlin Chen, Hao Yin, Lei Xie, and Zhifei Li. 2023. Promptspeaker: Speaker generation based on text descriptions. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–7. IEEE.

Table 5: Ablation experiments to explore the impact of the LLM and CFM on voice characteristics under different conditions. The ✓ indicates the input is a regular speaker embeddings, while the ✗ denotes random noise input.

Condition	LLM	CFM	SSIM
With Mel-spectrogram Prompt	✓	✓	62.76
	✗	✓	57.03
	✓	✗	29.39
	✗	✗	24.04
Without Mel-spectrogram Prompt	✓	✓	44.07
	✗	✓	34.51
	✓	✗	8.44
	✗	✗	4.42

Table 6: Performance of different models on the voice conversion task, where * indicates the absence of Mel-spectrogram prompt. Note that these results are not comparable to those in Table 3 due to different test samples

model	SSIM
Groud Truth	69.67
CosyVoice (Du et al., 2024)	72.63
CosyVoice*	43.59
FreeVC (Li et al., 2023a)	36.31
FACodec (Ju et al., 2024)	52.73

A Analysis about CosyVoice

A.1 Impact of the LLM and CFM modules on voice characteristics

In the zero-shot speech synthesis task, the speaker embeddings input to either the LLM or Flow were replaced with random tensors of the same size. For evaluation, 500 sentences from the LRS3 dataset were selected, and the speaker similarity between the generated speech and the source speech was computed, as shown in Table 5. The results indicate that, compared to CFM, LLM has a significantly smaller impact on voice characteristics due to the limited voice characteristics contained in semantic tokens. Additionally, the balance between semantic and voice characteristics within semantic tokens across different scenarios is worth further exploration.

Additionally, by comparing the performance under both conditions in Table 5, we found that the Mel-spectrogram prompt carries more voice information than the speaker embeddings. In fact, the Mel-spectrograms offers a more detailed representation of voice characteristics, while the speaker embeddings provides a coarser one. For multimodal voice alignment tasks, multimodal voice descriptions are inherently incomplete and imprecise (Leng et al., 2024; Sheng et al., 2024), with a one-to-many mapping to voice characteristics. Thus, a coarse speaker embeddings is sufficient to serve as an anchor for multimodal alignment.

A.2 Performance of CosyVoice on zero-shot Voice Conversion

Before self-distillation, we evaluated the zero-shot voice conversion performance of CosyVoice. We extracted semantic tokens from the source speech and speaker embeddings, along with the Mel-spectrogram prompt from the reference speech, as inputs for the CFM. The generated speech retained the content and prosody of the source while altering the speaker's identity. We randomly selected 500 sentences from the LibriTTS test set to evaluate the performance of the CosyVoice, FreeVC[4] (Li et al., 2023a), and FAcodec[5] (Ju et al., 2024) models, with the experimental results presented in Table 6. During inference, when both the Mel-spectrogram prompt and speaker embeddings were provided, CFM-generated speech surpasses real speech in objective metrics. This is attributed to the fact that voice characteristics can exhibit local variations driven by content, rhythm, and emotion. This suggests that the audio produced by CFM is independent of the speaker information in the source semantic tokens, achieving exceptional disentanglement of voice characteristics. This makes it well-suited for self-distillation.

Without the Mel-spectrogram prompt, performance was inferior to FAcodec, which can be attributed to the inconsistency between training and inference, as the model was trained with both the Mel-spectrogram prompt and speaker embeddings as input. After self-distillation, the performance relying solely on speaker embeddings showed a marked improvement, as indicated in Table 3.

A.3 Preliminary experiment on face-based voice description integration

In our preliminary experiments, we directly integrated face embeddings into the CFM of the official CosyVoice[6]. Specifically, we utilized a pre-trained face encoder to extract the global face embeddings, replacing the speaker embeddings in the CFM as input. We evaluated the trained model on the zero-shot voice conversion task and found that the resulting SSIM score was around 4.8, indicating that

[4] https://github.com/OlaWod/FreeVC
[5] https://github.com/Plachtaa/FAcodec
[6] https://github.com/FunAudioLLM/CosyVoice

the generated speech retained the identity information of the source speaker and did not achieve speaker conversion. This suggests that due to the cross-modal gap, CFM tends to extract voice information from the semantic tokens while neglecting the speaker information contained in the face. Therefore, to enable CFM to effectively utilize multimodal voice description integration, further voice disentanglement are necessary.

B Details of Voice Attribute Descriptions

For voice attribute description input, the model receives an input tuple consisting of two speech segments (A and B) and a text description t. The text description states that A exhibits a certain attribute more prominently than B. For example, t refers to "sounds more magnetic" meaning that voice characteritics of sample A is more magnetic than that of B. Following VoxEditor (Sheng et al., 2024), we first concatenate the speaker embeddings $\mathbf{s}_r^A, \mathbf{s}_r^B$ of two given speech samples and the text representation $\mathbf{s}_t$. Through MLP and Gaussian sampling, we predict the density difference $\alpha \in [0, 1]$ in attribute x between the two speeches samples, and then obtain the target speaker embeddings $\mathbf{s}_r$ via linear interpolation: $\mathbf{s}_r = (1 - \alpha) \cdot \mathbf{s}_r^B + \alpha \cdot \mathbf{s}_t$. During inference, we can control the density of the target voice attribute by adjusting α within a range of 0 to 1.

C Details about InfoNCE

Specifically, given a batch of N multimodal-voice speaker embeddings pairs $\{(\mathbf{v}_m^i, \mathbf{s}_r^i)\}_{i=1}^N$, the multimodal-voice similarity vector $p_i(\mathbf{v}_m, \mathbf{s}_r) = \{p_{ij}(\mathbf{v}_m, \mathbf{s}_r)\}_{j=1}^N$ and voice-to-multimodal similarity vector $p_i(\mathbf{s}_r, \mathbf{v}_m) = \{p_{ij}(\mathbf{s}_r, \mathbf{v}_m)\}_{j=1}^N$ can be calculated as follows,

$$
\begin{aligned}
p_{ij}(\mathbf{v}_m, \mathbf{s}_r) &= \frac{\exp\left(\operatorname{sim}\left(\mathbf{v}_m^i, \mathbf{s}_r^j\right)/\tau\right)}{\sum_{j=1}^N \exp\left(\operatorname{sim}\left(\mathbf{v}_m^i, \mathbf{s}_r^j\right)/\tau\right)}, \\
p_{ij}(\mathbf{s}_r, \mathbf{v}_m) &= \frac{\exp\left(\operatorname{sim}\left(\mathbf{s}_r^i, \mathbf{v}_n^j\right)/\tau\right)}{\sum_{j=1}^N \exp\left(\operatorname{sim}\left(\mathbf{s}_r^i, \mathbf{v}_n^j\right)/\tau\right)}
\end{aligned}
\tag{12}
$$

where τ is a learnable temperature coefficient, initialized to 0.07, and $\operatorname{sim}()$ denotes the dot product used to calculate similarity. The inter-modal alignment loss is computed using cross-entropy as follows,

$$
\begin{aligned}
\mathcal{L}_{\text{INTER}} = &\frac{1}{2N}\sum_{i=1}^{N}\mathcal{L}_{CE}\left(\mathbf{y}_i, p_i(\mathbf{v}_m, \mathbf{s}_r)\right) \\
&+\frac{1}{2N}\sum_{i=1}^{N}\mathcal{L}_{CE}\left(\mathbf{y}_i, p_i(\mathbf{s}_r, \mathbf{v}_m)\right)
\end{aligned}
\tag{13}
$$

where $\mathcal{L}_{CE}$ denotes the cross-entropy operation and $\mathbf{y}_i$ the one-hot label of i_{th} pair.

D Details of Datasets and Benchmark

D.1 Training Datasets

For the LRS3-TED video dataset, 100 facial images per speaker were randomly selected from the videos, and a facial attribute detection model FairFace[7] was used to further clean the data. Specifically, the speaker's age and gender were estimated based on the 100 images, calculating the mean and variance. If the variance was too large, indicating poor video quality for that speaker, all samples from that speaker were discarded. Anomalies in these 100 images, often blurry pictures or images of a different speaker, were also filtered out. During training, a random image from the given speaker's image set was selected as input. FFmpeg[8]was used to extract 16kHz audio from the video. Additionally, the LRS3 dataset is also utilized for self-distillation of the CFM. For libritts-p, following promptttts2 (Leng et al., 2024), we converted the these word-level annotations about voice characteristics into natural descriptive language using a language model.

D.2 Evaluation Datasets

To evaluate the effectiveness of the text descriptions, we used a language model to rewrite sentences from the validation set while maintaining their original meaning, as shown in Table 7. This approach allows us to assess the model's generalization ability while providing targeted audio for comparison. Additionally, we prompted a large language model to randomly generate 100 character descriptions and voice characteristics descriptions, which can be considered out-of-domain. To further validate out-of-domain face image, we selected an Asian face dataset[9] for testing, given that the LRS3 dataset was collected from TED. The generated speech are available on the demo website.

D.3 Evaluation Metrics

For SST, when performing FaceTTS, FaceVC, TextTTS, and TextVC tasks, a multimodal voice description is provided along with a corresponding target speech. This allows us to directly calculate the speaker similarity between the generated speech and the target speech. However, for the AVE task, as there are no real voice characteristics, we calculate the speaker similarity with the source speech. The AVE task aims to edit specific voice attributes while preserving other characteristics as much as possible, so SST is used to assess whether the edited speech retains the original voice characteristics. Therefore, we need to combine SST and MOS-Match to comprehensively evaluate the performance of AVE.

For SSD, we matched generated speech with voice descriptions for different speakers to calculate speaker similarity, and then averaged the results across the evaluation dataset. A smaller average indicates greater voice diversity within the dataset. Specifically, for the AVE task, the diversity of the generated speech is assessed by applying the same voice attribute editing with the same weights to different speech inputs.

For SSC, pairwise matching of different images of the same person was performed to calculate their speaker similarity. These values were then averaged across the entire evaluation dataset. A higher average indicates greater voice similarity between different photos of the same individual, suggesting that the model is robust to background noise and other variations in the images.

For MOS-Match and MOS-Nat, subjective evaluation were conducted on Amazon Mechanical Turk[10]. Twenty sentences were randomly selected, and 20 listeners were asked to score each generated utterance on a scale from 1 (completely mismatched or completely unnatural) to 5 (completely matched or completely natural) for both metrics.

E Comparative Methods

FaceTTS baselines:

- Imaginary Voice (Lee et al., 2023) is based on a score-based diffusion model, specifically Grad-TTS. Imaginary Voice used perceptual loss applied to the Mel-spectrograms to further align facial features and language.
- Face-StyleSpeech (Kang et al., 2023) proposes the disentangling of prosody and timbre, using facial features to control timbre and reference audio to control prosody. It also employs a contrastive learning to align facial and speaker embeddings.
- SYNTHE-SEES (Park et al., 2024) utilizes three types of losses—contrastive learning, speaker classification, and perceptual loss—to align face and speaker embeddings.

[7] https://github.com/dchen236/FairFace
[8] https://ffmpeg.org/
[9] https://github.com/X-zhangyang/Asian-Face-Image-Dataset-AFD-dataset
[10] https://www.mturk.com/

Table 7: An Example of using LLM to generate synonymous sentences.

Diglogue	
LLM prompts:	Rewrite the following sentence, keeping the meaning unchanged, with a variety of sentence structures and styles. Please replace key words with synonyms: Princess X is honored as a priestess of the winter sea god, portrayed as a woman imbued with deep nostalgia and melancholy, while also being a contemporary fashion designer who cherishes traditional craftsmanship.
Response:	Princess X is revered as a high priestess of the deity of the winter sea, depicted as a figure filled with profound wistfulness and sorrow, yet she is also a modern fashion designer who values artisanal traditions.

FaceVC baselines:

- FaceVC (Lu et al., 2021) employed a three stage training strategy, including face-voice reparameterization and facial-to-audio transformation, to align the face and voice characteristics.
- SP-FaceVC (Weng et al., 2023) first employed a bottleneck-free strategy for speech disentanglement. Then, multi-Scale discriminator and feature matching loss was proposed to improve performance.
- FVMVC (Sheng et al., 2023) used FaceNet to extract general face embeddings and employ the memory net to align the face embeddings and speaker embeddings.

TextTTS baselines:

- PromptSpeaker (Zhang et al., 2023) annotated an internal dataset of speaker descriptions on LibriTTS-R. Building on this dataset, PromptSpeaker employed a pre-trained BERT network in conjunction with a Glow model to achieve alignment with speaker embeddings.
- Promptttts++ (Shimizu et al., 2024) integrated a BERT network with a Gaussian mixture model to predict speaker embeddings based on text descriptions, utilizing cosine loss for alignment.
- CosyVoice-Instruct (Du et al., 2024) concatenated the speaker's description before the text content in the LLm module of CosyVoice during training.

TextVC baseline:

- PromptVC (Yao et al., 2024) utilized HuBERT and k-means clustering to represent semantic intermediate representations, and employed a diffusion model to predict style representations based on text input. Here, we replaced the dataset with ours to predict speaker embeddings using the diffusion model.

Table 8: The detailed model configurations of MVA.

Configuration	Value
Layer	8
Attention Dim	768
Attention Heads	16
Linear Dim	2048
Dropout	0.1
KV Size	128

AVE baseline:

- VoxEditor (Sheng et al., 2024) first annotated a dataset describing timbre characteristics and utilized a residual memory network to accomplish the voice attribute editing.

F Further Ablation Studies

We present the ablation experiment results for the TextTTS and TextVC tasks in Figure 9. This indicates that MVA and SoftCL are also beneficial for text-based timbre control. Additionally, we conducted ablation experiments on the size of learnable key-value vectors and the number of MVA layers, and found that within a certain range, the performance of voice control is not significantly affected, yet no clear patterns could be derived.

G Further Discussion

A unified voice space is constructed through a unified voice compressor. To validate the benefits of this shared space, voice interpolation on the speaker embeddings from different modalities is performed, allowing for manually adjusting the interpolation weights α. As shown in Figure **??**, we achieve

Table 9: The results of ablation studies on TextTTS and TextVC tasks, measured by SST, SSD.

Task	Methods	SST ↑	SSD ↓
TextTTS	UniSpeaker	23.09	21.10
	w.o. MVA	21.07	21.18
	w.o. SoftCL	22.57	34.51
TextVC	UniSpeaker	24.45	24.04
	w.o. MVA	21.50	24.26
	w.o. SoftCL	22.06	35.07

(a) FaceVC (b) FaceTTS

Figure 6: Average preference scores (%) of ABX tests about voice suitability in comparison, where participants were asked to select which of two speech samples—one generated based on the reference speaker's face image and one from the reference speaker's recording—better matched the speaker's appearance. "N/P" stands for "no preference". "Ground Truth" represents the real recording of the reference speaker.

voice control by interpolating the speaker embeddings obtained from face and textual descriptions. We observe that the voice characteristics vary as α changes, speech samples are available in the demo page.

By mapping multiple modalities to a unified voice space, we can leverage these different modalities to more comprehensively describe voice characteristics. Both face images and textual descriptions maintain a one-to-many relationship with the voice characteristics themselves. This means that given a face image or a textual description, the model can generate multiple matching voice characteristics. When both the target speaker's face and the textual voice description are input simultaneously, the generated voice characteristics that align with both modalities will better meet user expectations. Furthermore, we can editing specific voice attributes for more refined optimization. In the future, we will explore how to more finely utilize multiple modalities for voice control concurrently.

This is a section in the appendix.

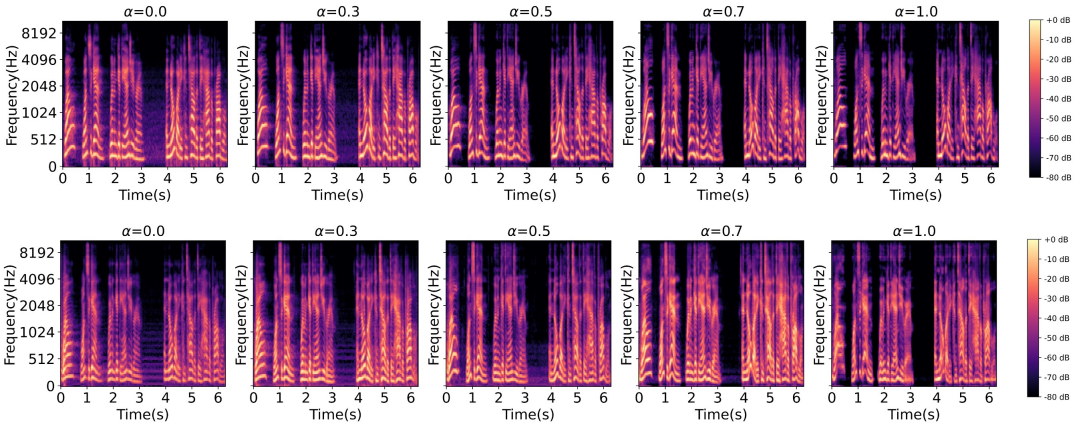

Figure 7: The voice characteristics controlled by both face and textual descriptions varies as α changes. When $\alpha = 0$, the voice characteristics are fully controlled by the face; when $\alpha = 1$, the voice characteristics are fully controlled by the textual description. We can observe the changes in voice characteristics and manually adjust α to achieve the desired voice characteristics.