Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

HiP-ETTS: Multimodal Emotional Text-to-Speech via Shared-Specific Emotion Modeling and Hierarchical Prior Projection and Refinement

Ziyu Zheng

zzhe232@aucklanduni.ac.nz

University of Auckland

xinfeng Ye

University of Auckland

Sathiamoorthy Manoharan

University of Auckland

Research Article

Keywords:

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11590288/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

HiP-ETTS: Multimodal Emotional Text-to-Speech via Shared-Specific Emotion Modeling and Hierarchical Prior Projection and Refinement

Ziyu Zheng, Xinfeng Ye, Sathiamoorthy Manoharan

School of Computer Science, The University of Auckland
Auckland, 1010, New Zealand

Abstract

Multimodal emotional prompts have broadened the controllability of emotional text-to-speech synthesis by allowing target emotions to be specified through diverse cues, including speech, text, and facial images. Nevertheless, existing methods still face two fundamental challenges: entangled emotional representations and the trade-off between content fidelity and emotional expressiveness. To address these challenges, we propose HiP-ETTS. It first decomposes emotional prompts into two fine-grained factors: an emotion-shared factor and a complementary emotion factor, which respectively represent VAD-related affective structure and complementary expressive information beyond the VAD-conditioned prediction. Conditioned on these emotion factors, HiP-ETTS then constructs a content-stable hierarchical emotional prior and adaptively projects it into an intermediate flow state for subsequent acoustic refinement, thereby enhancing emotional expressiveness while preserving linguistic content. Experiments across different emotional prompt conditions demonstrate that HiP-ETTS achieves lower or comparable word error rates and consistently higher emotional consistency scores than previous methods, while requiring fewer parameters and enabling faster inference.

Introduction

Emotional text-to-speech synthesis aims to generate speech that is natural, intelligible, and consistent with a target emotion, with broad applications in voice assistants, virtual characters, and media production (Xie et al. 2025). Although recent TTS models, such as NaturalSpeech (Tan et al. 2024), have approached the naturalness of human recordings, the generation of fine-grained and controllable emotional speech remains insufficiently addressed. We argue that this problem mainly stems from two fundamental challenges:

Entangled Emotional Representations. Recent emotional TTS studies have extended emotion control from single-modality conditioning to multimodal prompting. Representative methods such as MM-TTS (Guan et al. 2024) and FleSpeech (Li et al. 2025) employ multimodal prompting frameworks that allow textual, acoustic, and visual cues to condition emotional speech generation. Although such frameworks provide more flexible and diverse forms of emotion control, the emotional information carried by these cues still requires a more explicit structure.

In real emotional speech, the mapping from emotion to speech is not a fixed one-to-one relationship (Schuller 2018). Rather, it reflects both cross-speaker acoustic regularities and speaker-specific expressive variation. For example, speakers commonly exhibit increased loudness and faster speech rate when expressing anger, reflecting acoustic regularities associated with the same emotion (Banse and Scherer 1996; Sobin and Alpert 1999). However, the specific realization of these regularities is shaped by each speaker's habitual speaking style (Patel et al. 2011; Laukka, Neiberg, and Elfenbein 2014). Based on this observation, we decompose the emotional information carried by these prompts into two factors: an emotion-shared factor that captures cross-speaker acoustic regularities, and a complementary emotion factor that provides expressive information beyond the VAD-conditioned prediction. This decomposition formulates emotional speech synthesis as a shared-specific emotion modeling problem, in which the emotion-shared factor establishes a common emotional backbone, while the complementary emotion factor provides additional cues for speaker-conditioned prosody prediction.

Trade-Off Between Content Fidelity and Emotional Expressiveness. Previous emotional speech synthesis methods typically introduce emotional conditions through concatenation, feature fusion, or global modulation (Yao et al. 2024; Ning et al. 2023; Choi and Park 2025). However, emotional signals can induce substantial acoustic variations. Directly injecting them into the acoustic model may therefore improve emotional expressiveness at the expense of content fidelity (Yang et al. 2025b,c). This problem is further exacerbated by the heterogeneity of speech factors: content and prosody are represented as time-varying sequences, whereas timbre and emotion are modeled as utterance-level global conditions (Ren et al. 2024; Wu et al. 2025). Consequently, enhancing emotional expression without sacrificing content fidelity remains a key challenge in emotional speech synthesis.

To address this issue, we design a two-stage acoustic generation process. In the first stage, the model constructs a content-stable hierarchical emotional prior by organizing different speech factors according to their properties. In the second stage, conditioned on acoustic cues, the model projects this prior into an adaptive intermediate state on the flow trajectory and refines it into the final mel-spectrogram through

conditional flow matching. This design first anchors linguistic structure in the hierarchical prior and subsequently enriches emotional expression and acoustic detail during flow refinement.

In summary, we propose **HiP-ETTS**, a multimodal emotional TTS framework that enables more fine-grained and content-stable emotional speech synthesis by decomposing prompt-derived emotional signals and adopting two-stage acoustic generation. Our main contributions are summarized as follows:

- We propose a fine-grained emotional signal decomposition strategy that separates emotional information into emotion-shared and complementary emotion factors, enabling more speaker-congruent emotional expression.
- We propose a two-stage acoustic generation strategy that jointly preserves content fidelity and enhances emotional expressiveness by first establishing a content-stable hierarchical emotional prior and then adaptively projecting and refining it along the flow trajectory.
- Extensive experiments across different emotional prompt modalities demonstrate that HiP-ETTS achieves higher emotion consistency with lower or comparable word error rates, while enabling flexible and speaker-congruent emotional expression beyond fixed categorical emotion labels.

Related Work

Emotion Representation and Multimodal Prompting

A key challenge in emotional speech synthesis is how the target emotion is represented and provided to the model. Early methods commonly rely on discrete emotion labels, such as happiness, sadness, and anger. While these labels are intuitive and easy to use, they are limited to predefined categories and cannot capture expressive variations within the same category (Lee, Lee, and Rabiee 2017). Dimensional emotion models instead represent emotion using continuous coordinates, such as valence, arousal, and dominance (VAD), which characterize emotional polarity, activation, and perceived control, respectively (Russell 1980). Based on this representation, EmoSphere and EmoSphere++ control emotion category and intensity by manipulating coordinates in the emotion space (Cho et al. 2025, 2024). Meanwhile, in emotional understanding field, prior studies have investigated decomposing emotional representations into shared and modality-specific components to examine their relationships and complementary roles (Hazarika, Zimmermann, and Poria 2020). In emotional speech generation, EmoRES-TTS also explores a shared–residual decomposition to model commonalities across emotion categories and category-specific differences (Huang et al. 2026).

More recent emotional TTS studies have further broadened the sources of emotion control signals to include reference speech (Chen, Chen, and Hirschberg 2024), natural language descriptions (Yang et al. 2024), and facial images (Zhang et al. 2025). Extending this direction, MM-TTS and FleSpeech incorporate text, speech, and image inputs into multimodal prompting frameworks, allowing either a single modality or their combinations to provide emotional cues (Guan et al. 2024; Li et al. 2025). However, they generally do not explicitly separate cross-speaker emotional regularities from speaker-specific expressive variations. Consequently, the learned emotion condition may conflate speaker-specific expression with the target emotion. To address this limitation, we reformulate the representation of emotional prompts as a shared-specific modeling problem, in which VAD models the cross-speaker acoustic regularities and a high-dimensional residual provides complementary expressive information beyond the VAD-conditioned prediction.

Structured Priors and Factorized Speech Modeling

In flow-matching based TTS, acoustic generation typically starts from noise or a simple initial state and progressively evolves toward the final mel-spectrogram (Popov et al. 2021; Mehta et al. 2024). More recent methods have begun to introduce informative priors or intermediate states as the starting state for acoustic generation. For example, OZSpeech (Hieu et al. 2025) uses one-step OT-CFM from a learned prior distribution over factorized speech codes, while SFM (Yang et al. 2026a) constructs an adaptive intermediate state from coarse acoustic outputs and refines only the remaining trajectory. These studies suggest that a more informative starting state can facilitate acoustic generation.

Meanwhile, factorized speech modeling has become an important method for controllable speech synthesis. iEmoTTS (Zhang et al. 2023) separates emotion-related prosody from speaker timbre through dedicated emotion, prosody, and timbre modules, enabling cross-speaker emotion transfer while preserving speaker identity. NaturalSpeech 3 (Ju et al. 2024) uses a factorized neural codec to decompose speech into content, prosody, timbre, and acoustic-detail subspaces, and generates these speech factors with corresponding diffusion models. These methods demonstrate that assigning different speech factors to specialized representations facilitates controllable synthesis. Motivated by these findings, we construct a content-stable hierarchical emotional prior from factorized speech factors and use it as the basis for subsequent flow refinement.

Method

In this paper, we consider emotional speech synthesis under multimodal prompts. Given content, speaker, and multimodal emotional prompts, HiP-ETTS aims to generate speech that is consistent with the target emotion and aligned with the expressive style of the target speaker. The framework consists of two components: **Shared-Specific Multimodal Emotion Modeling**, which decomposes emotional prompts into emotion-shared and complementary emotion factors and uses them to condition prosody prediction; and **Hierarchical Emotional Prior Projection and Refinement**, which organizes content, prosody, timbre, and emotion according to their respective properties to construct a content-stable hierarchical emotional prior. The prior is subsequently projected into an adaptive intermediate state and refined into the target mel-spectrogram. The overall architecture is illustrated in Figure 1.

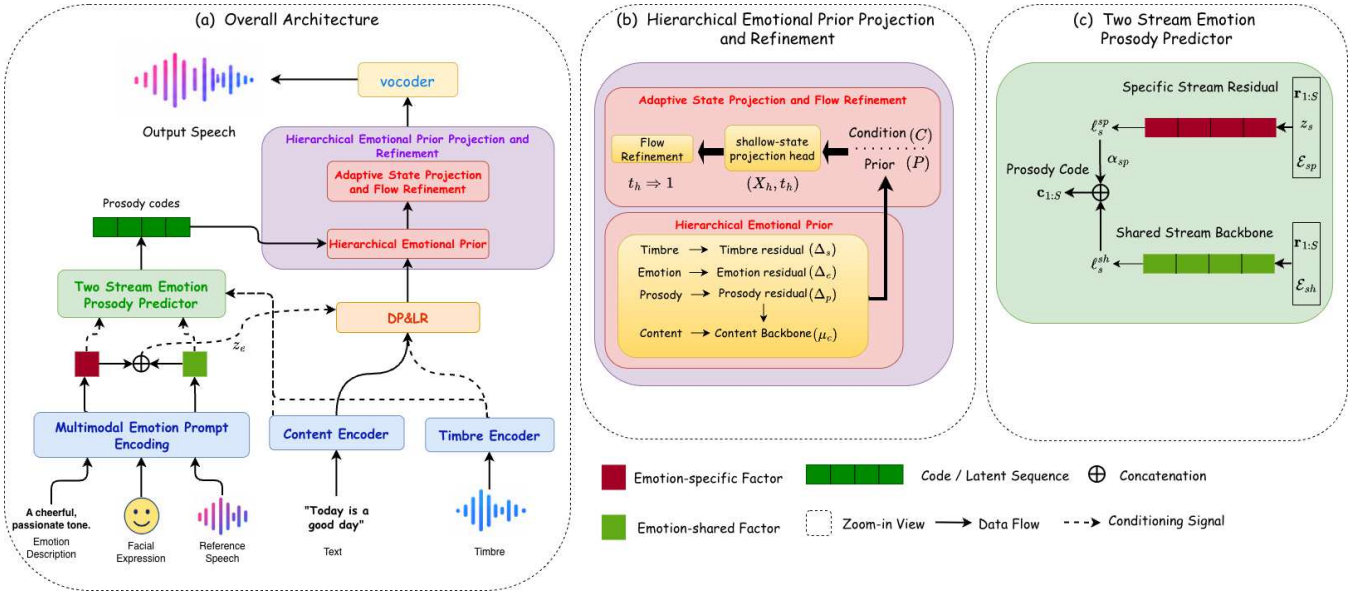


Figure 1: Overview of HiP-ETTS. (a) Emotional prompts are decomposed into emotion-shared and complementary emotion factors, which are used to predict prosody codes. The content, timbre, and emotion information are fed into the duration prediction and length regulation module. The resulting frame-level content representation, together with the prosody, timbre and emotion conditions, is then passed to the hierarchical emotional prior projection and refinement module to generate the final mel-spectrogram. (b) A coarse mel prior P is constructed from a content backbone and factor-specific residual offsets, projected to an adaptive state–time (X_h, t_h), and refined along the remaining flow trajectory. (c) The shared stream predicts a prosodic backbone, while the speaker-conditioned specific stream provides a residual correction. Their logits are combined using α_{sp}.

Hierarchical Emotional Prior Projection and Refinement

Hierarchical Emotional Prior. Conditioned on speech factors and duration information, the model first constructs a content-stable hierarchical emotional prior P. The prior has the same temporal resolution and dimensionality as the target mel-spectrogram and provides a coarse mel-space representation that preserves linguistic structure while incorporating expressive cues. It serves as the basis for adaptive state projection and flow refinement. Specifically, we organize different speech factors according to their functional roles and temporal granularity. Content and prosody are modeled as time-varying factors and represented at the frame-level as H_c^f and H_p^f, respectively (Ren et al. 2024). In contrast, timbre and emotion are represented as utterance-level global conditions z_s and z_e (Du et al. 2026).

We construct the prior in a hierarchical residual form. The content branch first predicts a base representation, while the remaining branches produce factor-specific residual offsets. Since prosody is a frame-level condition, it is directly combined with the content representation. Timbre and emotion, as global conditions, are instead incorporated through FiLM modulation (Perez et al. 2018):

$$\begin{aligned} \mu_c &= f_c(H_c^f), & \Delta_p &= f_p(H_c^f, H_p^f), \\ \Delta_s &= f_s(\mathrm{FiLM}(H_c^f, z_s)), & \Delta_e &= f_e(\mathrm{FiLM}(H_c^f, z_e)). \end{aligned} \tag{1}$$

Here, μ_c denotes the content backbone, while Δ_p, Δ_s, and Δ_e denote the prosodic, timbral, and emotional residual offsets, respectively. The hierarchical emotional prior is then formed by combining content backbone with these residual offsets:

$$P = \mu_c + \lambda_p \Delta_p + \lambda_s \Delta_s + \lambda_e \Delta_e, \tag{2}$$

where λ_p, λ_s, and λ_e are learnable scaling coefficients for the prosody, timbre, and emotion residuals, respectively. This hierarchical residual formulation preserves the content backbone while allowing the remaining factors to introduce controlled modifications. Accordingly, during training, the hierarchical emotional prior is optimized by:

$$\mathcal{L}_{\text{hep}} = \|P - Y\|_1 + \lambda_a \left\| P - P^{(0)} \right\|_1, \tag{3}$$

where Y is the target mel-spectrogram, $P^{(0)}$ denotes the prior without the emotion residual offset, and λ_a controls the anchor strength. The first term aligns P with the target mel structure, while the second constrains emotion-induced deviations from the content-preserving prior.

Adaptive State Projection and Flow Refinement. After obtaining the coarse mel-spectrogram prior P, our goal is to use it as the basis for generating the target mel-spectrogram. This coarse-to-fine perspective raises a key question: if acoustic refinement is viewed as a generative trajectory toward the target mel-spectrogram, where should the current prior P be located on this trajectory? Since the alignment between P and the target mel-spectrogram varies across utterances, the required refinement extent also varies accordingly. Therefore, using a fixed flow start time would require

the decoder to bridge different acoustic gaps under the same starting-time assumption.

To address this issue, inspired by SFM (Yang et al. 2026a), we introduce a shallow-state projection head that estimates an adaptive intermediate state and its corresponding flow time. Specifically, in parallel with the prior branch, we construct a frame-level decoder condition C. It is similar to the conditional representation used in flow-matching TTS models such as Matcha-TTS (Mehta et al. 2024), but further integrates prosody and emotion information. The shallow-state projection head jointly maps (P, C) to an intermediate state and its corresponding position on the flow trajectory:

$$H_h = f_h([P, C]), \quad X_h = P + g_x(H_h), \tag{4}$$

$$t_h = t_{\min} + (t_{\max} - t_{\min}) \cdot \mathrm{sigmoid}(g_t(\bar{H}_h)). \tag{5}$$

Here, H_h combines the coarse acoustic structure carried by P with the contextual refinement cues provided by C. The predicted offset $g_x(H_h)$ projects P to an intermediate state X_h, while t_h specifies its starting time on the flow trajectory and thus the remaining refinement interval. The flow time is estimated from the pooled representation $\bar{H}_h$ and constrained to $[t_{\min}, t_{\max}]$ through a sigmoid function.

During training, X_h and t_h are supervised using geometric projection targets in the normalized mel space. Let P_z and Y_z denote the normalized prior and target mel-spectrogram, respectively. We project P_z onto the direction defined by Y_z:

$$t^* = \mathrm{clip}\left(\frac{\langle P_z, Y_z\rangle}{\|Y_z\|_2^2}, t_{\min}, t_{\max}\right), \quad X_h^* = \mu_m + \sigma_m \cdot t^* Y_z. \tag{6}$$

Here, μ_m and σ_m denote the mean and standard deviation used for mel-spectrogram normalization, respectively. The state objective $\mathcal{L}_{\text{state}} = \|X_h - X_h^*\|_1$ encourages the predicted intermediate state to match its geometric target, while the time objective $\mathcal{L}_{\text{time}} = \|t_h - t^*\|_2^2$ aligns the predicted flow time with its corresponding target.

Given the intermediate state X_h and its associated flow time t_h, we formulate conditional flow matching over the remaining trajectory from $X(t_h) = X_h$ to the target mel-spectrogram $X(1) = Y$. Unlike standard CFM, which typically models a complete trajectory from a simple initial distribution (Le et al. 2023), our decoder starts from the intermediate state and models only the interval $[t_h, 1]$. During training, we construct a local linear path over this interval. We sample $u \sim \mathcal{U}(0, 1)$, map it to the flow time as $t = t_h + (1 - t_h)u$, and define the corresponding state as $X_t = (1 - u)X_h + uY$. Since $u = (t - t_h)/(1 - t_h)$, differentiating the local linear path with respect to t yields the target velocity $v^* = (Y - X_h)/(1 - t_h)$. The conditional flow decoder predicts the velocity field from the current acoustic state X_t, the flow time t, and the frame-level condition C. The flow objective is defined as:

$$\mathcal{L}_{\text{flow}} = \mathrm{SmoothL1}\left(v_\theta(X_t, t, C), v^*\right). \tag{7}$$

The flow objective $\mathcal{L}_{\text{flow}}$, together with $\mathcal{L}_{\text{state}}$ and $\mathcal{L}_{\text{time}}$, constitutes the adaptive projection and refinement objective $\mathcal{L}_{\text{arp}}$. At inference, the model first predicts (X_h, t_h) from the prior P and condition C. Starting from $X(t_h) = X_h$, the final mel-spectrogram is obtained by:

$$\hat{Y} = X(1) = X_h + \int_{t_h}^{1} v_\theta(X(t), t, C)\, dt. \tag{8}$$

This formulation converts full-trajectory generation into trajectory completion from an adaptive intermediate acoustic state. By modeling only the remaining transition from (X_h, t_h) to Y, the formulation encourages the flow decoder to allocate its capacity to acoustic naturalness and emotional detail while retaining the linguistic structure established by the hierarchical prior.

Shared-Specific Multimodal Emotion Modeling

Multimodal Emotional Prompt Decomposition. To support flexible multimodal emotion control, HiP-ETTS uses speech-derived emotion representations available during training as supervision to align textual and visual prompts with the acoustic emotion space through $\mathcal{L}_{\text{align}}$, which combines an L1 loss and a cosine loss. For each utterance, a VAD extractor[1] and an Emotion2Vec encoder[2] provide a spherical emotion representation and a high-dimensional emotion embedding, respectively. The former is encoded as the emotion-shared factor $\mathcal{E}_{sh}$, while the latter provides a richer representation of the reference speech. We therefore estimate the shared component of the high-dimensional representation explained by $\mathcal{E}_{sh}$ and treat the remaining residual as the complementary emotion factor:

$$\mathcal{E}_{sp} = \mathrm{LN}\left(f_{\text{emo}}(\mathbf{e}) - \mathrm{sg}\left(f_{v\to h}(\mathcal{E}_{sh})\right)\right), \tag{9}$$

$$\mathcal{L}_{\text{hVAD}} = \|f_{v\to h}(\mathcal{E}_{sh}) - \mathrm{sg}\left(f_{\text{emo}}(\mathbf{e})\right)\|_2^2. \tag{10}$$

Here, $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. The alignment objective $\mathcal{L}_{\text{hVAD}}$ encourages $f_{v\to h}$ to estimate the component of the high-dimensional emotion embedding explained by $\mathcal{E}_{sh}$, leaving the residual $\mathcal{E}_{sp}$ to provide complementary expressive information beyond the VAD-conditioned prediction.

Two-Stream Emotion Prosody Predictor. Prosody is a primary carrier of emotional expression in speech. Given the emotion-shared and complementary emotion factors, we introduce a two-stream emotion prosody predictor to generate segment-level prosody codes for acoustic synthesis. The shared stream models cross-speaker emotional prosodic structure, while the specific stream predicts a residual correction conditioned on the complementary factor and target speaker. We first use a VQ-based prosody encoder (Jiang et al. 2024) to extract reference prosody codes from the lowest 20 mel bins, which emphasize prosody over phonetic details, as training targets. The predictor models the code sequence in a GPT-style autoregressive manner:

$$p(\mathbf{c}_{1:S} \mid \Omega) = \prod_{s=1}^{S} p(c_s \mid c_{<s}, \Omega). \tag{11}$$

[1] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

[2] https://github.com/ddlBoJack/emotion2vec

Here, $\Omega = (r_{1:S}, E_{\text{sh}}, E_{\text{sp}}, z_s)$ denotes the conditioning tuple comprising the linguistic context, emotion-shared factor, complementary emotion factor, and speaker condition.

For each segment, the shared stream predicts a cross-speaker emotional prosodic backbone from the linguistic context and emotion-shared factor, while the specific stream predicts a residual correction conditioned additionally on the complementary emotion factor and speaker condition. The resulting logits and training objective are:

$$\ell_s^{sh} = F_{sh}(c_{<s}, \mathbf{r}_{1:S}, \mathcal{E}_{sh}), \ell_s^{sp} = F_{sp}(c_{<s}, \mathbf{r}_{1:S}, \mathcal{E}_{sp}, \mathbf{z}_s),$$
$$\mathcal{L}_{\text{pd}} = \sum_{s=1}^{S} \left[\text{CE}\left(c_s, \ell_s^{sh} + \alpha_{sp}\ell_s^{sp}\right) + \eta\, \text{CE}(c_s, \ell_s^{sh}) \right]. \tag{12}$$

where $\mathbf{r}_{1:S}$ denotes the segment-level context representation, the learnable coefficient α_{sp} controls the contribution of the specific stream, and η weights the auxiliary supervision on the shared stream.

Training Objectives. HiP-ETTS is trained in stages with the overall objective $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{hVAD}} + \mathcal{L}_{\text{orth}} + \mathcal{L}_{\text{pd}} + \mathcal{L}_{\text{hep}} + \mathcal{L}_{\text{arp}}$. Here, $\mathcal{L}_{\text{align}}$ aligns textual and visual prompts with the audio-derived emotion space, while $\mathcal{L}_{\text{hVAD}}$ supervises the shared-specific decomposition of the high-dimensional emotion representation. The cosine-similarity regularizer $\mathcal{L}_{\text{orth}}$ further reduces entanglement between timbre and emotion. $\mathcal{L}_{\text{pd}}$ trains the two-stream emotion prosody predictor, $\mathcal{L}_{\text{hep}}$ supervises the construction of the hierarchical emotional prior, and $\mathcal{L}_{\text{arp}}$ optimizes adaptive state projection and subsequent flow refinement.

Experiments

Experiment Setup

Dataset and Training Details. We use LibriTTS-clean (Zen et al. 2019) for base model pre-training and our proposed high-quality emotional speech dataset for emotional adaptation. Detailed statistics of the emotional speech dataset are provided in Appendix A of the Supplementary Document. Training is conducted in two stages. The base model is first trained on LibriTTS-clean for 350k steps using dynamic batching and then fine-tuned on the emotional dataset for 150k steps. The prosody predictor is trained separately for 6k steps. All experiments are performed on two NVIDIA 80GB A100 GPUs.

The content encoder is based on Conformer (Gulati et al. 2020) and consists of five layers with an embedding dimension of 512. The timbre encoder is based on a CAM++ (Wang et al. 2023) model. The prosody predictor adopts a two-stream architecture, with each stream comprising eight causal Transformer layers with a hidden dimension of 512. Finally, HiFi-GAN (Kong, Kim, and Bae 2020) is used to convert the mel-spectrograms into speech waveforms. We set the anchor strength to $\lambda_a = 0.30$, constrain the adaptive flow time to $[t_{\min}, t_{\max}] = [0.25, 0.90]$, and set the shared-stream loss weight to $\eta = 0.30$.

Evaluation Metrics. Traditional TTS evaluation commonly relies on subjective listening tests, whose results may vary with evaluation criteria, experimental settings, and individual listeners, making direct comparisons across studies difficult (Kirkland et al. 2023; Yang et al. 2026b). We therefore adopt a multidimensional automatic evaluation protocol. For content fidelity, we compute corpus-level word error rate (**WER**) using Whisper-Large V3[3]. Emotional consistency is evaluated from two complementary perspectives: the mean concordance correlation coefficient of VAD (**VAD-CCC**) measures consistency in the continuous emotion space, while **Emo-SIM** computes the cosine similarity between the emotion2vec (Ma et al. 2024) embeddings of generated speech and reference samples. Since speaker embeddings may be affected by prosodic variations, following Du et al. (Du et al. 2026), we evaluate speaker consistency (**Spk-SIM**) by comparing the embeddings of paired neutral and emotion-conditioned utterances generated under the same conditions. We use **UTMOS** (Saeki et al. 2022) to estimate acoustic quality, and **AutoPCP** (Barrault et al. 2023) together with the F0 Pearson correlation coefficient (**F0-PCC**) to assess prosodic consistency. To evaluate computational efficiency, we further report the total number of parameters (**Params**), real-time factor (**RTF**), and peak GPU memory usage (**Peak Mem.**), which measure model size, inference speed, and memory efficiency, respectively.

Baselines. We compare HiP-ETTS with representative and available zero-shot emotional TTS methods according to their emotion control conditions. For text prompts, we select the LLM-based EmoVoice (Yang et al. 2025a) and CosyVoice2 (Du et al. 2024). For audio prompts, we compare with MaskGCT (Wang et al. 2025b) and SparkTTS (Wang et al. 2025a), both of which can transfer the emotional style of reference speech. For facial prompts, we compare with FaceTTS (Lee, Chung, and Chung 2023), a diffusion-based TTS model conditioned on facial images. All baselines are evaluated using their officially released checkpoints and inference settings. Through these comparisons, we evaluate the emotional consistency and content fidelity of HiP-ETTS across different control modalities.

Performance Comparison

We compare HiP-ETTS with representative previous models for each emotion prompt, as shown in Table 1, with a focus on content fidelity and emotion consistency.

Content Fidelity: HiP-ETTS achieves lower or comparable WER across all emotion prompts. Compared with the next-best method, HiP-ETTS yields relative WER reductions of 55% and 79% for audio and visual prompts, respectively. This is attributed to its content-stable hierarchical emotional prior, which controllably incorporates emotional information while preserving linguistic content, thereby reducing prompt-induced interference. Under the text-prompt condition, HiP-ETTS achieves a WER comparable to that of CosyVoice2. The slight WER advantage of CosyVoice2 may stem from its large-scale text-conditioned training. However, the stronger emotion consistency of HiP-ETTS reflects a more favorable trade-off between content fidelity and emotional expression.

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

Prompt	Model	WER↓	VAD-CCC↑	Emo-SIM↑	Spk-SIM↑	UTMOS↑	AutoPCP↑	F0-PCC↑
–	GT	2.57%	–	–	–	3.156	–	–
Audio	**HiP-ETTS**	**2.98%**	**0.760**	**0.768**	**0.723**	3.297	**3.282**	0.433
	MaskGCT	6.59%	0.676	0.666	0.535	3.283	3.232	0.389
	SparkTTS	9.27%	0.641	0.611	0.527	**3.889**	3.081	**0.445**
Text	**HiP-ETTS**	2.99%	**0.730**	**0.669**	0.744	3.313	**3.257**	**0.451**
	EmoVoice	5.96%	0.434	0.584	**0.761**	3.774	3.228	0.371
	CosyVoice2	**2.56%**	0.383	0.587	0.687	**3.817**	3.178	0.400
Visual	**HiP-ETTS**	**3.11%**	**0.439**	**0.518**	**0.730**	**3.140**	**2.988**	**0.505**
	FaceTTS	14.50%	0.179	0.416	–	2.293	2.151	0.190

Table 1: Performance comparison of zero-shot emotional TTS models. "GT" refers to real speech samples. The best results are shown in bold, and the second-best results are underlined.

Model	Params↓	RTF↓	Peak Mem.↓
HiP-ETTS	236.9M	0.21	1.3 GiB
EmoVoice	675.0M	0.44	3.9 GiB
CosyVoice2	775.3M	0.64	2.6 GiB
MaskGCT	1645.0M	0.90	8.5 GiB
SparkTTS	662.9M	0.92	3.9 GiB
FaceTTS	**68.6M**	**0.09**	**0.5 GiB**

Table 2: Model size and inference efficiency. All efficiency metrics are measured under the same inference setting. The best results are shown in bold, and the second-best results are underlined.

Emotion Consistency: Across all prompts, HiP-ETTS achieves the highest VAD-CCC and Emo-SIM, demonstrating its effective modeling of continuous emotional states and fine-grained emotional expressions. Under audio prompt, which provide the richest emotional information, its VAD-CCC and Emo-SIM reach 0.760 and 0.768, respectively, clearly outperforming MaskGCT and SparkTTS and indicating more effective extraction of the target emotion from reference speech. For text and visual prompts, which lack direct acoustic emotion cues, HiP-ETTS also maintains strong performance, showing that its fine-grained emotion modeling can transform abstract semantic and static visual cues into effective emotional conditions for speech generation. These advantages are also reflected in the prosody that carries emotional expression. HiP-ETTS achieves the highest AutoPCP across all prompts and the highest F0-PCC under text and visual prompts, indicating better preservation of overall rhythm, stress, and intonation. Although SparkTTS obtains a slightly higher F0-PCC under audio prompts, HiP-ETTS performs better on AutoPCP, suggesting that our model captures the overall prosodic structure rather than merely reproducing frame-level pitch contours.

Speaker Consistency and Acoustic Quality: Under audio and text prompts, HiP-ETTS achieves the best or near-best Spk-SIM, indicating that speaker identity remains stable during emotional modulation. Since FaceTTS cannot vary emotion independently while keeping other conditions fixed, we do not report its Spk-SIM. HiP-ETTS does not show a clear overall advantage in UTMOS (we note that, among these baselines, higher UTMOS scores do not necessarily translate into better performance in content, emotion, or prosody). Compared with baselines that benefit from larger model scales and broader pretraining, HiP-ETTS obtains slightly lower UTMOS scores. Nevertheless, our primary objective is to achieve more expressive emotion control while maintaining low WER, and Hip-ETTS demonstrates clear advantages in content fidelity, emotional consistency, and prosodic modeling.

Inference Efficiency: As shown in Table 2, HiP-ETTS contains 236.9M parameters, achieves an RTF of 0.21, and requires only 1.3 GiB of peak GPU memory. Compared with the audio- and text-conditioned baselines, HiP-ETTS reduces the parameter count by 64%–86%, provides 2.2–4.5$\times$ faster inference, and substantially lowers memory consumption. Although FaceTTS is smaller and faster, it supports only visual prompting and performs considerably worse in content, emotion, and prosody metrics. Overall, HiP-ETTS supports unified multimodal emotion control at a relatively low computational cost while maintaining strong generation performance.

Ablation Study

We conduct ablation experiments on the key components of HiP-ETTS, with the results shown in Table 3. We discuss the results by component, focusing primarily on their effects on content fidelity and emotional consistency.

Emotion Prompt Decomposition. To analyze the role of the complementary emotion factor, we replace $\mathcal{E}_{sp}$ with that extracted from another speaker's utterance with the same emotion category and similar VAD values, thereby minimizing the effects of category and intensity conflicts. This intervention causes only minor degradation in WER and VAD-CCC, whereas Emo-SIM drops markedly from 0.768 to 0.707. This suggests that, even when the emotion category and continuous emotional state are preserved, a mismatched complementary emotion factor still weakens the reproduction of the target emotional expression. These results suggest that the complementary emotion factor provides reference-dependent expressive cues beyond the VAD-conditioned pre-

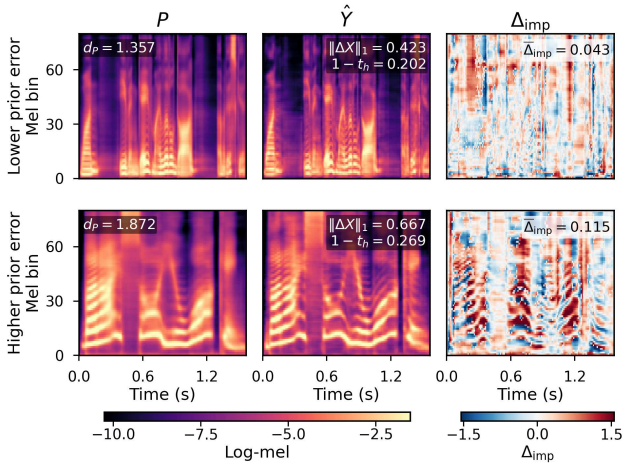


Figure 2: Representative cases near the 25th and 75th percentiles of prior error. The higher-error prior receives a larger state correction $\Delta X = X_h - P$ and a longer remaining flow interval $1 - t_h$. Positive Δ_{imp} indicates regions where refinement brings the output closer to the target.

diction, while the emotion-shared factor primarily represents the overall emotional state.

Two-Stream Emotion Prosody Predictor. To evaluate the two-stream emotion prosody predictor, we replace it with a separately trained single-stream predictor that jointly receives all emotion factors. Under the same budget, the single-stream variant only moderately increases WER from 2.98% to 3.51%, but substantially reduces VAD-CCC and Emo-SIM from 0.760 and 0.768 to 0.439 and 0.566, respectively. This suggests that the single stream preserves linguistic content but struggles to jointly capture continuous emotional states and fine-grained emotional expressions. In contrast, the two-stream design assigns cross-speaker emotional prosodic structure and complementary expressive variations to complementary streams, allowing the shared and specific components of the factorized emotion representation to be modeled at appropriate levels.

Hierarchical Emotional Prior. To examine the contribution of the emotion residual offset in the hierarchical prior, we remove Δ_e while keeping all other emotion pathways unchanged. This intervention causes only minor degradation in WER and VAD-CCC, whereas Emo-SIM drops markedly from 0.768 to 0.674. VAD-CCC mainly reflects coarse-grained global acoustic tendencies of emotion, such as overall energy, speaking rate, and pitch patterns, which can still be conveyed through the remaining emotion pathways. In contrast, Emo-SIM is more sensitive to fine-grained acoustic realizations. Its substantial decrease therefore indicates that injecting the emotion residual offset into the hierarchical prior contributes target-specific acoustic emotion details that cannot be fully recovered through the remaining pathways. The limited change in WER further suggests that this residual pathway enhances emotional realization while preserving linguistic content.

Variant	WER↓	VAD-CCC↑	Emo-SIM↑
Full	2.98%	**0.760**	**0.768**
Shuffled-Specific$^{\text{E}}$	3.13%	0.750	0.707
Single-Stream$^{\text{P}}$	3.51%	0.439	0.566
w/o Emo. Residual$^{\text{H}}$	3.11%	0.754	0.674
Fixed Time$^{\text{A}}$	3.23%	0.746	0.709
No State Offset$^{\text{A}}$	2.98%	0.728	0.666
Prior-Only$^{\text{F}}$	**2.62%**	0.653	0.663

Table 3: Ablation results of HiP-ETTS in terms of content fidelity and emotional consistency. Except for the separately trained Single-Stream variant, all ablations are implemented as inference-time interventions. The superscript attached to each variant identifies the component group being examined: E for emotion prompt decomposition, P for two-stream emotion prosody predictor, H for hierarchical emotional prior, A for adaptive projection, and F for flow refinement.

Adaptive Projection. To disentangle the roles of state and time prediction, we introduce two inference-time interventions. Fixed Time replaces the adaptive t_h with its test-set mean value while retaining X_h, whereas No State Offset sets $X_h = P$ while retaining t_h. Fixed Time increases WER from 2.98% to 3.23% and reduces Emo-SIM from 0.768 to 0.709, indicating that a uniform refinement interval cannot accommodate sample-dependent refinement demands. Removing the state offset further reduces Emo-SIM to 0.666 while leaving WER nearly unchanged, suggesting that shallow-state projection primarily corrects the prior state to facilitate subsequent emotional acoustic refinement without compromising linguistic content. Figure 2 further illustrates their joint adaptive behavior using representative cases near the 25th and 75th percentiles of prior error: the higher-error prior receives both a larger state correction and a longer remaining flow interval. Overall, shallow-state projection determines the acoustic starting state, while time prediction controls the remaining refinement interval.

Flow Refinement. To examine the contribution of the refinement stage as a whole, we directly feed the hierarchical prior P into the vocoder, bypassing both adaptive projection and flow generation. Prior-Only achieves a lower WER of 2.62%, but its VAD-CCC and Emo-SIM decrease from 0.760 and 0.768 to 0.653 and 0.663, respectively. This result shows that the hierarchical prior already provides a clear and content-stable acoustic structure, but remains conservative in emotional realization. The subsequent projection and flow refinement substantially enriches continuous emotion states and fine-grained emotional expression, while maintaining a low WER of 2.98%. Thus, the prior primarily anchors linguistic structure, whereas the downstream refinement stage converts it into a more expressive acoustic realization.

Conclusion

We presented HiP-ETTS, a multimodal emotional TTS framework that decomposes prompt-derived emotion into emotion-shared and complementary emotion factors for fine-

grained emotion modeling. By combining a content-stable hierarchical prior with adaptive state projection and flow refinement, HiP-ETTS improves emotional consistency while preserving linguistic content across different prompt modalities. Overall, HiP-ETTS provides an efficient solution for expressive multimodal emotional speech synthesis without sacrificing linguistic fidelity. Future work will examine robustness to noisy or adversarial multimodal cues while preserving content fidelity and emotion consistency.

References

Banse, R.; and Scherer, K. R. 1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3): 614.

Barrault, L.; Chung, Y.-A.; Meglioli, M. C.; Dale, D.; Dong, N.; Duppenthaler, M.; Duquenne, P.-A.; Ellis, B.; Elsahar, H.; Haaheim, J.; et al. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv preprint arXiv:2312.05187*.

Chen, H.; Chen, R.; and Hirschberg, J. 2024. Emoknob: Enhance voice cloning with fine-grained emotion control. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8170–8180.

Cho, D.-H.; Oh, H.-S.; Kim, S.-B.; Lee, S.-H.; and Lee, S.-W. 2024. Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. *arXiv preprint arXiv:2406.07803*.

Cho, D.-H.; Oh, H.-S.; Kim, S.-B.; and Lee, S.-W. 2025. EmoSphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Transactions on Affective Computing*.

Choi, H.-Y.; and Park, J. 2025. VoicePrompter: Robust Zero-Shot Voice Conversion with Voice Prompt and Conditional Flow Matching. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Du, H.; Shi, J.; Lu, S.; Zhou, G.; and Gao, Y. 2026. Sparse Autoencoders for Interpretable Emotion Control in Text-to-Speech. *arXiv preprint arXiv:2606.01479*.

Du, Z.; Wang, Y.; Chen, Q.; Shi, X.; Lv, X.; Zhao, T.; Gao, Z.; Yang, Y.; Gao, C.; Wang, H.; et al. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Guan, W.; Li, Y.; Li, T.; Huang, H.; Wang, F.; Lin, J.; Huang, L.; Li, L.; and Hong, Q. 2024. Mm-tts: Multi-modal prompt based style transfer for expressive text-to-speech synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18117–18125.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 1122–1131.

Hieu, N. H. N.; Nguyen, N. S.; Dang, H. N.; Vo, T.; Hy, T.-S.; and Nguyen, V. 2025. OZSpeech: One-step zero-shot speech synthesis with learned-prior-conditioned flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 21500–21517.

Huang, K.-P.; Liu, H.; Peng, P.; Wu, H.; Ni, Z.; yi Lee, H.; Lee, J.; and Chachra, N. 2026. EmoRES-TTS: Residual-Enhanced Vector Steering for Emotional Speech Generation. arXiv:2609.38157.

Jiang, Z.; Liu, J.; Ren, Y.; He, J.; Ye, Z.; Ji, S.; Yang, Q.; Zhang, C.; Wei, P.; Wang, C.; et al. 2024. Mega-tts 2: Boosting prompting mechanisms for zero-shot speech synthesis. In *International Conference on Learning Representations*, volume 2024, 57919–57939.

Ju, Z.; Wang, Y.; Shen, K.; Tan, X.; Xin, D.; Yang, D.; Liu, Y.; Leng, Y.; Song, K.; Tang, S.; et al. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Kirkland, A.; Mehta, S.; Lameris, H.; Henter, G. E.; Szekely, E.; and Gustafson, J. 2023. Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation. In *12th Speech Synthesis Workshop (SSW) 2023*.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33: 17022–17033.

Laukka, P.; Neiberg, D.; and Elfenbein, H. A. 2014. Evidence for cultural dialects in vocal emotion expression: acoustic classification within and across five nations. *Emotion*, 14(3): 445.

Le, M.; Vyas, A.; Shi, B.; Karrer, B.; Sari, L.; Moritz, R.; Williamson, M.; Manohar, V.; Adi, Y.; Mahadeokar, J.; et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36: 14005–14034.

Lee, J.; Chung, J. S.; and Chung, S.-W. 2023. Imaginary voice: Face-styled diffusion model for text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Lee, Y.; Lee, S.-Y.; and Rabiee, A. 2017. Emotional End-to-End Neural Speech Synthesizer. In *NIPS2017*. Neural Information Processing Systems Foundation.

Li, H.; Li, Y.; Wang, X.; Hu, J.; Xie, Q.; Yang, S.; and Xie, L. 2025. Flespeech: Flexibly controllable speech generation with various prompts. *arXiv preprint arXiv:2501.04644*.

Ma, Z.; Zheng, Z.; Ye, J.; Li, J.; Gao, Z.; Zhang, S.; and Chen, X. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, 15747–15760.

Mehta, S.; Tu, R.; Beskow, J.; Székely, É.; and Henter, G. E. 2024. Matcha-TTS: A fast TTS architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11341–11345. IEEE.

Ning, Z.; Xie, Q.; Zhu, P.; Wang, Z.; Xue, L.; Yao, J.; Xie, L.; and Bi, M. 2023. Expressive-vc: Highly expressive voice conversion with attention fusion of bottleneck and perturbation features. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Patel, S.; Scherer, K. R.; Björkner, E.; and Sundberg, J. 2011. Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, 87(1): 93–98.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; and Kudinov, M. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International conference on machine learning*, 8599–8608. PMLR.

Ren, Y.; Wang, T.; Yi, J.; Xu, L.; Tao, J.; Zhang, C. Y.; and Zhou, J. 2024. Fewer-token neural speech codec with time-invariant codes. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12737–12741. IEEE.

Russell, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161.

Saeki, T.; Xin, D.; Nakata, W.; Koriyama, T.; Takamichi, S.; and Saruwatari, H. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Schuller, B. W. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5): 90–99.

Sobin, C.; and Alpert, M. 1999. Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy. *Journal of psycholinguistic research*, 28(4): 347–365.

Tan, X.; Chen, J.; Liu, H.; Cong, J.; Zhang, C.; Liu, Y.; Wang, X.; Leng, Y.; Yi, Y.; He, L.; et al. 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6): 4234–4245.

Wang, H.; Zheng, S.; Chen, Y.; Cheng, L.; and Chen, Q. 2023. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*.

Wang, X.; Jiang, M.; Ma, Z.; Zhang, Z.; Liu, S.; Li, L.; Liang, Z.; Zheng, Q.; Wang, R.; Feng, X.; et al. 2025a. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Wang, Y.; Zhan, H.; Liu, L.; Zeng, R.; Guo, H.; Zheng, J.; Zhang, Q.; Zhang, X.; Zhang, S.; and Wu, Z. 2025b. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. In *International Conference on Learning Representations*, volume 2025, 47127–47150.

Wu, Z.; Kang, Y.; Cao, S.; Ma, L.; Li, Q.; and Yang, Q. 2025. MPE-TTS: Customized Emotion Zero-Shot Text-To-Speech Using Multi-Modal Prompt. *arXiv preprint arXiv:2505.18453*.

Xie, T.; Rong, Y.; Zhang, P.; Wang, W.; and Liu, L. 2025. Towards controllable speech synthesis in the era of large language models: A systematic survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 764–791.

Yang, D.; Cai, Y.; Saito, Y.; Wang, L.; and Saruwatari, H. 2026a. Shallow Flow Matching for Coarse-to-Fine Text-to-Speech Synthesis. *Advances in Neural Information Processing Systems*, 38: 76586–76618.

Yang, D.; Liu, S.; Huang, R.; Weng, C.; and Meng, H. 2024. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 2913–2925.

Yang, G.; Yang, C.; Chen, Q.; Ma, Z.; Chen, W.; Wang, W.; Wang, T.; Yang, Y.; Niu, Z.; Liu, W.; et al. 2025a. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 10748–10757.

Yang, J.; Li, S.; Shinozaki, T.; Saito, Y.; and Saruwatari, H. 2025b. Emotional Text-To-Speech Based on Mutual-Information-Guided Emotion-Timbre Disentanglement. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 567–572. IEEE.

Yang, Q.; Liu, Z.; Wang, J.; Du, Y.; Huang, P.; and Xiao, T. 2025c. RLAIF-SPA: Optimizing LLM-based Emotional Speech Synthesis via RLAIF. *arXiv preprint arXiv:2510.14628*.

Yang, Y.; Wang, H.; Han, B.; Liu, S.; Li, J.; Qin, Y.; and Chen, X. 2026b. Position: Towards Responsible Evaluation for Text-to-Speech. In *Forty-third International Conference on Machine Learning Position Paper Track*.

Yao, J.; Yang, Y.; Lei, Y.; Ning, Z.; Hu, Y.; Pan, Y.; Yin, J.; Zhou, H.; Lu, H.; and Xie, L. 2024. Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10571–10575. IEEE.

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

Zhang, G.; Qin, Y.; Zhang, W.; Wu, J.; Li, M.; Gai, Y.; Jiang, F.; and Lee, T. 2023. iEmoTTS: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 1693–1705.

Zhang, T.-H.; Zhang, J.; Wang, J.; Qian, X.; and Yin, X.-C. 2025. Facespeak: Expressive and high-quality speech synthesis from human portraits of different styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25922–25930.