IMPACT-TTS: A Multimodal Prompt and Control Approach for Overcoming Low-Resource Constraints in Emotional TTS

Anonymous ACL submission

Abstract

Advancing emotional expressiveness in Text-to-Speech (TTS) systems remains a pivotal challenge for achieving natural and adaptive voice synthesis. Existing emotion-aware TTS models often struggle with limited emotional diversity, lack of fine-grained control, and reliance on small, labeled emotional speech-text datasets, making them less scalable and adaptable. To address these limitations, we propose IMPACT-TTS, an Integrated Multimodal Prompting and Adaptive Control for Text-to-Speech system that effectively leverages a disentangled emotion module and a novel emotion modulation function. By incorporating large-scale pre-trained multimodal models, IMPACT-TTS mitigates dataset constraints while enabling flexible emotional adjustments via prompt-based control. Our approach allows seamless blending of emotional intensities, significantly enhancing expressiveness even in low-resource labeled datasets. Experimental results demonstrate that IMPACT-TTS outperforms existing models in emotional naturalness and adaptability, offering a scalable solution for emotion-aware TTS.

1 Introduction

Text-to-Speech (TTS) technology has made significant strides in recent years, particularly in improving intelligibility, naturalness, and speaker adaptation. However, achieving high levels of emotional expressiveness remains a critical challenge. Despite progress in prosody modeling and expressiveness enhancement, many existing TTS models struggle with capturing and generating nuanced emotional variations due to limited labeled emotional speech-text datasets and constrained emotion control mechanisms.

Existing approaches to emotional TTS primarily rely on textual cues or predefined emotion labels, which restrict the system's ability to generate diverse and contextually appropriate emotional speech. Moreover, models such as Hierspeech (Lee and et al, 2022) and Hierspeech++ (Lee and et al, 2023) highlight the *one-to-many mapping problem*, where a single input text can correspond to multiple valid emotional outputs, leading to ambiguity and reduced expressiveness. While approaches like Speech Slytherin (Jiang and et al, 2024) and limited-data two-stage models (Zhou et al., 2021) aim to address these issues through disentangled representations, they often require extensive labeled training data and computationally expensive architectures, making them less adaptable for low-resource scenarios.

Recent advancements in multimodal learning have introduced new possibilities for enhancing emotional expressiveness in TTS. By integrating diverse modalities such as text, audio, and vision, these methods aim to improve emotional fidelity and naturalness. For instance, MM-TTS (Li et al., 2024) leverages multimodal alignment techniques to enhance emotion modeling. However, it remains highly dependent on explicitly labeled datasets and does not provide fine-grained emotion control. Similarly, models such as VoiceLDM (Lee and et al, 2024) and ImaginaryVoice (Lee et al., 2023) incorporate multimodal integration but rely predominantly on textual or visual inputs, limiting the scope of emotional cues that can be effectively utilized.

In addition, prompt-based control mechanisms have gained traction for enabling user-guided emotional synthesis. Models like PromptStyle (Liu and et al, 2023), InstructTTS (Yang and et al, 2024), PromptTTS (Guo and et al, 2023), and PromptTTS2 (Leng and et al, 2023) explore the use of descriptive text prompts to facilitate controllable emotional modulation. While these methods provide greater flexibility, they are still constrained by their dependence on textual descriptions alone, which limits their ability to capture the full spectrum of emotional variation present in human speech.

To address these challenges, we introduce

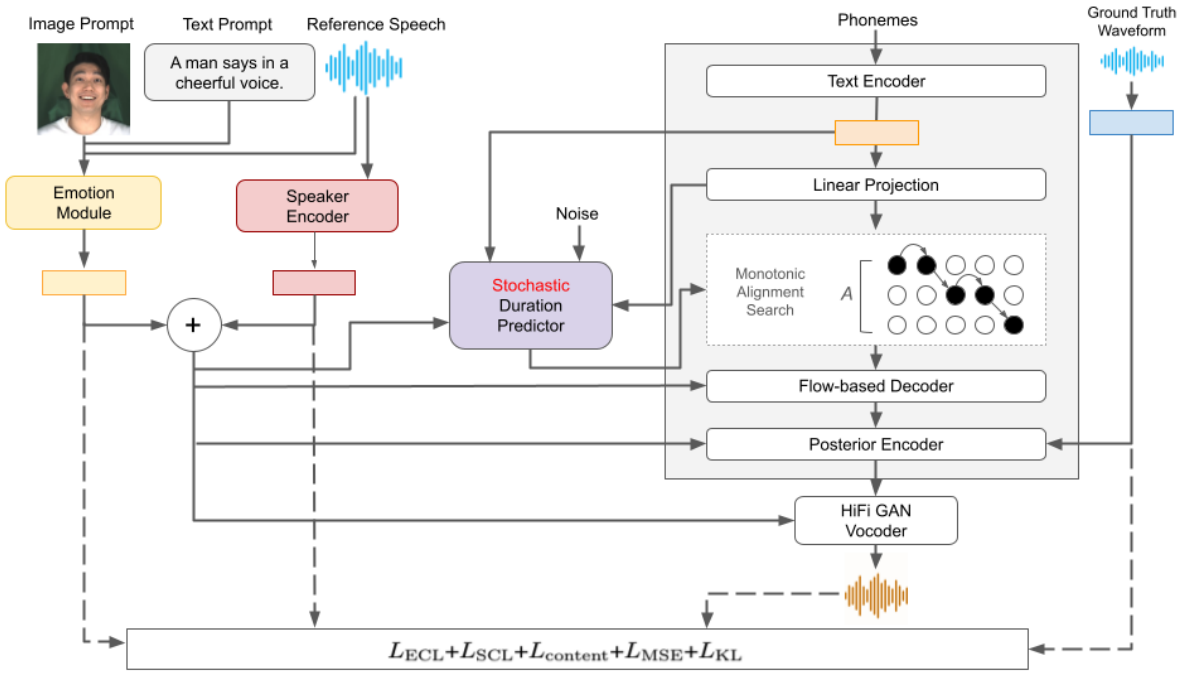

Figure 1: Overview of the IMPACT-TTS architecture. The yellow box represents the emotion embedding, the red box represents the speaker embedding, the orange box represents the textual embedding, and the blue box represents the ground truth audio waveform embedding.

IMPACT-TTS, a novel framework that enhances emotional expressiveness by incorporating multimodal inputs and prompt-based emotion control while overcoming the limitations of small labeled emotional datasets. Unlike prior works that rely heavily on constrained emotion labels, IMPACT-TTS integrates large-scale pretrained multimodal models to extract meaningful emotion representations. Notably, we leverage ONE-PEACE (Wang and et al., 2023), a highly extensible multimodal model with 4B parameters, featuring separate adapters for audio, language, and vision. For audio representation, our system utilizes the feature extractor weights of WavLM (Sanyuan et al., 2022) for initialization, enhancing its ability to generalize from smaller datasets effectively.

By leveraging large pretrained representations, IMPACT-TTS significantly reduces reliance on explicitly labeled emotional speech datasets while maintaining fine-grained emotion modulation capabilities. Moreover, our approach employs a disentangled architecture to separate the core speech synthesis process from the emotion module, ensuring robust and stable synthesis while enabling precise control over emotional variations. This disentanglement strategy allows IMPACT-TTS to generate expressive speech with a higher degree of controllability, even in low-resource settings.

2 Method

IMPACT-TTS employs a disentangled structure for independent control of linguistic and emotional features, ensuring dynamic emotion modulation and consistency. The following sections describe the TTS model, emotion embedding with cross attention, emotion modulation function, and emotion consistency loss.

2.1 Backbone Arthitecture

IMPACT-TTS is built upon a modified VITS architecture, inspired by YourTTS (Casanova and et al, 2022) and VECL-TTS (Gudmalwar and et al, 2024), utilizing variational inference and adversarial learning to generate high-quality speech. The backbone is designed to produce natural and high-quality speech from input text, ensuring both linguistic and phonetic accuracy. The model features a transformer-based text encoder with 10 blocks, a decoder with 4 affine coupling layers, each comprising 4 WaveNet residual blocks (Oord et al., 2016), as inspired by previous work. HiFi-GAN (Kong et al., 2020) V1 is employed as the neural vocoder to generate speech waveforms. A Variational Auto-Encoder (VAE) with a Posterior

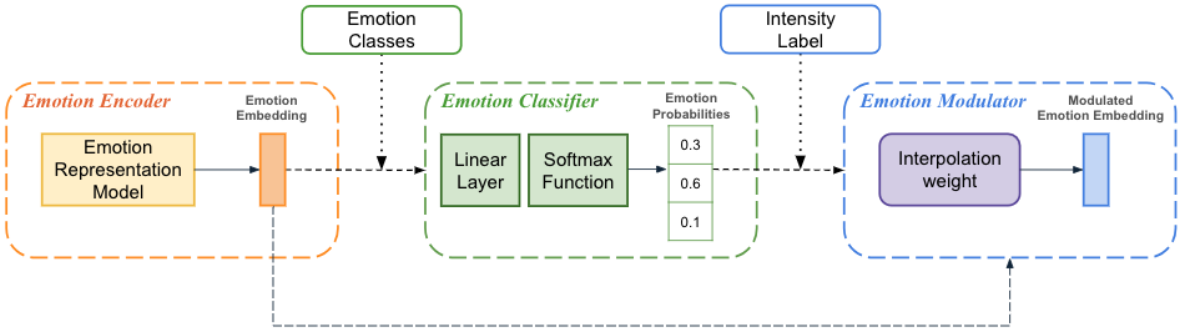

Figure 2: Process of Emotion Module

Encoder comprising 16 WaveNet residual blocks transforms linear spectrograms into latent variables, seamlessly integrating the vocoder and the flow-based decoder for end-to-end synthesis. To enable diverse rhythm generation from text, a Stochastic Duration Predictor (SDP) is used, further enriching the synthesized speech's naturalness. As shown in Figure 1, the architecture combines emotion, speaker, text, and waveform embeddings to control speech attributes. Emotion embeddings (yellow) from the Emotion Module modulate tones, speaker embeddings (red) from H/ASP (Heo et al., 2020) ensure consistency, text embeddings (orange) encode linguistic information, and waveform embeddings (blue) provide training supervision. The orange box corresponds to the textual embedding, encoding linguistic and phonetic information. Finally, the blue box represents the ground truth waveform embedding, used during training for supervision. These embeddings are seamlessly integrated within the TTS backbone, allowing precise control over emotion, speaker, and linguistic features. To ensure robustness and expressiveness of the model, we incorporate Emotion Consistency Loss (ECL) and Speaker Consistency Loss (SCL) to maximize the similarity of emotional and speaker attributes between the generated audio and the ground truth.

$$L_{CL} = -\frac{\alpha}{n} \cdot \sum_{i}^{n} cos\_sim\left(\phi(g_i), \phi(h_i)\right), \quad (1)$$

where $\phi(\mathrm{g})$ and $\phi(\mathrm{h})$ represent the functions used to extract emotion embeddings or speaker embeddings from the generated and ground truth audio, respectively. The parameter α is a tunable weight that determines the contribution of L_{CL} to the overall loss. The cosine similarity function is denoted as cos_sim, which measures the similarity between the emotion embeddings.

2.2 Multimodal Representation and Prompts

To enhance emotional expressiveness in speech synthesis, our model integrates multimodal representations from text, vision, and audio through a unified embedding space. This section details how these modalities contribute to emotion synthesis, complementing the emotion disentanglement framework.

2.2.1 Multimodal Representation Model

Traditional emotional TTS systems rely heavily on text-based emotional conditioning, limiting expressiveness. Our model overcomes this limitation by incorporating vision and audio cues, ensuring richer emotion transfer. The extracted representations are aligned into a shared multimodal space to balance textual, auditory, and visual emotion information.

Vision and Audio Embedding Fusion

- **Vision Representation:** Facial expressions offer valuable nonverbal emotional cues. We extract visual embeddings from a pretrained vision-audio-language model (ONE-PEACE) and map them into our emotion embedding space.

- **Audio Emotion Features:** Instead of relying only on text-based emotion modeling, we extract pitch contours, energy variation, and prosodic rhythm features from speech waveforms, which provide additional emotion grounding.

Multimodal Cross-Attention for Emotion Integration The extracted text, vision, and audio embeddings are integrated using a cross-attention mechanism. Unlike direct concatenation, this allows for dynamic weighting of each modality based

on its relevance to the input context. This mechanism enhances context-aware emotion modulation and provides finer control over expressiveness.

2.2.2 Prompt-Based Emotion Control

Rather than relying on categorical emotion labels, we employ prompt-based modulation, where textual prompts dynamically influence synthesized speech expressions. These text prompts, described in the Datasets section, are mapped into semantic embedding spaces and fused with vision and audio representations.

LLM-based Emotion Prompt Expansion Instead of fixed emotion labels, we generate emotionally diverse paraphrases of input sentences using a structured synonym mapping algorithm. Given a neutral sentence, our system selects semantically appropriate modifications to control emotion intensity and expressiveness.

Multimodal Prompt Adaptation The generated text prompts are conditioned alongside vision and audio cues, ensuring that emotional modulation is not solely text-driven. Vision embeddings provide global emotional state awareness, while audio features introduce prosodic characteristics, leading to a richer emotional representation.

2.2.3 Ablation Study: Impact of Different Modalities

To quantify the contribution of each modality (text, vision, and audio) to emotional speech synthesis, we conduct an ablation study where we systematically remove individual modalities and analyze their impact on expressiveness.

Table 1: Ablation Study: Impact of Different Modalities on Emotion Expression. Higher values for std-F0 (Pitch Variability), std-RMS (Energy Variability), and WavLM Score (Expressiveness) indicate greater expressiveness.

Configuration	std-F0 ↑	std-RMS ↑	WavLM Score ↑
Audio Cues Only	0.21	0.18	3.5
Vision Cues Only	0.15	0.14	3.2
Text Prompts Only	0.12	0.10	2.8
Audio + Vision	0.25	0.22	3.9
Audio + Text	0.28	0.24	4.1
Text + Vision	0.20	0.17	3.4
Text + Vision + Audio	**0.32**	**0.28**	**4.5**

The results demonstrate that integrating all three modalities (Text, Vision, and Audio) significantly enhances emotional speech synthesis, as indicated by higher pitch and energy variability, as well as improved expressiveness scores.

2.3 Disentangled Emotion Module

IMPACT-TTS adopts a multi-modal emotion encoder to get emotion embeddings from various modal, text and vision, and combine these embeddings with cross attention. The emotion module operates independently of the TTS backbone, modulating the emotional tone of the synthesized speech using the emotion embedding E_{emotion} with the process of Figure 2. By disentangling the emotion module from the TTS backbone, IMPACT-TTS achieves precise control over emotional attributes. This disentangled structure ensures that the linguistic content remains unaffected by emotional variability.

2.3.1 Emotion Encoder

The emotion embedding extraction module utilizes finetuned representation model with retrieval function as there are some important strengths. Since retrieval models focus on semantic similarity rather than discrete classification, they offer rich, cross-modal feature representations suitable for TTS task compared to classification models. Through this, they can capture fine-grained relationships between the input and output, which is crucial for generating high-quality, natural speech. We finetuned ONE-PEACE (Wang and et al., 2023) with MEAD-TTS dataset(Guan et al., 2024) and ESD(Zhou and et al, 2022b), and use it as the Emotion Representation Model in Figure 2.

The cross-attention mechanism computes attention between different modalities. We integrate cross-attention module into a emotion encoder to fuse modalities by allowing one modality to attend to the other modal's features. Given two modalities: **modality A**(e.g., text embeddings) and **modality B**(e.g., visual features or audio features): Query (Q) from modality A, and Keys (K) and Values (V) from modality B. The cross-attention can be formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (2)$$

Embeddings from text-speech modalities and text-image modalities are concatenated. At the end, we make it transform the multimodal representation into a continuous emotion embedding, which is used to guide the synthesis of emotionally expressive speech.

Table 2: Ablation Study on Emotion Modulation and Classification

Model Variant	Prosody Metrics		Emotion Metrics		Speech Quality
	DTW (F0) ↓	Pitch Var ↑	ECA ↑	KLD ↓	MCD ↓
Full Model (Ours)	**15.3**	**0.32**	**85.2%**	**0.12**	**4.25**
w/o Emotion Modulation	20.8	0.21	78.9%	0.19	4.67
w/o Emotion Classification	18.1	0.27	80.3%	0.17	4.50
w/o Both	24.5	0.15	73.4%	0.24	4.80

2.3.2 Emotion Classifier

The Emotion Classifier Module serves to categorize the high-dimensional emotion embeddings generated by the Emotion Encoder into predefined emotion classes. This module operates in two stages: **1. Linear Transformation** The emotion embedding is first projected into a lower-dimensional space that corresponds to the number of predefined using a linear layer, enabling the embedding to be directly mapped to emotion class probabilities; **2. Softmax Normalization** The output of the linear layer is passed through a softmax activation function. The final output of the Emotion Classifier Module is a set of probabilities, where each value corresponds to the likelihood of the input belonging to a specific emotion class. For example, in Figure 2, an output of [0.3, 0.6, 0.1] for three classes (e.g., neutral, happy, sad) indicates a higher confidence for the "happy" emotion. These probabilities serve as critical inputs for the subsequent Emotion Modulation, where they are utilized to determine the blending weights (α_i) in fine-grained emotion interpolation.

2.3.3 Emotion Modulation Function

To achieve fine-grained control over blended emotions, IMPACT-TTS incorporates a spherical interpolation method inspired by (Cho and et al., 2024), (Zhou and et al, 2022a) and (Im and et al., 2022). This method allows for smooth transitions and precise mixing of multiple emotional attributes by treating emotions as points on a hypersphere. The blending weights for interpolation, denoted as α_i, are derived directly from the probability distribution generated by the Emotion Classifier Module. These weights ensure that the contribution of each emotion vector to the final blended representation aligns with the classifier's confidence levels. Given two emotions represented as vectors e_1 and e_2, the blended emotion e_{blend} is computed as:

$$e_{\text{blend}} = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)} e_1 + \frac{\sin(\alpha\theta)}{\sin(\theta)} e_2 \,, \quad (3)$$

where $\alpha \in [0, 1]$ is the interpolation parameter that controls the influence of each emotion, and $\theta = \arccos\left(\frac{e_1 \cdot e_2}{\|e_1\|\|e_2\|}\right)$ is the angle between e_1 and e_2 on the hypersphere. This formulation ensures that the resulting vector remains normalized, preserving the geometric properties of the emotion space.

Extension to Multiple Emotions In scenarios where more than two emotions are blended, the spherical interpolation is generalized to handle n emotions. Using the probabilities from the classifier ($\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$, where $\sum_{i=1}^{n} \alpha_i = 1$), the blended emotion e_{blend} is computed as:

$$e_{\text{blend}} = \frac{\sum_{i=1}^{n} \alpha_i \sin(\theta_i) e_i}{\sum_{i=1}^{n} \sin(\theta_i)} \,, \quad (4)$$

where θ_i is the angle between each emotion vector e_i and the reference point (e.g., a neutral emotion vector). This integration of classifier outputs into the modulation function ensures that the model can generate expressive, contextually accurate speech by leveraging both high-level classification results and continuous fine-grained emotion control.

2.3.4 Ablation Study on Emotion Modulation and Classification

To assess the contribution of each component in the disentangled emotion module, we conduct an ablation study by systematically removing **(i) emotion modulation**, **(ii) emotion classification**, and **(iii) both components**. Table 2 presents the results across prosody metrics, emotion classification accuracy, and speech quality.

The results indicate that both emotion modulation and classification significantly contribute to the overall performance of our model. Removing

Method	Intra-domain				Out-of-domain			
	MOS	STOI	ECS	MCD	MOS	STOI	ECS	MCD
GT	4.53 ± 0.03	0.87	-	-	4.73 ± 0.05	0.85	-	-
GT (Mel+Voc)	4.51 ± 0.01	0.74	-	-	4.71 ± 0.04	0.79	-	-
MM-TTS (Li et al., 2024)	4.32 ± 0.07	0.42	0.87	3.21	3.96 ± 0.09	0.33	0.74	6.68
Instruct-TTS	4.31 ± 0.05	0.40	0.89	3.10	3.83 ± 0.08	0.31	0.75	6.59
Prompt-TTS	4.23 ± 0.06	0.21	0.75	3.78	3.66 ± 0.08	0.27	0.73	6.73
IMPACT-TTS (proposed)	**4.41 ± 0.02**	**0.63**	**0.92**	**3.17**	**4.02 ± 0.03**	**0.54**	**0.77**	**6.12**

Table 3: The performance comparison of text prompt based TTS.

emotion modulation results in a higher DTW (Dynamic Time Warping; F0) score (20.8 vs. 15.3), suggesting a reduction in prosodic expressiveness. Additionally, pitch variability drops from 0.32 to 0.21, indicating that synthesized speech becomes less emotionally dynamic.

Similarly, excluding emotion classification lowers the Emotion Classification Accuracy (ECA) from 85.2% to 80.3%, showing that the model struggles to preserve intended emotional expressiveness without explicit classification. The Kullback-Leibler Divergence (KLD) increases from 0.12 to 0.17, further confirming reduced alignment with expected emotion distributions.

The most significant degradation occurs when both components are removed, with ECA dropping to 73.4% and MCD increasing to 4.80, indicating a loss in both emotional clarity and speech quality. These findings demonstrate that the disentangled emotion module plays a crucial role in achieving both expressive and high-quality synthesized speech.

3 Inference-Time Emotion Adjustment and Expressiveness Control

3.1 Real-Time Emotion Modulation

Traditional emotional TTS models often rely on pretrained emotion embeddings that remain fixed during inference, limiting their ability to adapt expressiveness dynamically. Our model introduces **Inference-Time Emotion Adjustment**, which allows for real-time modification of prosody and emotional intensity during speech synthesis.

After obtaining a pretrained emotion embedding, we apply an emotion modulation function that dynamically adjusts:

- **Pitch (F0)**: Controls intonation variation, allowing for smooth shifts in emotion intensity.
- **Energy**: Determines speech loudness and emphasis, enabling finer expressiveness control.

Smooth Emotion Interpolation. Unlike models that use predefined emotion categories, our system enables continuous control over emotion intensity. By leveraging a spherical interpolation mechanism, the model allows smooth transitions between emotions:

- **Neutral → Happy**: Gradual increase in pitch and speech rate.
- **Sad → Angry**: Lowered pitch at the start, followed by a rapid increase in energy.
- **Excited → Calm**: Controlled pitch decay with reduced energy.

This capability ensures that synthesized speech adapts dynamically rather than being constrained by fixed, pre-trained emotion embeddings.

3.2 Comparison with Contrastive Learning-Based Emotion Alignment

Many multimodal emotional TTS models incorporate contrastive learning to pre-align emotion embeddings from text, vision, and audio into a shared representation space. This approach ensures consistency across modalities but relies on fixed emotion representations at inference time.

Our approach introduces an alternative strategy by integrating real-time emotion modulation, where emotion embeddings are adaptively adjusted at inference time based on synthesis needs. This enables:

- **Fine-Grained Emotion Interpolation** Blending between emotions smoothly rather than switching between predefined categories.

- **Intensity Scaling** Adjusting the strength of an emotion dynamically (e.g., slightly sad vs. deeply sad).

- **Context-Aware Expressiveness** Modulating prosody in real-time based on speaker intent.

4 Datasets

Pretraining The TTS backbone is initially trained on neutral speech data to establish a robust foundation for speech generation. We use VCTK(Yamagishi et al., 2019) dataset as the pretraining data, which comprises of 44 hours of speech from 109 different speakers.

Finetuning The emotion module is introduced, and the entire system is finetuned using datasets enriched with emotional speech samples to ensure consistent and expressive output. Specifically, we leverage two datasets for finetuning: ESD (Zhou and et al, 2022b), which covers 5 emotion classes (neutral, angry, happy, sad, and surprised), and MEAD-TTS (Guan et al., 2024) dataset, which encompasses 8 emotion classes (neutral, angry, contempt, disgusted, fear, happy, sad and surprised).

4.1 Generation of Emotional Text Prompt

Given a dataset annotated with emotional tags, we generate multiple emotionally nuanced variations of each sentence. This ensures that the text better reflects the speaker's emotional state, enhancing the system's ability to produce more accurate prosodic features during synthesis. The input data includes columns indicating the speaker's sex (male, female) and the associated emotion from the 8 classes. Using this information, we generate five emotionally enriched variations of the sentence as text prompts using GPT-4.0 for each input sentences based on (Primus et al., 2023).

Synonym Mapping for Diversity We employ a synonym mapping technique to introduce lexical variety in the emotion-contained sentences. Each emotion is associated with a set of semantically similar words, such as "angry", "furious", "irate", etc. This lexical variety helps diversify the emotional cues present in the input text, thereby improving the robustness of the generated speech.

4.2 Comparison of Emotion-Labeled Dataset Sizes

To analyze the effectiveness of IMPACT-TTS in low-resource scenarios, we compare the size of labeled emotional speech datasets used in our model with those of MM-TTS, InstructTTS, and PromptTTS. The dataset size significantly impacts the ability of a model to generalize emotional expressiveness effectively.

As shown in Table 4, IMPACT-TTS effectively reduces reliance on large labeled emotional speech datasets by leveraging pretrained multimodal models, whereas other models require substantially larger explicitly labeled datasets.

5 Experimental Results

We evaluate the generated speech quality and similarity by objective metrics and subjective evaluations. **Short-Time Objective Intelligibility (STOI)** is an objective measure of how understandable and clear speech is. The STOI score ranges from 0 to 1, where 1 indicates perfectly intelligible speech. **Emotion Cosine Similarity (ECS)** measures the similarity between the synthesized and target emotional embeddings, with higher ECS values indicating better alignment. **Mel Cepstral Distortion (MCD)** measures the spectral distance between the ground truth and synthesized Mel-spectrum features. As for subjective evaluations, we conduct 5-scale Mean Opinion Score (MOS) using MOSNet (Chen-Chou et al., 2019) for 5 times. We generate 50 speech samples for each model and select 50 samples from MEAD-TTS and LibriTTS testing sets for intra-domain and out-of-domain evaluation respectively. Table 3 presents the experimental results of text prompt-based speech synthesis, encompassing MOS, audio quality and classification accuracy for emotion and gender. For text prompt based TTS, we conduct experiments on the following systems : 1) GT: This is the ground-truth recording; 2) GT (Mel + Voc): This is the speech synthesized using pretrained HiFi-GAN vocoder for GT Mel-spectrogram; 3) MM-TTS (Guan et al., 2024): This is a model for multi-modal prompt based TTS, which prompt encoder based on CLIP; 4) IMPACT-TTS: This is the proposed model for multi-modal prompt based expressive TTS. In the context of prompts, IMPACT-TTS model surpasses the baseline model in terms of audio naturalness and classification performance on the MEAD-TTS dataset as well as out-of-domain datasets. These

Model	Emotion-Labeled Speech Data (Hours)
IMPACT-TTS (Ours)	65
MM-TTS (Li et al., 2024)	100
InstructTTS (Yang and et al, 2024)	300
PromptTTS (Guo and et al, 2023)	152

Table 4: Comparison of emotion-labeled dataset sizes across emotional TTS models. The dataset sizes (in hours) are estimated. MM-TTS also utilizes 15,433 emotion-labeled images from the RAF-DB dataset.

results highlight the effectiveness of the method in accurately capturing and extracting emotion attributes from various text prompts.

6 Limitation

One limitation of the current model is the lack of vision datasets for facial expressions, leading to confusion with prompts like "sorrowful eyes," where happy emotions may be generated due to overlapping features like watery eyes. In the future, this limitation could be addressed by incorporating more diverse and abundant datasets and employing fine-grained image detection techniques to distinguish specific facial components. Another challenge lies in the large size of the representation model, which requires substantial server capacity for effective processing. Future work will focus on exploring lightweight alternatives, such as compressing the emotion encoder through pruning and quantization or using parameter-efficient fine-tuning methods.

7 Conclusion and Future Work

In this paper, we introduced IMPACT-TTS, a multimodal prompt-based TTS system that enhances emotional expressiveness while overcoming the limitations of small labeled datasets. By integrating large-scale pretrained models, generating text prompts with diverse synonyms and employing spherical interpolation for emotion modulation, IMPACT-TTS offers fine-grained control over emotional expression with minimal reliance on explicit emotion labels.

Future work will focus on improving efficiency by developing lightweight versions of the multimodal components. Additionally, we aim to extend the framework to support a broader range of emotional variations, including nuanced styles and speaker-adaptive expressions. We will also explore integrating ethical safeguards, such as audio watermarking for authenticity verification and usage monitoring, to prevent misuse. Incorporating broader ethical considerations into the design and deployment process will strengthen the trustworthiness and social responsibility of emotional TTS systems like IMPACT-TTS.

References

Edresson Casanova and et al. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. *International Conference on Machine Learning. PMLR*.

Lo Chen-Chou, Fu Szu-Wei, and Huang Wen-Chin. 2019. Mosnet: Deep learning-based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*.

Deok-Hyeon Cho and et al. 2024. Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. *INTERSPEECH 2024*.

Wenhao Guan, Yishuang Li, Tao Li, and Hukai Huang. 2024. Mm-tts: Multi-modal prompt based style transfer for expressive text-to-speech synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38.

Ashishkumar Gudmalwar and et al. 2024. Vecl-tts: Voice identity and emotional style controllable cross-lingual text-to-speech. *INTERSPEECH 2024*.

Zhifang Guo and et al. 2023. Prompttts: Controllable text-to-speech with text descriptions. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*.

Hee Soo Heo, Bong-Jin Lee, and Jaesung Huh. 2020. Clova baseline system for the voxceleb speaker recognition challenge 2020. *arXiv preprint arXiv:2009.14153*.

Chae-Bin Im and et al. 2022. Emoq-tts: Emotion intensity quantization for fine-grained controllable emotional text-to-speech. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*.

Xilin Jiang and et al. 2024. Speech slytherin: Examining the performance and efficiency of mamba for

speech separation, recognition, and synthesis. *arXiv preprint arXiv:2407.09732*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *NeurIPS 2020*.

Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung. 2023. Imaginary voice: Face-styled diffusion model for text-to-speech. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*.

Sang-Hoon Lee and et al. 2022. Hierspeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis. *Advances in Neural Information Processing Systems 35*, pages 16624–16636.

Sang-Hoon Lee and et al. 2023. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *arXiv preprint arXiv:2311.12454*.

Yeonghyeon Lee and et al. 2024. Voiceldm: Text-to-speech with environmental context. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*.

Yichong Leng and et al. 2023. Prompttts 2: Describing and generating voices with text prompt. *arXiv preprint arXiv:2309.02285*.

Xiang Li, Zhi-Qi Cheng, Jun-Yan He, and XiaoJiang Peng. 2024. Mm-tts: A unified framework for multimodal, prompt-induced emotional text-to-speech synthesis. *arXiv preprint arXiv:2404.18398*.

Guanghou Liu and et al. 2023. Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions. *arXiv preprint arXiv:2305.19522*.

Aaron van den Oord, Sander Dieleman, and Heiga Zen. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Paul Primus, Khaled Koutini, and Gerhard Widmer. 2023. Advancing natural-language based audio retrieval with passt and large audio-caption data sets. *arXiv preprint arXiv:2308.04258*.

Chen Sanyuan, Wang Chengwi, and Chen Zhengyang. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv preprint arXiv:2110.13900*.

Peng Wang and et al. 2023. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. The vctk corpus: The centre for speech technology research (cstr) voice cloning toolkit. https://datashare.ed.ac.uk/handle/10283/3443. Version 0.92.

Dongchao Yang and et al. 2024. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Kun Zhou and et al. 2022a. Emotion intensity and its control for emotional voice conversion. *IEEE Transactions on Affective Computing 14.1*, pages 31–48.

Kun Zhou and et al. 2022b. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18.

Kun Zhou, Berrak Sisman, and Haizhou Li. 2021. Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training. *INTERSPEECH 2021*.