



# EATS-Speech: Emotion-Adaptive Transformation and Priority Synthesis for Zero-Shot Text-to-Speech

Jingyuan Xing<sup>1</sup>, Zhipeng Li<sup>1</sup>, Shuaiqi Chen<sup>2</sup>, Xiaofen Xing<sup>1,\*</sup>, Xiangmin Xu<sup>1,3</sup>

<sup>1</sup>South China University of Technology, China

<sup>2</sup>Meituan, China

<sup>3</sup>Pazhou Lab, China

202320163324@mail.scut.edu.cn, eeleezp@mail.scut.edu.cn, chenshuaiqi03@meituan.com, xfxing@scut.edu.cn, xmxu@scut.edu.cn

## Abstract

Zero-shot text-to-speech (TTS) supports diverse speech synthesis without speaker-specific data but struggles to accurately transfer emotions from reference to target text. Traditional approaches treat emotion as part of a global style, leading to inconsistent emotional expressiveness. To address this, we propose EATS-Speech, an Emotion-Adaptive Transformation Synthesis framework. EATS-Speech employs Emotion Priority Synthesis through a parallel pipeline that decomposes speech into non-emotion style, emotion, and content. It prioritizes emotion generation to enhance expressiveness. Furthermore, it introduces Emotion-Adaptive Transformation Synthesis, where an LLM-based converter learns text-emotion mapping patterns from the reference speech and transfers them to the target text. Experiments on the LibriTTS dataset demonstrate the improvements in emotional expressiveness and accurate emotion adaptation. Speech demos are available at <https://test1341.github.io/demo/>.

**Index Terms:** Zero-shot TTS, Emotion-Adaptive Transformation Synthesis, Emotion Priority Synthesis, LLM-Based Conversion

## 1. Introduction

Zero-shot text-to-speech (TTS) is a powerful paradigm for synthesizing natural speech in diverse styles without requiring speaker-specific training data [1] [2] [3]. By conditioning on reference speech, zero-shot TTS models can capture and transfer various style attributes, such as prosody, pitch, and speaker identity, to generate speech for unseen speakers or styles [4] [5] [6]. Its capability has enabled multiple applications, from personalized voice assistants to expressive audiobook narration. However, it remains a significant challenge for the TTS model to accurately model and transfer emotion.

In traditional zero-shot TTS approaches, emotion is one of the style attributes embedded within a global representation of the reference speech [7] [8]. While the traditional approach effectively captures speaker identity and prosodic patterns, it introduces two key issues. Firstly, emotion, a perceptually salient factor for human listeners, is often overshadowed by other style attributes when modeled as part of a unified representation [9] [10]. It leads to insufficient emotional expressiveness in the synthesized speech. Secondly, emotion is closely tied to the semantic content of the text [11] [12]. Directly using the reference emotion without considering its alignment with the target text can result in emotionally inconsistent or incorrect outputs, as illustrated in Figure 1 (A).

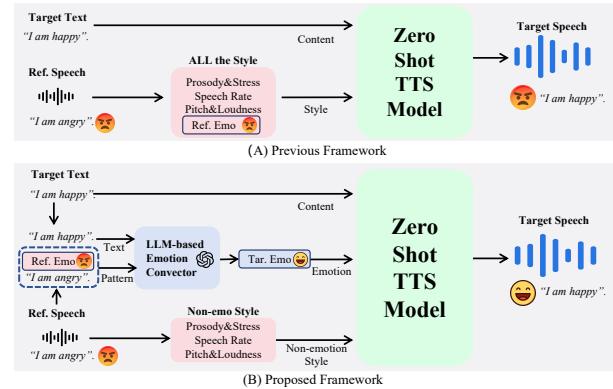


Figure 1: Comparison Between Previous and Proposed Zero-Shot TTS Frameworks.

To address the first issue, considering the significant role of emotion in human perception [13] [14], decoupling emotion from other style perception and prioritizing its synthesis is imperative [15] [16]. By isolating emotion-specific features from non-emotional style components, the synthesis process can concentrate on generating emotionally expressive speech before incorporating additional attributes. Regarding the second issue, a mechanism is required to dynamically adapt the reference emotion to the semantic expression of the target text. By harnessing the semantic understanding capabilities of large language models (LLMs) [17] [18], one can construct a transformation model that converts the reference emotion into an emotion congruent with the target content.

We proposed EATS-Speech, Emotion-Adaptive Transformation, and Priority Synthesis for Zero-Shot Text-to-Speech. As shown in Figure 1 (B), firstly, it decomposes the reference speech into emotion representations and non-emotion style attributes. These emotion representations are then transformed to match the target text's emotion using an LLM-based emotion conversion model, which takes the reference text, emotion representation, and target text as input. Finally, the target emotion representation, non-emotion style features, and target text are combined to generate the target speech that preserves the reference style while aligning emotional expressions with the content. The contributions are as follows:

- **Emotion Priority Synthesis:** EATS-Speech proposes a novel zero-shot speech synthesis framework based on emotion disentanglement. It decouples speech into three parallel streams: non-emotion style, emotion, and content, enabling parallel synthesis. The emotion branch employs fine-grained emotion representation at a granular level, prioritizing emo-

\*Corresponding author.

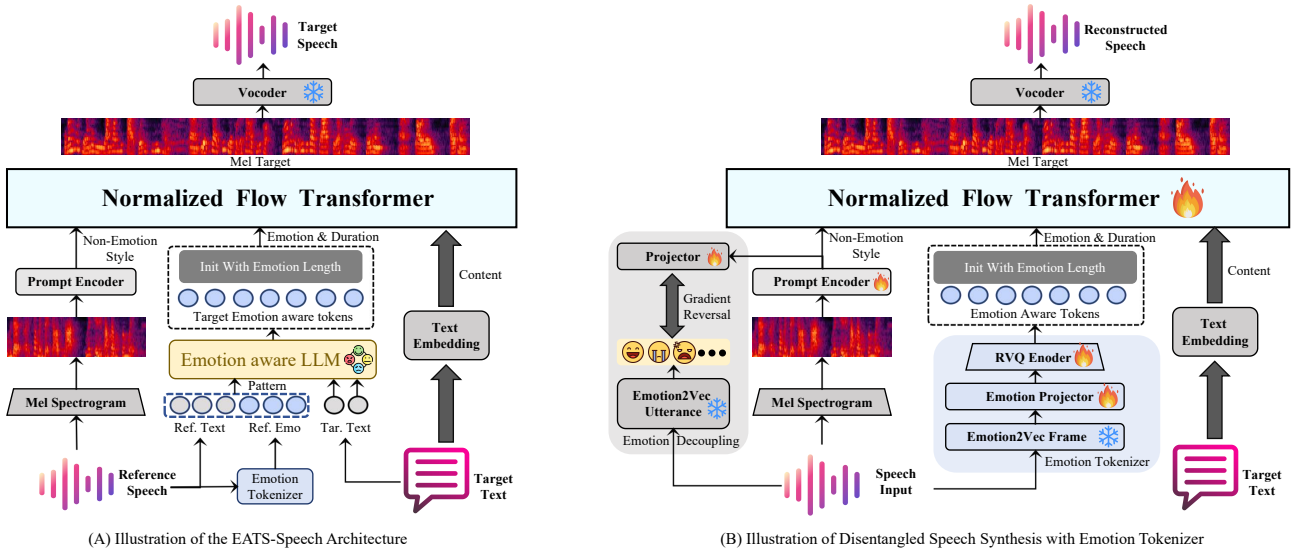


Figure 2: The overview of our proposed model. (A) Illustration of the EATS-Speech Architecture; (B) Illustration of Disentangled Speech Synthesis with Emotion Tokenizer.

tional processing to enhance expressiveness.

- **Emotion-Adaptive Transformation Synthesis:** EATS-Speech leverages the semantic understanding capabilities of LLMs by introducing an emotion-aware LLM converter into the emotion branch of the parallel synthesis framework. The converter learns emotion expression patterns from reference text-emotion pairs, ensuring that the synthesized speech’s emotions align with the target text rather than the reference speech, thereby improving emotional consistency.
- Experiments on the LibriTTS [19] datasets demonstrate that EATS-Speech enhances emotional expressiveness in synthesized speech, effectively adapting emotion to the target content.

## 2. Method

EATS-Speech, as illustrated in Figure 2(A), is built upon a disentangled speech synthesis framework with an Emotion Tokenizer, which decouples speech into three parallel streams: non-emotion style, emotion, and content. The non-emotion style represents speaker-specific characteristics unrelated to emotion, such as speaking habits and timbre. The emotion branch captures fine-grained, duration-aware emotional attributes at the frame level, considering the intrinsic relationship between speaking speed and emotional expression. The content branch is derived directly from the target text, ensuring accurate linguistic content generation. To address the challenge of guaranteeing emotional consistency between synthesized speech and its corresponding text, an Emotion-Aware LLM is integrated into the emotion branch. The emotion-aware LLM takes the emotion tokens extracted from the reference speech, the content of the reference text, and the target text as input. It first learns emotion-expression patterns from the reference emotion-text pairs and subsequently transfers these patterns to the target text, enabling emotion-adaptive synthesis.

The framework is built upon two essential components:

- **Disentangled Speech Synthesis with Emotion Tokenizer,** which prioritizes emotional attributes in the synthesis process, improving emotional expressiveness.

- **Emotion-Aware LLM for Emotion Transformation,** which transfers emotion-expression habits from the reference to the target text, ensuring emotional consistency.

The detailed methodologies for these components are described as follows.

### 2.1. Disentangled Speech Synthesis with Emotion Tokenizer

The Emotion Tokenizer extracts emotion tokens from speech. It comprises three components: a pre-trained frame-level emotion2vec module, an emotion projector, and a residual vector quantization (RVQ) encoder.

The training framework of the Emotion Tokenizer, shown in Figure 2(B), is based on a normalized flow transformer-based speech reconstruction model [1]. To achieve emotion-prioritized speech synthesis, we transform the synthesis process into parallel streams. The speech comprises three components: non-emotion style, emotion, and content. Since speech duration is closely linked to emotional expression, the duration is allocated to the emotion branch. The Emotion Tokenizer is a component of the emotion branch.

In the emotion branch, the input speech is first processed through the emotion2vec [20] module, which extracts frame-level emotional features at 50 Hz with a dimensionality of 768. These features are then projected to a lower dimensionality of 512 via the emotion projector, maintaining the same 50 Hz frequency. Subsequently, the features are passed through a 3-bit RVQ encoder. A codebook of 1024 entries represents each RVQ encoding.

For the non-emotion style branch, the input speech is initially converted into the mel spectrogram. It is followed by a prompt encoder to extract the fixed 512-dimensional style condition features. The prompt encoder’s architecture mirrors the condition encoder in Tortoise-TTS [21]. A gradient reversal mechanism is applied based on utterance-level emotional classification, which is performed using emotion2vec to ensure that this branch remains emotion-independent. Finally, the style condition features are projected onto the corresponding emotion

categories (8 classes) via the projector, ensuring the separation of emotion.

In the content branch, the input text is first passed through a text embedding layer. This embedding is used as a content condition and input into the normalized flow transformer. The Multi-reference Timbre Encoder (MRTE) [22] is then employed to align the text length with the length of the emotion token.

The loss function for training the Emotion Tokenizer denoted as  $\mathcal{L}_{ET}$ , is defined as:

$$\mathcal{L}_{ET} = \mathcal{L}_{RVQ} + \mathcal{L}_{GR} + \mathcal{L}_{NF} + \mathcal{L}_{Vocoder} \quad (1)$$

Where:

- $\mathcal{L}_{RVQ}$  is the RVQ reconstruction loss. It is computed as the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{RVQ} = \text{MSE}(\hat{e}, e) \quad (2)$$

Here,  $\hat{e}$  represents the decoded emotion tokens, and  $e$  represents the original emotion features.

- $\mathcal{L}_{GR}$  is the gradient reversal loss [23]. It is computed as the cross-entropy loss between the projected non-emotion style features and the emotion2vec utterance classification, with a gradient reversal applied:

$$\mathcal{L}_{GR} = -\log P(\hat{y}|f_{\text{style}}) \quad (3)$$

where  $\hat{y}$  is the predicted emotion classification, and  $f_{\text{style}}$  represents the non-emotion style features.

- $\mathcal{L}_{NF}$  is the reconstruction loss of the mel spectrogram. It is computed as the MSE loss between the mel spectrogram predicted by the Normalized Flow Transformer and the actual mel spectrogram corresponding to the emotion tokens:

$$\mathcal{L}_{NF} = \text{MSE}(\hat{m}, m) \quad (4)$$

Here,  $\hat{m}$  represents the predicted mel spectrogram, and  $m$  represents the ground truth mel spectrogram.

- $\mathcal{L}_{Vocoder}$  is the vocoder loss. It is computed by randomly cropping a segment from the predicted speech and comparing it to the corresponding segment from the ground truth speech using MSE loss:

$$\mathcal{L}_{Vocoder} = \text{MSE}(\hat{s}_{\text{rand}}, s_{\text{rand}}) \quad (5)$$

Here,  $\hat{s}_{\text{rand}}$  represents the randomly cropped predicted speech segment, and  $s_{\text{rand}}$  is the corresponding ground truth segment.

## 2.2. Emotion-Aware LLM for Emotion Transformation

The structure of the Emotion-Aware LLM is shown in Figure 3. The input to the model consists of the emotion tokens extracted from the reference speech, the reference text corresponding to the reference speech, and the target text to be synthesized.

This module is built on a retrained GPT-2 [17] decoder-only Transformer. We first extract the reference emotion tokens from the speech using the pre-trained Emotion Tokenizer. The reference emotion tokens, the reference text, and the target text are projected into an embedding space with the exact dimensions. All streams of the speech sequence are processed separately and then combined. The text and emotion sequences are augmented with different learnable positional embeddings. This embedding sequence is then processed through a stack of causal Transformer layers.

It employs autoregressive prediction, generating each token sequentially based on previous tokens. The key difference

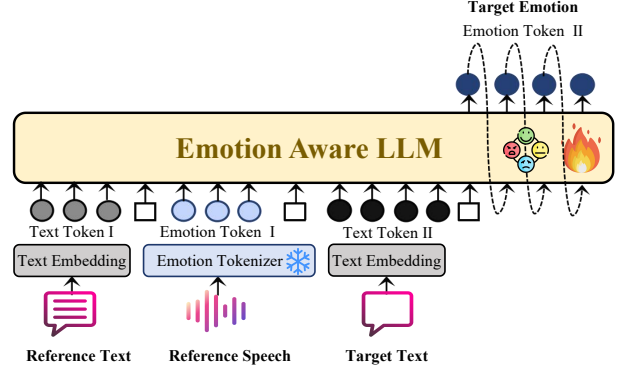


Figure 3: The overview of the Emotion-Aware LLM.

lies in using emotion tokens, which emphasize the emotional features of speech. Thus, this large model is designed to be emotion-aware, as it integrates both semantic and emotional cues in its generation process.

The training of this module consists of two steps. The first step focuses on training the model to map from text to emotion tokens. In this step, the model learns to predict the next emotion token given a sequence of text and its corresponding emotion tokens. The loss can be formalized as:

$$\mathcal{L}_{\text{step1}} = -\sum_{t=1}^T \log P(\hat{e}_t | \mathbf{e}_{<t}, \mathbf{t}) \quad (6)$$

where  $\hat{e}_t$  represents the predicted emotion token at time step  $t$ ,  $\mathbf{e}_{<t}$  denotes the previous emotion tokens, and  $\mathbf{t}$  represents the text sequence. The model learns to predict the next emotion token  $\hat{e}_t$  based on the previous tokens and the text.

The second step involves training the model for emotion transformation, where the reference emotion tokens  $\mathbf{e}_{\text{ref}}$  and reference text  $\mathbf{t}_{\text{ref}}$  are given, and the target emotion tokens  $\mathbf{e}_{\text{target}}$  are generated based on the target text  $\mathbf{t}_{\text{target}}$ . It is done autoregressively, where each target emotion token is generated sequentially based on the previous tokens, the reference emotion tokens, and the reference text. The loss for this step can be formulated as follows:

$$\mathcal{L}_{\text{step2}} = -\sum_{t=1}^T \log P(\hat{e}_t | \mathbf{e}_{<t}, \mathbf{e}_{\text{ref}}, \mathbf{t}_{\text{ref}}, \mathbf{t}_{\text{target}}) \quad (7)$$

This formulation allows the model to learn the mapping from the reference speech's emotional features to the target speech's emotional features, conditioned on the target text. The Emotion-Aware LLM effectively learns to transform emotional features through these two stages.

## 3. Experiments

### 3.1. Dataset

The LibriTTS [19] dataset is used. All comparison and ablation experiments are conducted at a sampling rate of 16 kHz.

LibriTTS is a multi-speaker English corpus. It amounts to 585 hours and over 2300 speakers. Train-clean-100, train-clean-360, and train-other-500 are merged as the training set. Dev-clean and dev-other are merged as a development set. Test-clean and test-other are merged as the test set.

Table 1: Comparison Experiments on LibriTTS Datasets. SBS refers to SpeechBERTScore, and ED refers to Emotion Discrepancy. EMOS and SMOS scores are reported with 95% confidence intervals. The best results are **bold**.

Model	EMOS ( $\uparrow$ )	SMOS ( $\uparrow$ )	SBS ( $\uparrow$ )	UTMOS ( $\uparrow$ )	WER ( $\downarrow$ )	ED ( $\downarrow$ )
<b>LibriTTS Development Set</b>						
GT	4.15 $\pm$ 0.10	4.28 $\pm$ 0.14	1.000	4.087	0.163	0.000
YourTTS	3.34 $\pm$ 0.14	3.38 $\pm$ 0.16	0.788	3.119	0.192	1.289
TransferTTS	3.35 $\pm$ 0.12	3.40 $\pm$ 0.12	0.830	3.304	0.216	1.243
VALL-E	3.28 $\pm$ 0.17	3.61 $\pm$ 0.12	0.798	3.681	0.264	1.215
E2-TTS	3.71 $\pm$ 0.15	3.67 $\pm$ 0.13	0.829	3.610	0.181	1.221
CosyVoice	3.57 $\pm$ 0.09	3.99 $\pm$ 0.10	0.823	3.894	0.220	1.014
EATS-Speech	<b>4.05 <math>\pm</math> 0.12</b>	<b>4.01 <math>\pm</math> 0.14</b>	<b>0.831</b>	<b>3.902</b>	<b>0.170</b>	<b>0.594</b>
<b>LibriTTS Test Set</b>						
GT	4.31 $\pm$ 0.15	4.35 $\pm$ 0.14	1.000	4.203	0.141	0.000
YourTTS	3.36 $\pm$ 0.16	3.36 $\pm$ 0.16	0.781	3.146	0.198	1.288
TransferTTS	3.32 $\pm$ 0.13	3.38 $\pm$ 0.15	0.821	3.065	0.264	1.346
VALL-E	3.38 $\pm$ 0.13	3.39 $\pm$ 0.15	0.695	3.220	0.276	1.399
E2-TTS	3.51 $\pm$ 0.15	3.52 $\pm$ 0.13	0.818	3.553	0.195	0.802
CosyVoice	3.78 $\pm$ 0.16	3.82 $\pm$ 0.16	0.819	3.734	0.238	1.023
EATS-Speech	<b>3.96 <math>\pm</math> 0.16</b>	<b>3.88 <math>\pm</math> 0.14</b>	<b>0.825</b>	<b>3.882</b>	<b>0.192</b>	<b>0.571</b>

### 3.2. Experiments Setup

For EATS-Speech training, we first download the pre-trained weights for emotion2vec base [20]<sup>1</sup>. Using the pre-trained weight, we train the Emotion Tokenizer on the LibriTTS training sets, each for 300k steps on an NVIDIA A800 GPU with a batch size of 16. The AdamW optimizer is used with an initial learning rate of  $2.0 \times 10^{-4}$ , which decays by a factor of  $0.999^{1/8}$  every epoch. After training, we freeze the emotion tokenizer and train the emotion-aware LLM for 500k steps under the same conditions. Finally, we freeze the emotion-aware LLM and the Emotion Tokenizer and fine-tune the Normalized Flow Transformer module of the EATS-Speech by 20k steps.

### 3.3. Comparison Experiments

Since EATS-Speech is fundamentally a zero-shot TTS model with emotion-adaptive capabilities, its core architecture and operational paradigm remain consistent with conventional zero-shot TTS frameworks. To ensure a fair and rigorous evaluation, we restrict our comparisons to models that operate within the zero-shot TTS paradigm. Furthermore, given that EATS-Speech incorporates both AR and NAR components in its design, we include state-of-the-art models representing both AR and NAR approaches in our comparative analysis.

For this purpose, we compare EATS-Speech against five prominent zero-shot TTS models: YourTTS [1], TransferTTS [24], VALL-E [25], E2-TTS [26], and CosyVoice [27]. These models were carefully selected to represent both AR and NAR paradigms, ensuring a balanced comparison that reflects the architectural innovations of EATS-Speech. To maintain experimental fairness and eliminate potential biases, all comparative models are retrained on the same dataset.

### 3.4. Subjective Evaluations

In the subjective evaluation, 20 human evaluators, trained on unannotated data to understand emotion and speech quality concepts, rate randomly selected samples on a scale of 1 to 5 for naturalness and emotion alignment.

We use the Emotion Mean Opinion Score (EMOS) to assess how well the synthesized speech’s emotional expressiveness aligns with the text content and the Speaker Similarity Mean Opinion Score (SMOS) to assess how closely the synthesized speech resembles the target speaker. The results are

<sup>1</sup>[https://huggingface.co/emotion2vec/emotion2vec\\_plus\\_base](https://huggingface.co/emotion2vec/emotion2vec_plus_base)

Table 2: Ablation Study. SBS refers to SpeechBERTScore, and ED refers to Emotion Discrepancy.

Model	SMOS ( $\uparrow$ )	EMOS ( $\uparrow$ )	SBS ( $\uparrow$ )	UTMOS ( $\uparrow$ )	WER ( $\downarrow$ )	ED ( $\downarrow$ )
<b>LibriTTS Development Set</b>						
EATS-Speech	<b>4.05 <math>\pm</math> 0.12</b>	<b>4.01 <math>\pm</math> 0.14</b>	<b>0.831</b>	<b>3.902</b>	<b>0.170</b>	<b>0.594</b>
-w/o Decoupling	3.55 $\pm$ 0.14	3.83 $\pm$ 0.13	0.723	3.862	0.184	0.926
-w/o Conversion	3.62 $\pm$ 0.17	3.87 $\pm$ 0.15	0.756	3.772	0.181	0.920
<b>LibriTTS Test Set</b>						
EATS-Speech	<b>3.96 <math>\pm</math> 0.16</b>	<b>3.88 <math>\pm</math> 0.14</b>	<b>0.825</b>	<b>3.882</b>	<b>0.192</b>	<b>0.571</b>
-w/o Decoupling	3.45 $\pm$ 0.15	3.68 $\pm$ 0.14	0.706	3.704	0.204	0.939
-w/o Conversion	3.54 $\pm$ 0.14	3.73 $\pm$ 0.14	0.738	3.660	0.206	1.056

shown in Table 1, where **GT** represents the ground truth speech.

### 3.5. Objective Evaluations

We evaluate EATS-Speech using four key metrics: SpeechBERTScore (SBS) [28], UTMOS [29], and Word Error Rate (WER). SBS measures the semantic alignment between the synthesized speech and ground truth. UTMOS evaluates the overall naturalness and perceptual quality, while WER, based on a Wav2Vec 2.0-large ASR model [30], assesses word-level synthesis accuracy. The results are presented in Table 1.

To specifically validate the emotion-adaptive capabilities of EATS-Speech, we introduce Emotion Discrepancy (ED) as a dedicated metric. It is computed using the utterance-level embeddings from emotion2vec large [20] to measure the alignment of emotional expressiveness between the synthesized speech and the ground truth. ED is quantified using Euclidean distance, where a smaller value indicates higher emotional consistency between the synthesized and reference speech.

Considering both **subjective** and **objective** metrics, the results demonstrate EATS-Speech’s effectiveness in emotion alignment, with significant improvements in emotional expressiveness and consistency. Additionally, enhancements in some other metrics indicate that resolving emotional consistency improves the coordination of diverse speech attributes, further optimizing synthesis quality and speaker similarity.

### 3.6. Ablation Study

Ablation studies are performed to evaluate the effectiveness of emotion decoupling and emotion conversion in the EATS-Speech framework. The experimental setup follows the same procedure as in the comparison experiments.

**-w/o Decoupling** refers to training EATS-Speech without emotion decoupling, where the emotion tokenizer is not used for gradient reversal of the emotion during the style conditioning process. **-w/o Conversion** refers to training EATS-Speech without the emotion-aware LLM, where only the emotion2vec model is used to assist in the zero-shot TTS synthesis, omitting the emotion conversion step. The results are presented in Table 2. The results show that emotion decoupling and emotion conversion contribute significantly to the model’s performance, with their absence leading to reduced emotion alignment and expressiveness in the synthesized speech.

## 4. Conclusion

This paper proposes a novel framework, EATS-Speech, introducing Emotion-Adaptive Transformation and Emotion Priority Synthesis. With Disentangled Speech Synthesis and Emotion-Aware LLM, our approach ensures emotionally expressive and consistent speech synthesis. Experimental results on the LibriTTS dataset demonstrate the superiority of EATS-Speech in achieving emotional adaptation and expressiveness.

## 5. Acknowledgements

The work is supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2022B0101010003; in part by Guangdong Basic and Applied Basic Research Foundation (2025A1515011203); in part by Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004).

## 6. References

- [1] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, “Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.
- [2] Y. A. Li, C. Han, and N. Mesgarani, “Styletts: A style-based generative model for natural and diverse text-to-speech synthesis,” *arXiv preprint arXiv:2205.15439*, 2022.
- [3] —, “Styletts-vc: One-shot voice conversion by knowledge transfer from style-based tts models,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 920–927.
- [4] X. Zhu, Y. Lv, Y. Lei, T. Li, W. He, H. Zhou, H. Lu, and L. Xie, “Vec-tok speech: speech vectorization and tokenization for neural speech generation,” *arXiv preprint arXiv:2310.07246*, 2023.
- [5] M. Kim, M. Jeong, B. J. Choi, D. Lee, and N. S. Kim, “Transduce and speak: Neural transducer for text-to-speech with semantic token prediction,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–7.
- [6] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang *et al.*, “Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models,” *arXiv preprint arXiv:2403.03100*, 2024.
- [7] D. Yang, D. Wang, H. Guo, X. Chen, X. Wu, and H. Meng, “Simplestts: Towards simple and efficient text-to-speech with scalar latent transformer diffusion models,” *arXiv preprint arXiv:2406.02328*, 2024.
- [8] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, X. Wu *et al.*, “Uniaudio: An audio foundation model toward universal audio generation,” *arXiv preprint arXiv:2310.00704*, 2023.
- [9] M. Kang, S. Kim, and I. Kim, “Unitts: Residual learning of unified embedding space for speech style control,” *arXiv preprint arXiv:2106.11171*, 2021.
- [10] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, “The ambiguous world of emotion representation,” *arXiv preprint arXiv:1909.00360*, 2019.
- [11] R. A. Patamia, P. E. Santos, K. N. Acheampong, F. Ekong, K. Sarpompong, and S. Kun, “Multimodal speech emotion recognition using modality-specific self-supervised frameworks,” in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2023, pp. 4134–4141.
- [12] Z. Wu, Y. Lu, and X. Dai, “An empirical study and improvement for speech emotion recognition,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [13] W.-T. Lin, K.-W. Liang, and P.-C. Chang, “Auditory physiology based emotion recognition system,” in *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2020, pp. 156–157.
- [14] J. A. Prado, C. Simplicio, and J. Dias, “Robot emotional state through bayesian visuo-auditory perception,” in *Technological Innovation for Sustainability: Second IFIP WG 5.5/SOCOLNET Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2011, Costa de Caparica, Portugal, February 21-23, 2011. Proceedings 2*. Springer, 2011, pp. 165–172.
- [15] A. Wilf and E. M. Provost, “Towards noise robust speech emotion recognition using dynamic layer customization,” in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.
- [16] Y. Chen, L. Yang, Q. Chen, J.-H. Lai, and X. Xie, “Attention-based interactive disentangling network for instance-level emotional voice conversion,” *arXiv preprint arXiv:2312.17508*, 2023.
- [17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [19] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “Libritts: A corpus derived from librispeech for text-to-speech,” *arXiv preprint arXiv:1904.02882*, 2019.
- [20] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, “emotion2vec: Self-supervised pre-training for speech emotion representation,” *Proc. ACL 2024 Findings*, 2024.
- [21] J. Betker, “Better speech synthesis through scaling,” *arXiv preprint arXiv:2305.07243*, 2023.
- [22] Z. Jiang, J. Liu, Y. Ren, J. He, C. Zhang, Z. Ye, P. Wei, C. Wang, X. Yin, Z. Ma *et al.*, “Mega-tts 2: Zero-shot text-to-speech with arbitrary length speech prompts,” *arXiv preprint arXiv:2307.07218*, 2023.
- [23] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.
- [24] M. Kim, M. Jeong, B. J. Choi, S. Ahn, J. Y. Lee, and N. S. Kim, “Transfer Learning Framework for Low-Resource Text-to-Speech using a Large-Scale Unlabeled Speech Corpus,” in *Proc. Interspeech 2022*, 2022, pp. 788–792.
- [25] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, “Neural codec language models are zero-shot text to speech synthesizers,” *arXiv preprint arXiv:2301.02111*, 2023.
- [26] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan *et al.*, “E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts,” *arXiv preprint arXiv:2406.18009*, 2024.
- [27] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, “Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens,” *arXiv preprint arXiv:2407.05407*, 2024.
- [28] T. Saeki, S. Maiti, S. Takamichi, S. Watanabe, and H. Saruwatari, “Speechbertscore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics,” *arXiv preprint arXiv:2401.16812*, 2024.
- [29] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, “Utmos: Utokyo-sarulab system for voicemos challenge 2022,” *arXiv preprint arXiv:2204.02152*, 2022.
- [30] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.