Emotional Text-to-Speech via Style Decoder with Emotion Shared Styleformer Block and RoPE Prior Encoder

Wenhan Yao[1], Fen Xiao[1], Ye Xiao[1], Zexin Li[1], Xiarun Chen[2], and Weiping Wen[2(✉)]

[1] XiangTan University, Xiangtan, Hunan, China
wenhanyao@smail.xtu.edu.cn
[2] Peking University, Beijing, China
weipingwen@pku.edu.cn

Abstract. Emotional Text-to-Speech (E-TTS) aims to generate speech that not only sounds natural but also conveys rich emotional expressions. Unlike traditional TTS, E-TTS must capture complex elements such as pitch, prosody, rhythm, and timbre variations to accurately convey emotions. Recently, some classical deep learning-based methods, such as Tacotron2, Transformer-TTS, FastSpeech2, and VITS, have significantly improved speech synthesis quality. However, these models still face challenges like alignment instability, strict duration constraints, and difficulties in generalizing across emotions and styles. The VITS model, while capable of high-quality speech synthesis, struggles with integrating emotional information due to its complex architecture. To address this, we propose RoStyleVITS, an end-to-end emotional TTS model built on VITS. RoStyleVITS incorporates emotion-infused styleformer blocks and replaces the standard attention layer with a self-attention layer using Rotary Position Embedding (RoPE) to enhance text sequence modeling. Our method outperforms existing state-of-the-art emotional speech synthesis models in both subjective and objective evaluations, demonstrating improved emotional expression and synthesis quality.

Keywords: Emotion Speech · VITS · Style Transfer · RoPE

1 Introduction

Emotional Text-to-Speech (E-TTS) is a crucial branch of text-to-speech [1–6] that aims to generate speech that not only exhibits high naturalness but also conveys rich emotional expressions. Unlike traditional Text-to-Speech (TTS) systems, E-TTS needs to model more complex speech elements, including pitch, prosody, rhythm, and timbre variations [7,8], to accurately express target emotions. This technology has broad applications in intelligent voice assistants, virtual character dubbing, gaming, and education.

In recent years, research on emotional speech synthesis has mainly focused on two categories: parametric modeling methods and deep learning-based methods. Parametric modeling approaches, such as Hidden Markov Models (HMMs) [9]

W. Senn et al. (Eds.): ICANN 2025, LNCS 16070, pp. 180–191, 2026.
https://doi.org/10.1007/978-3-032-04549-2_15

and rule-based methods like PSOLA [10], are difficult to achieve high naturalness and expressive speech. In contrast, most deep learning methods are based on several classical speech synthesis architectures [11–19], such as Tacotron2 [1], Transformer-TTS [4], FastSpeech2 [3], and VITS [5]. E-TTS methods that are built on these models have significantly improved synthesis quality and efficiency. However, these methods still suffer from certain limitations. For example, Tacotron2 and Transformer-TTS rely on an attention mechanism for alignment, which can lead to instability in synthesized speech, such as frame skipping and misalignment. FastSpeech2, while addressing alignment issues, enforces strict duration constraints, making it difficult to control speech timing flexibly. Moreover, existing approaches often struggle with cross-emotion generalization and multi-style speech modeling. The VITS model is trained based on variational autoencoder (VAE) theory, achieving exceptionally high-quality synthesized speech and allowing for stochastic prediction of speech duration. However, due to its complex architecture, it is not easy to integrate emotional information.

We aspire to create high-quality synthesized speech, so this paper chooses VITS as the foundation for our proposed model. To address the challenge of integrating emotional styles, we propose RoStyleVITS, an end-to-end novel emotional text-to-speech model that combines the advantages of both autoregressive and non-autoregressive models. Previous studies have demonstrated the effectiveness of self-attention models in speech generation [20]. Accordingly, we proposed to utilize emotion-shared double styleformer blocks on both sides of the decoder's MRF layer for emotion infusing. Additionally, we replace the standard attention layer in the original text prior encoder with a self-attention layer based on Rotary Position Embedding (RoPE). This substitution improves the effectiveness of text sequence modeling. Experiments demonstrate that our proposed method outperforms existing state-of-the-art emotional speech synthesis methods in both subjective and objective metrics.

2 Background

2.1 Text-to-Speech

Modern deep learning-based speech synthesis models can be broadly categorized into autoregressive methods (which learn $P_\theta([y_{t+1}]|[y_t, w_1, w_2, ..., w_N])$ and non-autoregressive methods (which learn $P_\theta([y_1, y_2, ..., y_T]|[w_1, w_2, ..., w_N]))$ based on their implementation principles. The y_t denotes a speech signal point, and the w_n denotes a text unit. Tacotron2 [1], Large Language Model based TTS (LLM-TTS) [6], and Transformer-TTS [4] fall into the autoregressive category, while VITS [5] and FastSpeech2 [3] belong to the non-autoregressive category. Emotional speech synthesis needs to generate speech with highly variable emotional prosody, so models that allow flexible control over speech duration are better suited for this task.

2.2 VITS

VITS (Variational Inference Text-to-Speech) is a deep learning-based model for text-to-speech (TTS) synthesis. It combines the strengths of variational autoencoders (VAEs), normalizing flows, and recurrent neural networks to generate high-quality, natural-sounding speech. VITS is designed to handle both prosody and content simultaneously, making it more efficient and capable of generating expressive and diverse speech. By using a probabilistic framework, VITS can model the uncertainty in speech synthesis, allowing for more flexible and robust TTS systems. Its end-to-end architecture eliminates the need for complex, hand-engineered features, enabling seamless and fast training. Since the duration of the generated speech can vary randomly (depending on the speaker or emotional control conditions), the VITS model is well-suited as an emotional TTS architecture.

3 Method

3.1 Framework Based on VITS

Our proposed framework is built on VITS, as shown in 1, We added a style encoder (G_{senc}) and the well-designed decoder in VITS as the style decoder (G_{sdec}). The other modules include the Normalizing Flow ($G_{f\theta}$) [21], Monotonic Alignment Search (MAS) [5], the text encoder (G_{text}, including a projection layer), the Stochastic Duration Preditor (SDP), and the posterior encoder(G_q). Then, we introduce the framework's overview and individual modules. The whole model G_θ can accept an input text x_{text} and the emotional reference utterance y_{ref}, the model synthesizes speech y_{syn} with emotional style taken from the reference speech. The text's true waveform is y.

Overview. Our proposed framework can be formulated as a conditional VAE, aiming to maximize the variational lower bound (also known as the evidence lower bound, or ELBO) of the intractable marginal log-likelihood of the $log[G_\theta(y_{syn}|x_{text}, y_{syn})]$:

$$log[G_\theta(y|x_{text}, y_{ref})] \geq E_{G_q}\left[log[G_{sdec}(y|zq, y_{syn})] - log\frac{G_q(zq|y)}{G_{text}(zp|x_{text})}\right] \quad (1)$$

where $G_{text}(zp|x_{text})$ denotes a prior distribution of the text's latent variables zp. The $G_{sdec}(y|zq, y_{syn})$ is the likelihood function of a true waveform with the emotion condition y_{syn}, making it generate target emotion speech. In the VAE, first, we incorporated a posterior encoder $G_q(zq|y)$, which estimates the posterior distribution zq of real speech, helping it achieve point-to-point probability distribution space alignment between true speech and the text.

Posterior Encoder. The posterior encoder G_q consists of multiple residual blocks used in Glow-TTS [22]. A Glow-TTS residual block consists of layers of dilated convolutions with a gated activation unit and skip connection. We add speaker embedding into the residual blocks as global conditioning van2016wavenet. In the output step, the linear projection layer above the blocks produces the mean and variance of the normal posterior distribution μ_q, σ_q, which means $zq = \mu_q + \epsilon \cdot \sigma_q$.

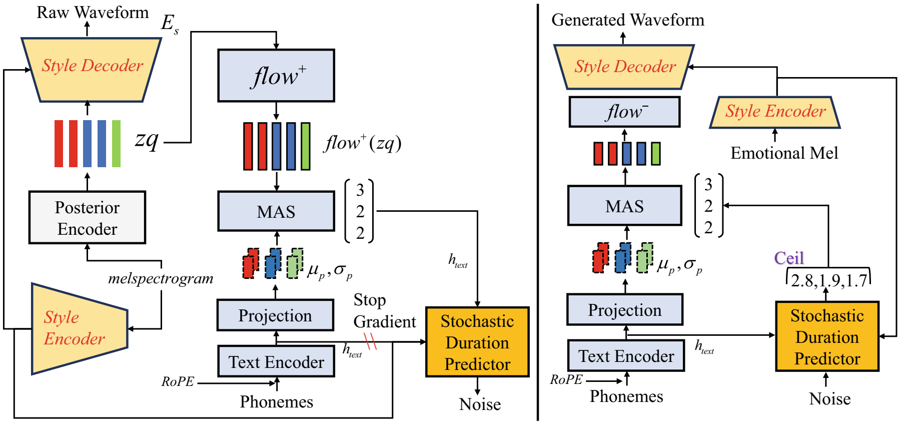

Fig. 1. Proposed framework.

Prior Modules. The prior encoder consists of a text encoder G_{text} and a normalizing flow $G_{f\theta}$. The text encoder accepts the phoneme index sequences and outputs the linguistic hidden representation's prior distribution μ_p, σ_p with a linear projection. We constructed the G_{text} and $G_{f\theta}$ by Rofomer's encoder [23] that utilizes Rotary Position Embedding (RoPE) to enhance the linguistic relationships. Under the guidance of RoPE, the Transformer-based speech emotion synthesis model achieves lower sequence loss, meaning that the reconstruction quality of the speech spectrogram is improved.

Other Modules. Monotonic Alignment Search (MAS) to align input phonemes with the target waveform by maximizing the ELBO rather than the exact log-likelihood. This redefined MAS ensures the alignment remains monotonic and non-skipping, reflecting natural speech patterns. Additionally, we introduce a Speaker and Emotion-Dependent Projection (SDP) module, where a linear layer integrates speaker and emotion embeddings—extracted from the style decoder—into the text representation, enabling the model to generate expressive speech conditioned on emotional prompts.

3.2 Style Encoder and Decoder

Our proposed core modules are the style encoder that encodes the input utterance's emotion information into the embedding and the style decoder that expands the lengths of zq and squeezes the dimension to a single-channel waveform.

Style Encoder. We apply the CAM++ model [24] as the emotion encoder. The style encoder accepts a speech spectrogram and outputs an emotion embedding. The entire model consists of two parts: a residual convolutional network as the

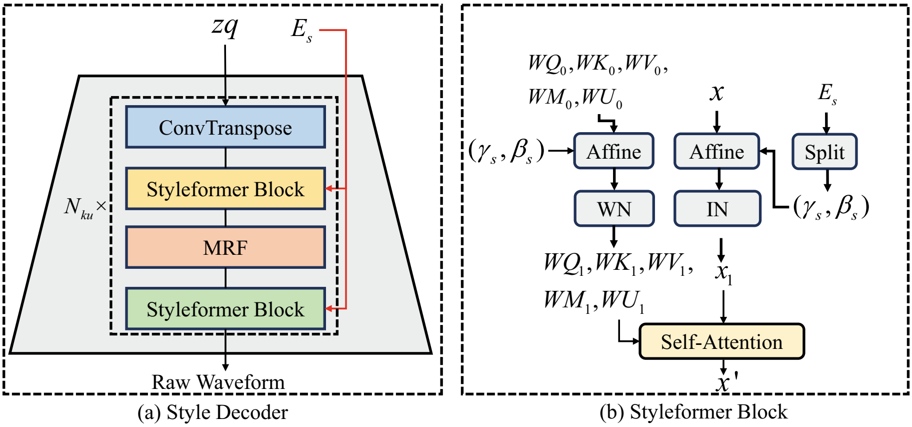

Fig. 2. The style decoder and the styleformer block. The decoder and the generator of HiFi-GAN have the same architecture. The styleformer block uses emotion embedding to compute an attention mechanism with emotional style.

front end and a time-delay neural network structure as the backbone. The front-end module is a 2D convolutional structure designed to extract more localized and detailed time-frequency features. The backbone module uses dense connections to reuse hierarchical features and improve computational efficiency. At the same time, each layer incorporates a lightweight context-aware mask module, which extracts contextual information at multiple scales through pooling operations. The mask helps remove irrelevant noise from the features while retaining key emotion information.

Style Decoder. As shown in Fig. 2, the style decoder consists of multiple expanding layers which include a transposed convolution, a front styleformer block, a multi-receptive field fusion (MRF) module, and a shared styleformer block. Based on the innovations in the StyleFormer [20], we share the emotional style embedding vector with the style transfer modules of the two twoStyle-Formerr blocks as shown in Fig. 2, indicated by the red arrow. Then, we will describe the styleformer block.

We assume the style block accepts input spectrograms $x \in R^{b,t,d}$ and emotion embedding $E_s \in R^{b,2d}$, the E_s is splitter split into two equal-length vectors, which serve as affine coefficients $(\gamma_s, \beta_s) \in R^{b,1,d}$. We create some trainable weights $\{WQ_0, WK_0, WV_0, WM_0, WU_0\} \in R^{b,d,d}$ during the model initialization phase. Then, these weight matrices are infused with the emotional style of the target speech through linear operations (Using WQ_0 matrix as an example) and used the weight normalization [25] (WN) for better training convergence:

$$WQ' = \gamma_s \cdot WQ_0 + \beta_s \tag{2}$$

$$WQ_1 = \frac{WQ'}{\sqrt{\sum_{k=0}^{d} WQ'^2_{i,j,k}}} \tag{3}$$

Then the input x is also infused with the emotional style:

$$x = \gamma_s \cdot \left(\frac{x - \mu_{b,t}}{\sigma^2_{b,t} + \epsilon} \right) + \beta_s \tag{4}$$

where $\mu_{b,t}$ denotes the mean of x along the d-dimension, $\sigma^2_{b,t}$ denotes the variance of x along the d-dimension, and ϵ represents a small constant (e.g., 1^{-5}) to prevent division by zero. Finally, we adopt the self-attention process with the style weights and style input sequence:

$$x' = \left[\frac{x \cdot (WQ_1) \cdot (x \cdot WK_1)^t}{\sqrt{d}} \cdot (x \cdot WV_1) \right] \cdot WM_1 + (x \cdot WU_1) \tag{5}$$

The style block outputs the stylized hidden spectrogram x'. We set two style-former blocks before and after the MRF in each layer for learning the acoustic emotion information from the emotional target speech, which shares the same emotion embedding. Considering the transposed convolutions and MRFs, we set the same weight parameters as the HiFi-GAN generator. Thus, the style decoder can generate the emotional raw waveform with the reference speech.

4 Experiments Setup and Metrics

4.1 Dataset

A successfully trained speech synthesis model generally requires a dataset with a total duration of more than 3 h, and each speaker should have at least 50 utterances. Based on this, we chose to use the **ESD** [8]. The ESD database consists of 350 parallel utterances spoken by 10 native English and 10 native Chinese speakers and covers 5 emotion categories (neutral, happy, angry, sad and surprise). Each audio sample has a sampling rate of 16 kHz and a 32-bit floating-point depth. Beside, **EmoV-DB** [26] is also considered. The database covers 5 emotion classes for four speakers, containing two males and two females.

4.2 Training Configuration

Since the input to the posterior encoder is the Short-Time Fourier Transform (STFT) spectrograms, the parameters for extracting the STFT spectrograms are as follows. We use a window length of 1024 (equal to FFT banks) and a hop length of 160. The absolute amplitude of each audio sample is constrained between 0 and 1. During training, all the learning rates and other parameters are kept consistent with the settings of VITS.

4.3 Baseline Models

We chose the following baselines. (1) TP-GST [27] + FS2 [3]. The model is based on FastSpeech2 and a global style-token emotion encoder. (2) EmoQ-TTS [12]. (3) MsEmoTTS [18]. (4) METTS [15]. Most of these models are based on the FastSpeech2 model and have advanced capabilities for stylized emotional speech synthesis.

4.4 Metrics

Subject Metrics. The subjective evaluation metrics mainly reflect human perception of the naturalness and quality of generated speech. We use the **Mean Opinion Score (MOS)** and the **A/B preference test** to assess the naturalness of generated emotional speech and human emotional preference for different models. The MOS is measured by some annotators to evaluate sound quality on a 5-point scale. The higher the speech quality, the higher the MOS score. In the test, the subjects are asked to choose which of the two speeches in the same sentence by different models is perceptually more expressive.

Objective Metrics. Objective metrics evaluate the quality and emotional category of speech through speech feature comparison and model-based automatic classification. We computed **Mel Cepstral Distortion (MCD)** [28], **Root Mean Squared Error of Log *f0* ($RMSE_{f0}$)** [29] metrics for objective evaluation. The *f0* denotes the pitch. The MCD measures the spectral difference between generated and real speech, while $RMSE_{f0}$ measures the fundamental frequency difference. The lower these two metrics are, the better the quality of the generated speech.

Additionally, we use a series of pre-trained models to evaluate the quality of generated utterances. We employ the NISQA model [30] to automatically predict the objective MOS score of utterances, referred to as **N-MOS**, which reflects the objective quality of the utterances. Furthermore, a speech emotion recognition model CAM++ [24], which is pre-trained on the ESD dataset, is used to predict the emotional category of generated utterances, and we compute the **Emotion Classification Macro F1 Score (Emo-F1)**, which reflects the accuracy of emotional expression in the generated utterances. Finally, we use the Whisper [31] model to calculate the **Word Error Rate (WER)** of the generated speech, which reflects the overall generation quality of the synthesis model. The higher the N-MOS and Emo-F1 metrics, the closer the speech's objective quality and emotional tendency are to real speech. Conversely, the lower the WER metric, the more successful the training of the speech synthesis model.

5 Ablation Study

As shown in the bottom two rows of the Tables 1 and 2, we constructed some model configurations for the proposed model: (1) **eVITS**. A VITS model that only additionally includes the speaker encoder layer. (2) **StyleVITS (w/o RoPE)**. The VITS with style modules but does not contain the RoPE text encoder. (3) **RoVITS (w/o style)**. The VITS with RoPE text encoder but does not contain the style modules. (4) **RoStyleVITS**. Proposed model, which contains style modules and RoPE text encoder.

Table 1. Results on ESD dataset

Model	MOS/N-MOS↑	MCD↓	$RMSE_{f0}$ ↓	Emo-F1↓ ↑	WER↓
Ground Truth	4.25/4.02	–	–	99.51%	2.10%
TP-GST + FS2 [3]	3.68/3.44	4.89	54.43	98.16%	3.64%
EmoQ-TTS [12]	3.72/3.53	4.81	53.15	99.39%	3.52%
MsEmoTTS [18]	4.02/3.83	4.76	53.04	99.51%	3.48%
METTS [15]	4.02/3.84	4.74	52.67	99.48%	3.41%
eVITS	3.52/3.48	4.96	54.56	97.89%	3.70%
StyleVITS (w/o RoPE)	3.79/3.67	4.87	49.56	98.54%	3.21%
RoVITS (w/o style)	3.84/3.72	4.67	54.37	97.51%	3.57%
RoStyleVITS	4.12/3.96	4.45	47.98	99.65%	2.78%

Table 2. Results on EmoV-DB dataset

Model	MOS/N-MOS↑	MCD↓	$RMSE_{f0}$ ↓	Emo-F1↓ ↑	WER↓
Ground Truth	4.27/4.01	–	–	99.47%	2.08%
TP-GST + FS2 [3]	3.65/3.42	4.82	53.98	98.12%	3.59%
EmoQ-TTS [12]	3.70/3.52	4.79	53.19	99.35%	3.48%
MsEmoTTS [18]	4.04/3.78	4.75	53.10	99.53%	3.38%
METTS [15]	4.07/3.90	4.70	52.43	99.59%	3.37%
eVITS	3.48/3.39	4.90	54.72	93.72%	3.82%
StyleVITS (w/o RoPE)	3.75/3.67	4.75	48.76	98.66%	3.16%
RoVITS (w/o style)	3.82/3.70	4.68	55.73	96.89%	3.37%
RoStyleVITS	4.10/4.02	4.39	47.56	99.42%	2.74%

6 Results

6.1 Objective Metrics Results

As shown in Tables 1 and 2, the experiments demonstrate that the MOS scores of our proposed method reached approximately 4.10, surpassing those of the more advanced baseline models. This indicates that the emotional speech generated by our proposed method is perceived as high-quality by human listeners. In the MOS evaluation, our analysis takes the ESD dataset as an example, as it contains a longer total duration of speech data, making it more representative for drawing accurate conclusions. The eVITS model lacks a proper emotional vector fusion module and suffers from convergence difficulties, resulting in lower synthesized speech quality. As a result, it only achieves a MOS score of 3.52. The StyleVITS (w/o RoPE) model does not include the RoPE module in its prior encoder, but it features a style transfer module that promotes model convergence and emotional fusion, achieving a better MOS score of 3.79 compared to eVITS. The RoVITS (w/o style) model lacks the emotional fusion module but includes the RoPE module, which enhances the effectiveness of text encoding, thereby reaching a MOS score of 3.84. This demonstrates that the RoPE module positively influences the quality of synthesized speech. The RoStyleVITS model incorporates both of our proposed key modules and thus achieves a higher MOS score of 4.12, outperforming the baseline models. Through these experiments, we demonstrate that both the RoPE and style blocks contribute positively to the quality of speech synthesis. A similar conclusion can be derived from the EmoV-DB dataset. Due to the shorter total duration of speech in EmoV-DB, the final experimental metrics are lower than those obtained on the ESD dataset.

In the Fig. 3, we present the A/B preference test between the baseline model and the proposed method. In the test, the subjects are asked to choose which of the two speeches for the same sentence, by different models, is perceptually more expressive. The experiment proves that RoStyleVITS is preferred over the baseline models.

6.2 Subjective Metrics Results

As shown in Tables 1 and 2, in the objective metric experiments, lower MCD, RMSE, and WER values indicate higher generated speech quality, while a higher Emo-F1 score demonstrates the model's ability to successfully generate utterances with reference emotion. We analyze the emotion-related metrics. The experimental results show that models lacking the effective emotional fusion blocks, such as eVITS and RoVITS (w/o style), perform poorly on $RMSE_{f0}$ and emo-F1 metrics. This indicates that the generated target emotional styles are either less distinguishable or exhibit noticeable deviation, highlighting the importance of our proposed style transfer block.

Next, we consider the semantics-related metrics. The results show that both the RoPE and the emotional style transfer block lead to improvements in these metrics. This indicates that they improve reconstruction quality in two ways: efficient textual sequence encoding and weight normalization.

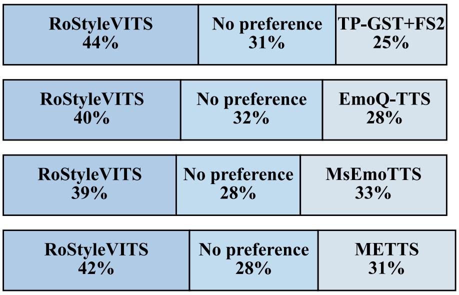

Fig. 3. A/B preference test

7 Conclusion

This paper proposes RoStyleVITS, a model built upon VITS that enhances text prior encoding using RoPE. Additionally, it employs a styleformer module with shared emotional embedding to construct an end-to-end multi-emotion and multi-speaker text-to-speech model. We propose replacing the standard attention layer with the RoPE attention layer to enhance the text encoder. Additionally, we design a dual-layer styleformer module with shared emotional embeddings in the decoder of the VITS model, which effectively injects the style of the reference speech during the decoding process. Experimental results show that our proposed method outperforms the baseline model in both subjective and objective metrics, indicating that our model can synthesize high-quality emotional speech. Furthermore, due to the use of the emotional encoder, it also has zero-shot generation capabilities.

References

1. Shen, J., et al.: Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4779–4783. IEEE (2018)
2. Ren, Y., et al.: Fastspeech: fast, robust and controllable text to speech. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
3. Ren, Y., et al.: Fastspeech 2: fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558 (2020)
4. Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M.: Neural speech synthesis with transformer network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 6706–6713 (2019)
5. Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International Conference on Machine Learning, pp. 5530–5540. PMLR (2021)
6. Wang, C., et al.: Neural codec language models are zero-shot text to speech synthesizers. arXiv preprint arXiv:2301.02111 (2023)

7. Chan, C.H., Qian, K., Zhang, Y., Hasegawa-Johnson, M.: Speechsplit2. 0: unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6332–6336. IEEE (2022)
8. Zhou, K., Sisman, B., Liu, R., Li, H.: Emotional voice conversion: theory, databases and ESD. Speech Commun. **137**, 1–18 (2022)
9. Nose, T., Yamagishi, J., Masuko, T., Kobayashi, T.: A style control technique for hmm-based expressive speech synthesis. IEICE Trans. Inf. Syst. **90**(9), 1406–1413 (2007)
10. Toma, Ş.-A., Târşa, G.-I., Oancea, E., Munteanu, D.-P., Totir, F., Anton, L.: A td-psola based method for speech synthesis and compression. In: 2010 8th International Conference on Communications, pp. 123–126. IEEE (2010)
11. Bott, T., Lux, F., Vu, N.T.: Controlling emotion in text-to-speech with natural language prompts. arXiv preprint arXiv:2406.06406 (2024)
12. Im, C.-B., Lee, S.-H., Kim, S.-B., Lee, S.-W.: Emoq-TTS: emotion intensity quantization for fine-grained controllable emotional text-to-speech. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6317–6321. IEEE, 2022
13. Diatlova, D., Shutov, V.: Emospeech: guiding fastspeech2 towards emotional text to speech. arXiv preprint arXiv:2307.00024 (2023)
14. Cui, C., et al.: Emovie: a mandarin emotion speech dataset with a simple emotional text-to-speech model. arXiv preprint arXiv:2106.09317 (2021)
15. Zhu, X., et al.: Metts: multilingual emotional text-to-speech by cross-speaker and cross-lingual emotion transfer. IEEE/ACM Trans. Audio Speech Lang. Process. **32**, 1506–1518 (2024)
16. Li, X., et al.: Umetts: a unified framework for emotional text-to-speech synthesis with multimodal prompts. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2025)
17. Liu, R., Sisman, B., Li, H.: Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability. arXiv preprint arXiv:2104.01408 (2021)
18. Lei, Y., Yang, S., Wang, X., Xie, L.: Msemotts: multi-scale emotion transfer, prediction, and control for emotional speech synthesis. IEEE/ACM Trans. Audio Speech Lang. Process. **30**, 853–864 (2022)
19. Alemayehu, Y., Yadav, R.K., Mohammed, A.R., Thapa, S., Chauhan, S.: Infusing emotion in text to speech model. In 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 712–716. IEEE (2024)
20. Park, J., Kim, Y.: Styleformer: transformer based generative adversarial networks with style vector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8983–8992 (2022)
21. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International Conference on Machine Learning, pp. 1530–1538. PMLR (2015)
22. Kim, J., Kim, S., Kong, J., Yoon, S.: Glow-TTS: a generative flow for text-to-speech via monotonic alignment search. In: Advances in Neural Information Processing Systems, vol. 33, pp. 8067–8077 (2020)
23. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
24. Wang, H., Zheng, S., Chen, Y., Cheng, L., Chen, Q.: Cam++: a fast and efficient network for speaker verification using context-aware masking. arXiv preprint arXiv:2303.00332 (2023)

25. Salimans, T., Kingma, D.P.: Weight normalization: a simple reparameterization to accelerate training of deep neural networks. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
26. Adigwe, A., Tits, N., Haddad, K.E., Ostadabbas, S., Dutoit, T.: The emotional voices database: Towards controlling the emotion dimension in voice generation systems. arXiv preprint arXiv:1806.09514 (2018)
27. Wang, Y., et al.: Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis. In: International Conference on Machine Learning, pp. 5180–5189. PMLR (2018)
28. Kubichek, R.: Mel-cepstral distance measure for objective speech quality assessment. In: Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, vol. 1, pp. 125–128. IEEE (1993)
29. Chai, T., Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)?-arguments against avoiding RMSE in the literature. Geoscientific Model Dev. **7**(3), 1247–1250 (2014)
30. Mittag, G., Naderi, B., Chehadi, A., Möller, S.: Nisqa: a deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets (2021)
31. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 492–518. PMLR (2023)