# EFTTS: Zero-Shot Emotional Speech Synthesis via Conditional Flow Matching and Self-Supervised Representations

Haoyu Wang, Jiale Chen, Jiaxun Li, Sizhe Shan and Yuehai Wang◇
Zhejiang University, China
E-mail: {22331169, 22431155, 22460307, 22360174, wyuehai}@zju.edu.cn

***Abstract*—Text-to-speech (TTS) technology has made significant advancements. However, generating high-quality, controllable emotional speech remains challenging due to the complex nature of emotions and speaker variability. In this paper, we propose EFTTS, a controllable, zero-shot emotional TTS model. EFTTS extracts emotional features using a self-supervised emotion representation model and utilizes conditional flow matching to model emotion features with text inputs. This approach allows for precise emotional control and decouples emotion from other paralinguistic features. Additionally, we introduce zero-shot emotion transfer to generate emotionally appropriate speech that is closely aligned with the input text and reference speech. Experimental results show that EFTTS outperforms existing methods in terms of emotional expressiveness, naturalness, and synthesis quality, offering a promising solution for high-quality, controllable emotional speech synthesis. Demo samples are available at https://bigbroes.github.io/EFTTS/.**

## I. Introduction

Text-to-speech (TTS) technology aims to synthesize high-quality speech. In recent years, TTS has made significant progress, with some end-to-end models achieving human-level performance[1]–[3]. With the increasing application of speech synthesis technology in fields such as voice assistants and text-to-speech reading, the focus of research has shifted towards generating more expressive and stylistically rich emotional speech.

Although significant progress has been made in emotional speech synthesis, achieving high-quality and controllable speech synthesis remains a challenge. This is primarily due to the varying ways in which emotions are expressed across different speakers and text transcriptions. Some earlier works in emotional speech synthesis have employed sequence-to-sequence models[4], where attention mechanisms allow the model to focus on emotionally salient parts of the speech. However, Seq2Seq models face typical challenges of auto-regressive models, such as long-term dependencies and repetition issues. With the development of end-to-end models, some have demonstrated their powerful capabilities in text-to-speech. For example, [5] uses categorical labels to select predefined emotional embeddings to control speech generation. [6] selects the center of emotion clusters based on labels and uses them as emotional embeddings to synthesize speech. However, such approaches fail to capture the rich emotional variations present in speech, resulting in synthesized speech that exhibits overly averaged emotional expressions. Reference-based methods have made significant advancements [7]. Global Style Tokens (GST)[8], enables speech synthesis without requiring explicit style labels. GST calculates the similarity between reference audio and style tokens using an attention module, which allows the model to generate style embeddings that guide the synthesis process. Building upon this, [9], [10] have extended this approach, transferring style control from random styles to emotional speech generation. However, this approach lacks the ability to directly control the emotion in advance. Furthermore, [11] employs multi-level, fine-grained style modeling to capture the style in the reference speech by combining utterance-level and frame-level emotional embeddings. Despite these advancements, traditional reference-based methods that rely on a simple reference encoder still struggle to separate emotion from other paralinguistic features, such as speaker identity and prosody. Additionally, since the content of the reference audio often differs from the speech that needs to be generated, these methods typically produce synthesized speech that is overly similar to the reference in both style and emotional expression, leaving little room for flexibility. This limits the model's ability to adapt to the emotional nuances of the input text, making the generated speech sound more like a replication of the reference, rather than a proper adjustment to the emotional context conveyed by the text.

With the rise of generative models like diffusion[12] and flow matching[13], [14] in TTS, emotional speech generation has also made substantial progress. [15] uses an emotional classifier to guide the diffusion process in generating emotional speech, while [16] leverages features from speech recognition models (SER) as conditions to mix different emotional types, enabling controllable speech synthesis in terms of emotional variety and intensity. However, these methods still face challenges in terms of synthesis speed and capturing emotion that is tightly linked to the input text.

In response to these challenges, we propose EFTTS, a controllable emotional speech synthesis and zero-shot emotion transfer model based on a self-supervised emotional representation model and conditional flow matching. The main contributions of our work are as follows:

◇Corresponding author

979-8-3315-7206-8/25/$31.00 

- We extract unique emotional features for each speech sample through a self-supervised speech emotion representation model (Emotion2Vec) rather than learning only the average representation of each emotion, which also decouples emotion from other paralinguistic information.
- We predict the emotional features corresponding to the input phonemes through conditional flow matching, modeling emotion features that are closely aligned with the text, rather than relying on a global average representation, thereby enabling high-quality and controllable emotional generation.
- Combining the reference encoder and emotional features from conditional flow matching in zero-shot emotion transfer. This approach enables emotion transfer that is more naturally aligned with the input text, rather than rigidly copying the emotion in reference speech.

To ensure author anonymity, the link to the audio demo page will be added after the review process.

## II. Proposed Method

In this paper, we introduce **EFTTS**, an end-to-end controllable zero-shot emotional text-to-speech model, which utilizes a self-supervised pre-trained model to extract emotional representations and employs conditional flow matching (CFM) [17] to model and sample from the emotional distribution. This enables the model to predict corresponding emotional features for speech from text input. **Fig. 1** and **Fig. 2** illustrates the training and inference frameworks of the model, respectively. During zero-shot generation, we combine the emotional features from the emotion encoder and CFM to obtain richer emotional characteristics. The specific details of the model will be introduced in the following sections.

### A. Overview of EFTTS

Our model is built upon the StyleTTS2[18] framework, which follows an encoder-decoder structure for the TTS backbone. An improved ISTFTNet[19] decoder accepts aligned text embeddings, emotional features, pitch ($F_0$), and energy to directly generate waveforms, as represented by $\text{Decoder}(e, F_0, N, h_{\text{align}})$, thus avoiding the two-stage loss typically associated with generating mel spectrograms.

Similar to StyleTTS2, we integrate a phoneme-level BERT [20] text encoder to enrich the prosodic information within the phoneme embeddings. These embeddings are then used to predict the duration, $F_0$, and energy for speech synthesis. The ground truth duration is obtained from a pre-trained ASR model, while the $F_0$ energy is derived from a pitch extractor.

For speech emotion representation, we employ the Emotion2Vec base model, a self-supervised model designed to capture speech emotion features. We use it to extract utterance-level emotional features for each speech sample. Conditional Flow Matching (CFM) is then applied to model the vector field, representing the transition from noise to the emotional feature distribution. The predicted emotional features are obtained by solving the ordinary differential equation (ODE) governing this flow. These features are subsequently integrated into the duration predictor, pitch predictor, and decoder through an adaptive normalization method (AdaIN)[21]. This integration ensures that the synthesized emotional speech is more closely aligned with the context and content of the original text, producing a more natural and contextually appropriate emotional expression.

We employ both a Multi-Period Discriminator (MPD) and a Multi-Resolution Discriminator (MRD)[22] to evaluate the quality of the synthesized speech across multiple temporal periods and frequency resolutions, ensuring that both fine-grained details and global structures are effectively captured. By optimizing the adversarial training losses based on these evaluations, we improve the overall quality of the synthesized speech, leading to more natural and realistic outputs.

All modules are jointly trained using the following loss function:

$$\mathcal{L}_{\text{jointly}} = \mathcal{L}_{\text{mel}} + \mathcal{L}_{\text{dur}} + \mathcal{L}_{F_0} + \mathcal{L}_N + \mathcal{L}_{\text{CFM}} + \mathcal{L}_{\text{GAN}} \quad (1)$$

This comprehensive training strategy ensures the generation of high-quality and emotionally expressive speech.

### B. Emotion Encoder

In our approach, we utilize Emotion2Vec[23], a state-of-the-art speech emotion representation model, as the emotion encoder. Emotion2Vec has demonstrated exceptional performance across a range of downstream tasks in speech emotion recognition, establishing it as a versatile and robust tool for extracting emotional features from speech data. The model is trained using a large, unlabeled collection of emotional speech datasets in an unsupervised manner[24], [25]. Through this process, it learns to capture highly generalized emotional features that are not only applicable to a specific emotion but can also generalize across different emotional expressions and contexts. Emotion2Vec employs a combination of online distillation and a masked prediction approach, which enhances its capacity to extract specialized emotional features from each individual speech sample. Online distillation allows for real-time adaptation and learning from new data, while the masked prediction strategy encourages the model to learn to predict missing emotional information from context, thereby improving its ability to model complex emotional expressions. This combination helps mitigate the common issue of overly averaged emotional representations, where traditional models tend to smooth out subtle emotional variations, making it difficult to capture the finer nuances of emotion in speech. Emotion2Vec, by contrast, preserves the richness and diversity of emotional expression, enabling more accurate and controllable emotional speech synthesis. It also effectively decouples emotion from other acoustic features, avoiding the issues of insufficient and non-generalizable emotion feature extraction that arise from training an emotion encoder from scratch.

By leveraging the power of Emotion2Vec, we obtain emotion features that are most aligned with the speech's semantic context, effectively avoiding the issue of overly averaged

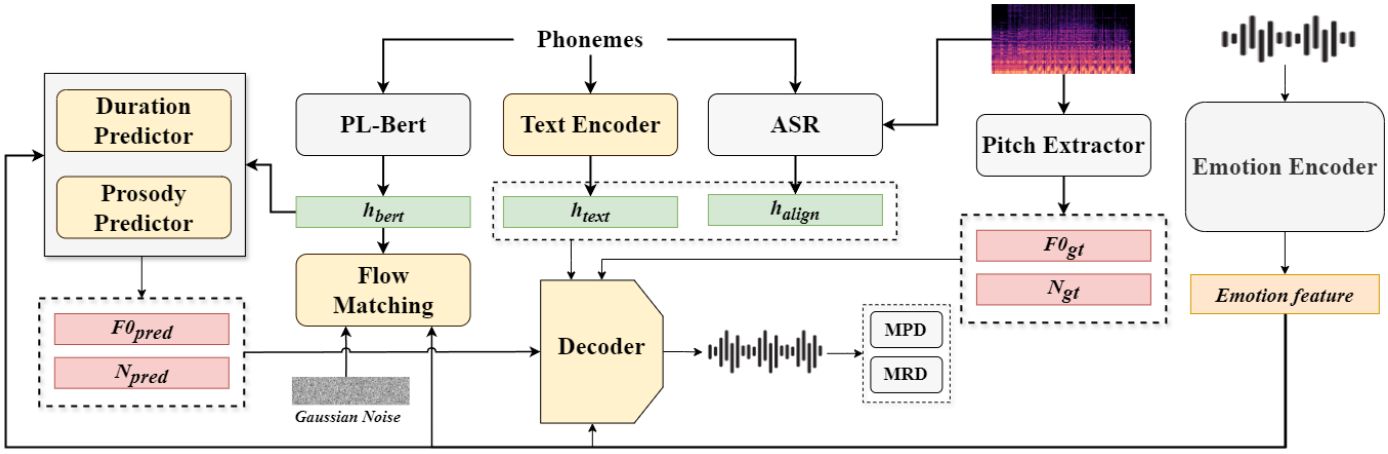


Fig. 1. Training diagram of our model. The pitch extractor, PL-bert, and ASR modules are frozen during training. The emotion encoder extracts the utterance-level features of speech.

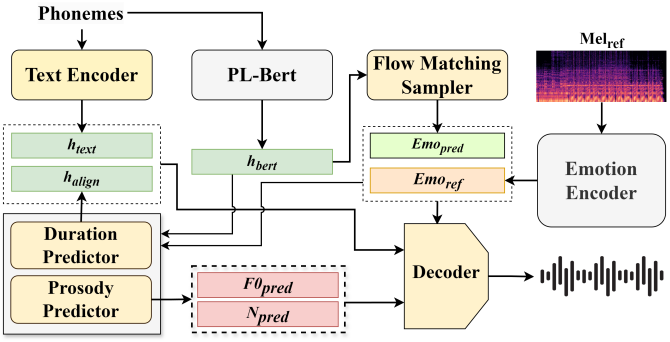


Fig. 2. Inference structure of the EFTTS.

emotions. At the same time, the self-supervised model's capabilities allow us to decouple emotion from other paralinguistic information, ensuring a more precise and contextually appropriate emotional representation.

### C. Conditional Flow Matching

Flow Matching (FM) is a continuous flow process that learns a time vector field $v_\theta(x,t)$ from an initial distribution to a target distribution, used to describe changes in probability flow. The data distribution is represented as $p_1(x)$, the prior distribution as $p_0(x)$, and the true vector field $u(x,t)$ represents the transformation from the prior distribution to the data distribution. The flow is defined by the ordinary differential equation (ODE):

$$\frac{dx(t)}{dt} = u(x,t) \tag{2}$$

If a learnable neural network can accurately fit the true vector field, it will be able to obtain the probability path $p_t$ for transforming the distribution. The loss function for Flow Matching is given by:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,p_t(x)}\|v_\theta(x,t) - u(x,t)\|^2 \tag{3}$$

where $\theta$ represents the learnable parameters, and $t \sim U[0,1]$. However, the Eq.(3) is often difficult to compute because $p_t(x)$ and $u(x,t)$ cannot be directly determined. To address this issue, Conditional Flow Matching (CFM) incorporates the data sample $x_1$ as a condition into the true vector field $u_t(x)$, transforming it into a conditional vector field $u_t(x)$. By learning this conditional vector field, it is ultimately possible to obtain the conditional probability path $p_t(x|x_1)$, with the loss function being

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,p_1(x_1),p_t(x|x_1)}||v_\theta(x,t) - u_t(x|x_1)||^2 \tag{4}$$

Voiceflow[13] demonstrated that the conditions in the Eq.(4) can be further generalized. Any condition $c$ (text or speech) has the same optimization objective for the conditional probability path as in the Eq.(4). We integrate the text and emotion labels into the model condition c, whose probability distribution is $p(c)$. The generation target is the speech feature vector $x_1$ that contains emotional information, while the noise $x_0$ is sampled from the standard normal distribution $\mathcal{N}(0,I)$. $x_1$ and $x_0$ are the boundaries of the conditional distribution $p_t(x|c)$. The conditional vector field $u(x,t|c)$ can be defined by the ordinary differential equation $dx = u_t(x|c)dt$, and our optimization objective can be expressed as:

$$\min_\theta \mathbb{E}_{t,p_t(\boldsymbol{x}_t|\boldsymbol{c}),p(\boldsymbol{c})}\|\boldsymbol{v}_\theta(\boldsymbol{x},t,\boldsymbol{c}) - u_t(x|c)\|^2 \tag{5}$$

## III. Experiments

### A. Experimental Setup

We utilized the English portion of the Emotional Speech Dataset (ESD) [4] for all our experiments. The dataset consists of recordings from 10 English speakers, each expressing five emotional states: neutral, happy, angry, sad, and surprised, resulting in 17500 samples in total. Each speaker provides 350 parallel utterances per emotional state, resulting in approximately 1.2 hours of speech data per speaker. To ensure consistency and reproducibility, we allocated 20 samples per

emotion for each speaker for validation and test set, which were exclusively used to evaluate the model's generalization and robustness on unseen data. This setup allowed us to rigorously assess the model's ability to learn and generalize across diverse emotional expressions.

To process the audio, we first extracted Mel-spectrograms by applying a short-time Fourier transform (STFT) with a hop size of 256, a window size of 1,024, and a fast Fourier transform (FFT) size of 1,024. The resulting spectrograms were then mapped to 80 Mel frequency bins to represent the energy distribution across the Mel scale. In addition, we utilized Phonemizer[26] to convert the text into phoneme sequences. This tool preserves punctuation and stress markers, allowing us to obtain phoneme representations that contain more rhythmic and prosodic information which are crucial for conveying emotion.

The parameters of Emotion2Vec were frozen throughout the training process. The features from Emotion2Vec are projected to a 128-dimensional space through a linear layer. The duration and prosody predictor is a sequence prediction model based on LSTM, which predicts the duration, $F0$, and energy of the speech. The model leverages the $h_{\text{bert}}$ from the PL-bert to inform the prediction of these prosodic features.

The model was jointly trained on 4 NVIDIA RTX 4090D GPUs for 100 epochs, using the AdamW optimizer [27] with $\beta_1 = 0$, $\beta_2 = 0.99$, weight decay $\lambda = 10^{-4}$, learning rate $\gamma = 5 \times 10^{-5}$ and a batch size of 16.

For the Conditional Flow Matching (CFM) module, the number of function evaluations (NFE) for the ODEs was randomly sampled from 1 to 4 during training for computational efficiency and fixed to 3 during inference to balance speed and quality.

### B. Evaluation

For subjective metrics, we employed mean opinion scores to evaluate the naturalness (nMOS) and emotional similarity (eMOS) of the synthesized speech. 20 participants who majored in English were recruited for the evaluation, and 20 samples were randomly selected from each emotional category in the test set. Participants were asked to rate the speech on a scale from 1 to 5 based on its naturalness and emotional similarity. The results are presented with a 95% confidence interval.

For objective metrics, we calculated the Mel Cepstral Distortion (MCD)[28], which reflects the spectral difference between the synthesized speech and the reference speech, serving as an indicator of the generated speech's acoustic quality and fidelity. Additionally, we used the Emotion2Vec+large model to compute the Emotion Recognition Accuracy (ERA).

### C. Comparison

We selected Emodiff[15] and EmosphereTTS[29] for comparison to evaluate the results of controllable emotional speech synthesis.

TABLE I
RESULTS OF CONTROLLABLE EMOTIONAL SPEECH SYNTHESIS WITH EMOTION LABELS.

| Emotion | nMOS(CI)↑ | eMOS(CI)↑ | MCD↓ | ERA↑ |
|---|---|---|---|---|
| Angry | 4.23(±0.07) | 4.24(±0.7) | 4.80 | 86.60% |
| Happy | 3.98(±0.06) | 4.03(±0.10) | 4.96 | 71.34% |
| Netural | 4.10(±0.08) | 4.18(±0.09) | 4.93 | 81.00% |
| Sad | 4.11(±0.07) | 4.13(±0.06) | 4.83 | 82.34% |
| Surprised | 4.23(±0.06) | 4.20(±0.06) | 4.81 | 85.67% |
| Average | 4.13(±0.07) | 4.16(±0.08) | 4.87 | 81.40% |

TABLE II
CONTROLLABLE EMOTIONAL SPEECH SYNTHESIS COMPARED TO BASELINE MODELS.

| Model | nMOS(CI)↑ | eMOS(CI)↑ | MCD↓ | ERA↑ |
|---|---|---|---|---|
| Ground truth | 4.46(±0.06) | 4.39(±0.06) | — | 97.34% |
| Emodiff | 3.76(±0.06) | 3.62(±0.05) | 5.98 | 68.77% |
| Emosphere | 3.92(±0.07) | 3.96(±0.06) | 5.63 | 72.34% |
| Ours | **4.13(±0.07)** | **4.16(±0.08)** | **4.87** | **81.40%** |

**Emodiff**: This is a diffusion-based TTS model where emotion can be manipulated by a proposed soft-label guidance technique derived from classifier guidance.

**EmosphereTTS**: This model employs Cartesian-Spherical Transformation to extract emotion vectors, enabling the synthesis of controllable emotional speech.

## IV. RESULTS

### A. Controllable Emotion Speech Synthesis

To achieve controllable emotional speech synthesis without reference audio, we use the emotion label and prosodic text embedding from PL-BERT to guide CFM, generating corresponding emotional features. In the experiments, we selected a fixed speaker and generated 300 speech samples for each emotion. The subjective metrics were obtained from 50 samples. The experimental results are shown in Table I. It can be observed that the five emotions perform well in both subjective and objective metrics. EFTTS achieves an average MCD of 4.87, with an average nMOS of 4.13 and sMOS of 4.16, demonstrating human-level performance in emotional speech synthesis. As shown in **Fig. 3**, the t-SNE visualization of the emotional features generated by CFM demonstrates that our approach effectively models emotion distribution. **Table II** demonstrates that our model outperforms the baselines in terms of both subjective and objective metrics, showcasing its robust capability for emotional speech synthesis.

### B. Zero-shot Cross-speaker Emotion Transfer

To achieve zero-shot emotion transfer, we first input the reference audio into both the emotion encoder and Conditional Flow Matching (CFM) model to obtain the emotion features as $e_{\text{ref}}$ and $e_{\text{pred}}$, respectively. These two features are then combined in a weighted manner, with the ratio controlled by a parameter $\beta$. The resulting emotion feature $\hat{e}$ is given by the following equation:

$$\hat{\mathbf{e}} = \beta \cdot \mathbf{e}_{\text{pred}} + (1 - \beta) \cdot \mathbf{e}_{\text{ref}} \tag{6}$$

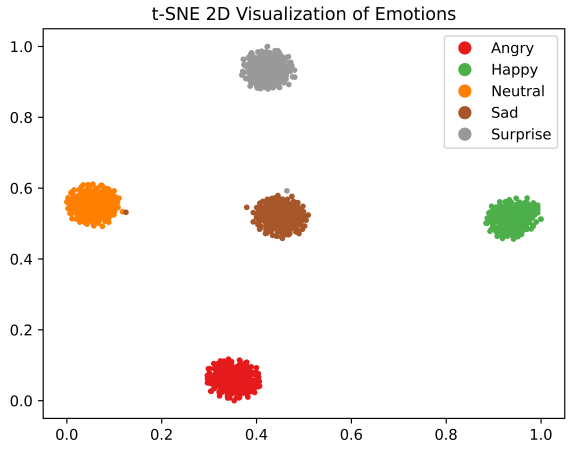


Fig. 3. **t-SNE visualization of the emotional features generated by CFM.** The distinct clustering of emotions shows the model's ability to capture and separate different emotional features.

TABLE III
RESULTS OF ZERO-SHOT EMOTION TRANSFER.

| Emotion | nMOS(CI)↑ | eMOS(CI)↑ | MCD↓ | ERA↑ |
|---|---|---|---|---|
| Angry | 4.25(±0.10) | 4.29(±0.12) | 4.73 | 97.67% |
| Happy | 4.05(±0.09) | 4.02(±0.14) | 4.91 | 84.64% |
| Netural | 4.08(±0.12) | 4.20(±0.14) | 4.87 | 97.34% |
| Sad | 4.17(±0.14) | 4.19(±0.12) | 4.86 | 99.34% |
| surprised | 4.19(±0.09) | 4.18(±0.13) | 4.75 | 98.67% |
| Average | 4.14(±0.13) | 4.18(±0.10) | 4.83 | 95.33% |

Through our experiments, we found that the best emotion transfer performance occurs when $\beta = 0.5$, as CFM predicts the corresponding emotion for the input phonemes, while the emotion encoder extracts the emotion from the reference. Both features play a crucial role in enhancing the quality of synthesized emotional speech. We selected one reference speech for each of the five emotions and randomly sampled 300 transcriptions to perform emotion transfer. Similarly, the subjective metrics were calculated from 50 samples. The subjective and objective metrics shown in **Table III** demonstrate that EFTTS achieves precise and high-quality emotion transfer with minimal loss in speech quality.

### C. Ablation Study

To validate the effectiveness of the proposed method, we conducted ablation experiments to evaluate the contributions of Emotion2Vec as the emotion encoder and CFM in generating emotional features. As a baseline, when ablation of Emotion2Vec, we employed a 3-layer simple convolutional downsampling network to transform the reference Mel spectrogram to a 128-dimension vector. In the case of ablating CFM, we used only the emotional features from the Emotion2Vec encoder, i.e., $\beta = 0$. It is important to note that, in this setting, the model still requires the input of the reference speech. Therefore, the full model results are based on the emotion transfer task. The experimental results, as shown in **Table IV**, demonstrate a significant decline in both subjective and objective metrics, with statistical significance, thereby validating the effectiveness of our method.

TABLE IV
COMPARISON METRICS OF ABLATION STUDY

| Model | nMOS | sMOS | MCD | ERA |
|---|---|---|---|---|
| Full Model | — | — | — | — |
| w/o Emotion2Vec | -0.26 | -0.33 | +0.46 | -15.67% |
| w/o CFM | -0.07 | -0.13 | +0.02 | -4.43% |

## V. CONCLUSION

In this paper, we introduced EFTTS, a novel controllable emotional Text-to-Speech model designed to address the challenges in generating high-quality, emotionally expressive speech. Our approach leverages self-supervised emotion representation through the Emotion2Vec model and utilizes conditional flow matching to model and transfer emotion features from input text. This combination allows EFTTS to produce speech that not only captures a wide range of emotional expressions but also closely aligns with the content and context of the input text, making it suitable for a variety of real-world applications in voice assistants, entertainment, and accessibility tools.

Our experimental results demonstrate that EFTTS outperforms existing TTS models in terms of both subjective and objective evaluation metrics. In particular, EFTTS achieves notable advancements in emotional expressiveness and naturalness, as evidenced by subjective and objective metrics. Furthermore, the zero-shot emotion transfer capability of EFTTS, where emotional features are transferred from reference speech highlights its potential for flexible and scalable emotional speech synthesis.

## REFERENCES

[1] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[2] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, PMLR, 2021, pp. 5530–5540.

[4] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[5] C.-B. Im, S.-H. Lee, S.-B. Kim, and S.-W. Lee, "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)*, IEEE, 2022, pp. 6317–6321.

[6] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 7254–7258.

[7] H.-S. Oh, S.-H. Lee, and S.-W. Lee, "Diffprosody: Diffusion-based latent prosody generation for expressive speech synthesis with prosody conditional adversarial training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[8] Y. Wang, D. Stanton, Y. Zhang, *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International conference on machine learning*, PMLR, 2018, pp. 5180–5189.

[9] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2019, pp. 623–627.

[10] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 5734–5738.

[11] H. Tang, X. Zhang, N. Cheng, J. Xiao, and J. Wang, "Ed-tts: Multi-scale emotion modeling using cross-domain emotion diarization for emotional speech synthesis," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 146–12 150.

[12] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8599–8608.

[13] Y. Guo, C. Du, Z. Ma, X. Chen, and K. Yu, "Voiceflow: Efficient text-to-speech with rectified flow matching," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 121–11 125.

[14] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, "Matcha-tts: A fast tts architecture with conditional flow matching," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 341–11 345.

[15] Y. Guo, C. Du, X. Chen, and K. Yu, "Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[16] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Emomix: Emotion mixing via diffusion models for emotional speech synthesis," in *INTERSPEECH 2023*, 2023, pp. 12–16. DOI: 10.21437/Interspeech.2023-1317.

[17] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[18] Y. A. Li, C. Han, V. Raghavan, G. Mischler, and N. Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "Istftnet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6207–6211.

[20] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, "Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[21] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[22] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022.

[23] Z. Ma, Z. Zheng, J. Ye, *et al.*, "Emotion2vec: Self-supervised pre-training for speech emotion representation," *arXiv preprint arXiv:2312.15185*, 2023.

[24] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[25] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[26] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021.

[27] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[28] M. Shannon, *Mel cepstral distortion (mcd) computations in python*, 2021. [Online]. Available: https://github.com/MattShannon/mcd.

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