



# EME-TTS: Unlocking the Emphasis and Emotion Link in Speech Synthesis

Haoxun Li<sup>1</sup>, Leyuan Qu<sup>1,\*</sup>, Jiaxi Hu<sup>1</sup>, Taihao Li<sup>1,\*</sup>

<sup>1</sup>Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, China

lihaoxun23@mailsucas.ac.cn, leyuan.qu@ucas.ac.cn,  
hujiaxi23@mailsucas.ac.cn, lith@ucas.ac.cn

## Abstract

In recent years, emotional Text-to-Speech (TTS) synthesis and emphasis-controllable speech synthesis have advanced significantly. However, their interaction remains underexplored. We propose Emphasis Meets Emotion TTS (EME-TTS), a novel framework designed to address two key research questions: (1) how to effectively utilize emphasis to enhance the expressiveness of emotional speech, and (2) how to maintain the perceptual clarity and stability of target emphasis across different emotions. EME-TTS employs weakly supervised learning with emphasis pseudo-labels and variance-based emphasis features. Additionally, the proposed Emphasis Perception Enhancement (EPE) block enhances the interaction between emotional signals and emphasis positions. Experimental results show that EME-TTS, when combined with large language models for emphasis position prediction, enables more natural emotional speech synthesis while preserving stable and distinguishable target emphasis across emotions. Synthesized samples are available on-line<sup>1</sup>.

**Index Terms:** Emotional Speech Synthesis, Emphasis Control, Emotion Expressiveness

## 1. Introduction

With the advancement of deep learning, Text-to-Speech (TTS) systems have significantly improved in terms of quality, clarity, and naturalness, leveraging architectures such as Transformers [1, 2], normalizing flows [3, 4], and diffusion models [5, 6]. However, conventional TTS systems often struggle with expressiveness, producing monotonous and mechanical speech. To address this issue, researchers have explored emotionally expressive speech synthesis [7], employing explicit emotion labels [8, 9, 10] and reference-based approaches [11, 12, 13] to enhance speech expressiveness.

A key aspect of expressive speech is emphasis, which highlights prominent prosodic regions through variations in pitch, phoneme duration, and spectral energy [14, 15, 16]. Several studies have introduced emphasis control in TTS [17, 18, 19], ranging from handcrafted feature integration in systems based on Hidden Markov Models (HMM) [17] to leveraging intermediate acoustic cues like pitch range [18] and variance-based features [19].

Despite significant progress in emotional TTS and emphasis-controllable synthesis, the interplay between emotion and emphasis in speech remains largely unexplored. Emphasis and emotion are intrinsically linked—emphasis modulates emotional perception by influencing prosodic patterns, while emotional states naturally determine which words are accentuated

in speech. This study develops a speech synthesis model capable of simultaneously controlling both emotion and emphasis. Specifically, we seek to answer two key questions: How can emphasis be effectively leveraged to enhance the expressiveness of emotional speech? And how can emphasis clarity and stability be preserved across different emotional conditions?

To answer the first question, we examine how predefined emphasis positions influence emotional speech synthesis. Emotional speech is shaped by both prosody and semantics, where shifts in emphasis position and intensity can alter an utterance’s emotional interpretation. While prior works such as EE-TTS [20] predicts emphasis using textual and grammatical cues, we argue that Large Language Models (LLMs) [21] can infer emphasis more effectively. Instead of focusing on semantic-based emphasis prediction, we investigate how predefined emphasis positions influence emotional speech synthesis. To this end, we annotate emphasis pseudo-labels on an emotional speech dataset and integrate an improved variance-based emphasis modeling approach. During inference, an LLM predicts emphasis positions based on emotion labels and text input. Evaluations show that our model generates more emotionally expressive speech, especially when contextual information is present.

To answer the second question, we introduce the Emphasis Perception Enhancement (EPE) block to improve emphasis clarity and stability across emotions. This block refines the interaction between emotion control signals and emphasis locations, reducing unintended interference from global emotional effects. Additionally, this design mitigates artifacts that often arise at emphasized positions, enhancing both perceptual emphasis clarity and synthesis quality.

To summarize, the main contributions of this work are as follows:

- Emotionally Controllable Emphasis Modeling: We incorporate variance-based emphasis features into an emotional TTS framework, and use weakly supervised learning with Emphasis-Class [22] pseudo-labeling on the ESD dataset [23].
- Refined Emotion-Emphasis Interaction: We propose EPE block to modulate emphasis prominence effectively, which ensures stable and clear emphasis across emotions.
- To the best of our knowledge, this study is the first to systematically investigate the relationship between emotion and emphasis in speech synthesis. EME-TTS enhances emotional expressiveness and preserves emphasis clarity with predefined emphasis positions and emotion labels.

\*Corresponding author

<sup>1</sup>[https://wd-233.github.io/EME-TTS\\_DEMO/](https://wd-233.github.io/EME-TTS_DEMO/)

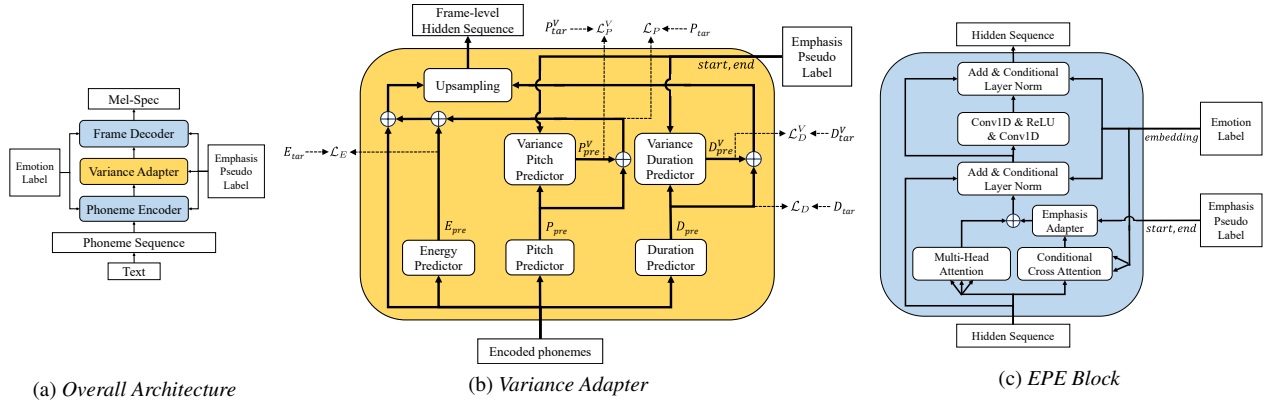


Figure 1: The entire framework of our proposed model. (a) is the overall architecture diagram. (b) and (c) show the detailed structure of variance adapter and Emphasis Perception Enhancement (EPE) block, respectively.

## 2. Proposed Method

### 2.1. Overview

The overall architecture of EME-TTS is shown in Figure 1a. We use EmoSpeech [9] as the base architecture of the acoustic model, using the embedding of emotions and the position of the emphasis as conditions. These conditions are obtained through the emotion label and the emphasis pseudo label respectively. EME-TTS consists of a phoneme encoder, a variance adapter, and a frame decoder, working in conjunction to generate emotionally expressive speech with controllable emphasis. The phoneme encoder processes the input phoneme sequence extracted from text, encoding phonetic features that serve as the foundation for speech synthesis. The variance adapter models prosodic variations through duration, pitch, and energy while integrating variance-based emphasis features for explicit emphasis control. The resulting frame-level hidden representations are then converted into a mel-spectrogram by the frame decoder.

### 2.2. Weakly Supervised Emphasis Pseudo-Labeling

Determining whether a word is emphasized within a sentence involves significant subjectivity, as multiple valid emphasis positions may exist. Consequently, collecting large-scale labeled emphasis data is both challenging and costly. Unlike EE-TTS [20], which employs a wavelet-based prosody toolset [24] to compute prominence scores via Continuous Wavelet Transform (CWT) using pitch, energy, and duration signals, we leverage the EmphAssess dataset and the EmphaClass emphasis recognizer [22]. EmphaClass fine-tunes a pre-trained Self-Supervised Learning (SSL) speech model for frame-level classification, and aggregates these scores to determine word-level emphasis with high accuracy. By utilizing EmphaClass to annotate emphasis in the ESD dataset [23], which underpins our TTS experiments, we obtain highly reliable emphasis pseudo-labels without the need for extensive manual annotation.

### 2.3. Variance Adapter

In order to integrate both fundamental prosodic predictors and variance-based emphasis modeling, the variance adapter is designed to model and regulate prosodic variations. As shown in Figure 1b, it consists of predictors for pitch, duration, and energy, which predict their respective prosodic features from encoded phonemes. Additionally, it includes variance-based

pitch and duration predictors, which refine these predictions by capturing local deviations in emphasized regions. The upsampling mechanism ensures that phoneme-level prosodic features are aligned with the frame-level hidden sequence before being passed to the decoder.

To explicitly control emphasis, we incorporate variance-based prosodic features as local modulation signals. Following the assumption in [19] that pitch and duration are the primary indicators of emphasis, with energy having a lesser impact, we focus on modeling only pitch and duration variance features, omitting energy. The computation of these variance features is formulated as follows:

$$\text{Pitch Variance} = W_{F_0} - S_{F_0} \quad (1)$$

$$\text{Duration Variance} = W_{\text{dur}} - S_{\text{dur}} \quad (2)$$

where  $W_{F_0}$  and  $W_{\text{dur}}$  represent the average pitch and duration of phonemes in emphasized regions, while  $S_{F_0}$  and  $S_{\text{dur}}$  denote the sentence-level averages. These values are derived within the Variance Pitch Predictor and Variance Duration Predictor using the emphasis pseudo-labels, which provide the *start, end* positions of emphasized words.

During training, the predicted pitch feature  $P_{\text{pre}}$  is used to compute the pitch variance feature  $P_{\text{pre}}^V$  via Equation (1), and its loss is calculated against the target pitch feature  $P_{\text{tar}}$ :

$$\mathcal{L}_P = \text{MSE}(P_{\text{pre}} + P_{\text{pre}}^V, P_{\text{tar}}) \quad (3)$$

where Mean Squared Error (MSE) measures the average squared difference between predicted and target values. Similarly, the pitch variance loss is:

$$\mathcal{L}_P^V = \text{MSE}(P_{\text{pre}}^V, P_{\text{tar}}^V) \quad (4)$$

For duration modeling, we adopt the same approach:

$$\mathcal{L}_D = \text{MSE}(D_{\text{pre}} + D_{\text{pre}}^V, D_{\text{tar}}) \quad (5)$$

$$\mathcal{L}_D^V = \text{MSE}(D_{\text{pre}}^V, D_{\text{tar}}^V) \quad (6)$$

where the variance features  $P_{\text{pre}}^V$  and  $D_{\text{pre}}^V$  are only applied to emphasized regions, with non-emphasized regions set to zero. Since direct variance calculations may introduce extreme values or negative emphasis scores, we normalize the features to the range of [0,2] based on data distribution and apply regularization for stability. The aforementioned losses, along with the energy loss, are included in the final total loss calculation.

## 2.4. Emphasis Perception Enhancement Block

The phoneme encoder and frame decoder in EME-TTS are composed of multiple stacked Emphasis Perception Enhancement (EPE) blocks, replacing the original feed-forward transformer blocks, as illustrated in Figure 1c. Each EPE block refines the modeling of emphasis perception by integrating Multi-Head Attention (MHA), Conditional Cross Attention (CCA), and Emphasis Adapter (EA), ensuring that the synthesized speech maintains clear and stable emphasis across emotions. The input hidden sequence is processed through self-attention and conditional normalization layers before being further refined by convolutional layers. Simultaneously, the emotion embedding and emphasis position can serve as external conditions.

Among these components, MHA captures global dependencies within the hidden sequence by computing multiple attention heads, allowing the model to extract contextual relationships between phonemes. CCA re-weights self-attention by incorporating emotional cues, allowing attention distributions to be adjusted based on the given emotion embedding  $c$ . CCA re-weights self-attention using:

$$Q = W_q \cdot h, \quad K = W_k \cdot c, \quad V = W_v \cdot c \quad (7)$$

$$w = \text{softmax} \left( \frac{Q \cdot K^T}{\sqrt{d}} \right) \quad (8)$$

$$\text{CCA} = w \cdot V \quad (9)$$

where  $h$  represents the input feature, and  $w$  denotes computed attention weights. In expressive speech, the prominence of certain words naturally varies with emotion, causing inconsistencies in emphasis perception. This can lead to unintended shifts in emphasis positions or even the suppression of intended emphasis under strong emotional conditions. Additionally, increasing the prominence of emphasized words may introduce artifacts and degrade synthesis quality.

To mitigate these issues, we introduce EA that refines attention distributions in predefined emphasis regions. EA explicitly enhances emphasis perceptibility while minimizing interference from emotional variations. Given an initial attention weight  $w$ , the adjusted weight is computed as:

$$w_{\text{adjusted}} = w + \Delta w, \quad \Delta w = \text{strength} \cdot \text{mask}(\text{start}, \text{end}) \quad (10)$$

where  $\text{mask}(\text{start}, \text{end})$  identifies the designated emphasis positions, and  $\text{strength}$  scales the emphasis intensity. This enhancement serves two key purposes: first, it ensures that emphasized words remain perceptually distinct, even in highly expressive speech; second, it refines emphasis representation through attention modulation rather than direct energy or pitch manipulation, thereby reducing synthesis artifacts.

For Conditional Layer Normalization (CLN), we adopt a design similar to AdaSpeech4 [25] to integrate emotional context into the normalization process, ensuring adaptive prosodic control across different emotional conditions.

## 3. Experiments

### 3.1. Experimental Setup

For our experiments, EME-TTS utilizes the English portion of the Emotional Speech Database (ESD) [23], which comprises recordings from 10 speakers across five emotions: *angry*, *happy*, *sad*, *surprise*, and *neutral*. Each speaker contributes 350 utterances per emotion, resulting in approximately 1,750 utterances and 1.2 hours of speech per speaker. We follow the

train/validation/test splits established in EmoSpeech [9], where the validation and test sets consist of 19 and 31 utterances per emotion per speaker, leading to a total of 950 and 1,550 utterances, respectively. Emphasis positions in the training data are labeled using EmphaClass [22].

The model utilizes iSTFTNet [26] as the vocoder, which is trained on the English subset of the ESD dataset [23]. The strength in EA is set to 0.2. Training is conducted on 2 Nvidia A100 GPUs and 8 RTX 4090 GPUs, with a batch size of 64 for 100,000 steps. The Adam optimizer [27] is used with a learning rate of 0.0001 and  $(\beta_1, \beta_2) = (0.5, 0.9)$ .

### 3.2. Evaluation Metrics

The evaluation framework comprises both objective and subjective assessments. The objective evaluation includes audio quality assessment and emotional accuracy measurement. The subjective evaluation consists of four tasks:

- **Emphasis Accuracy Test (EAT):** Measures how well predicted emphasis positions match listener perception.
- **Emotion Accuracy Test (EAT-EMO):** Assesses emotion recognition accuracy by comparing perceived and intended emotions.
- **Emotional Expressiveness Preference Test (EEPT):** Listeners select the most emotionally expressive sample from multiple outputs, with preference scores indicating expressiveness strength.
- **Mean Opinion Score (MOS) Rating:** Evaluates naturalness and quality on a five-point scale.

A total of 11 listeners participated in all evaluations, with each participant completing all four tasks. For all tasks, higher scores indicate better performance.

#### 3.2.1. Emphasis Accuracy

To demonstrate EME-TTS’s ability to preserve the perceptual clarity and stability of target emphasis across different emotions, we designed a more challenging subjective test as task 1 (EAT), distinct from previous studies on emphasis control [19]. Instead of rating the degree of emphasis at a predefined position, participants were asked to identify the emphasized words in 80 randomly shuffled speech samples. These samples were generated from texts in the test set using EME-TTS w/o EPE and EME-TTS. The results presented in Table 1 indicate that the emphasis produced by our proposed model was clearly perceivable across different emotions.

Notably, compared to other emotions, the *surprise* emotion posed a greater challenge for listeners in accurately identifying the emphasis position. This is attributed to the frequent pitch rise at the end of *surprise* speech, which often led participants to mistakenly perceive the emphasized word as being at the sentence’s end. The integration of the EPE effectively mitigated this issue by enhancing the prominence of the intended emphasis position. At the same time, it improved the distinctiveness of emphasis recognition across all emotions.

Table 1: *Emphasis Recognition Accuracy of Different Models.*

Model	Mean	Neutral	Angry	Happy	Sad	Surprise
EME-TTS w/o EPE	0.73	0.77	0.75	<b>0.82</b>	<b>0.75</b>	0.55
EME-TTS	<b>0.78</b>	<b>0.80</b>	<b>0.82</b>	<b>0.82</b>	<b>0.75</b>	<b>0.64</b>

### 3.2.2. Emotion Accuracy

To demonstrate that controlling emphasis in EME-TTS enhances the accuracy of perceived emotions in synthesized speech, we conducted both objective and subjective evaluations. For comparison, we included EmoSpeech [9] and CosyVoice2-0.5B-Instruct (CosyVoice2) [28] as baseline models. CosyVoice2 provides multiple inference modes; To ensure fairness, all CosyVoice2-generated samples were conditioned on a *neutral* reference speaker’s audio and a textual emotion prompt, ensuring that only the speaker’s identity and emotion labels were provided as input. During inference, our proposed model consistently utilized a large language model [29] to predict suitable emphasis positions, which were then used as input for testing.

Table 2: Objective Evaluation of Different Models on Emotion Accuracy.

Model	Mean	Neutral	Angry	Happy	Sad	Surprise
CosyVoice2 [28]	0.68	<b>0.99</b>	0.51	<b>0.73</b>	0.52	0.36
EmoSpeech [9]	0.72	<b>0.99</b>	<b>0.91</b>	0.69	0.54	<b>0.48</b>
EME-TTS w/o EPE	<b>0.74</b>	<b>0.99</b>	0.90	0.71	0.60	0.47
<b>EME-TTS</b>	0.73	<b>0.99</b>	0.87	0.70	<b>0.61</b>	0.47

From an objective perspective, we utilized the Emotion2vec-plus-large model [30] to evaluate the emotional accuracy of 1,550 synthesized audio samples from each model in the test set. The recognition outcome was assigned a score of 1 for correct classifications and 0 for incorrect ones, from which the overall accuracy was computed. The results shown in Table 2 indicate that while local emphasis control did not introduce substantial changes in global prosody, it notably improved the recognition of *sad* emotions in objective evaluation. This improvement is attributed to the increased duration of emphasized regions, which effectively enhanced emotion perception.

On the subjective side, in task 2 (EAT-EMO), participants were asked to identify the emotions of 80 randomly shuffled audio samples. Similarly, scores of 1 and 0 were assigned for correct and incorrect classifications respectively, to assess the emotional accuracy of the synthesized speech. However, as shown in Table 3, the subjective evaluation revealed unexpected shifts in emotion perception due to the introduction of emphasis. First, emphasis increased the likelihood of *neutral* speech being perceived as emotional, leading to a slight decrease in the accuracy of *neutral* emotion expression. Second, unlike CosyVoice2 [28], which struggled to synthesize *angry* and *sad* emotions in this inference mode, EME-TTS produced speech that made these emotions more perceptible, outperforming both CosyVoice2 [28] and EmoSpeech [9]. For *happy* and *surprise* emotions, EME-TTS achieved accuracy levels comparable to other models. These results demonstrate that EME-TTS achieves higher emotion accuracy in synthesized speech compared to baseline models, highlighting its overall effectiveness in generating emotionally expressive speech.

Table 3: Subjective Evaluation of Different Models on Emotion Accuracy.

Model	Mean	Neutral	Angry	Happy	Sad	Surprise
CosyVoice2 [28]	0.48	<b>0.93</b>	0.07	0.34	0.36	<b>0.70</b>
EmoSpeech [9]	0.58	0.86	0.48	<b>0.36</b>	0.50	<b>0.70</b>
EME-TTS w/o EPE	0.58	0.68	0.57	0.27	0.80	0.59
<b>EME-TTS</b>	<b>0.67</b>	0.80	<b>0.75</b>	0.32	<b>0.82</b>	0.68

### 3.2.3. Improvement of Emotional Expressiveness Through Emphasis

To assess how emphasis enhances emotional expressiveness in EME-TTS, we conducted a ranking experiment as task 3 (EEPT). Participants evaluated 30 sets of speech samples, each containing outputs from four models, based on perceived emotional expressiveness. Among them, 10 sets were derived from a short passage with contextual information. Within each set, samples were ranked from 1 (least expressive) to 4 (most expressive). Figure 2 illustrates the ranking distribution of emotional expressiveness across different TTS models. Participants evaluated speech samples based on perceived emotional expressiveness, with (a) assessing individual sentences and (b) ranking sentences within a contextualized passage. The results indicate that EME-TTS consistently received the highest rankings, especially in the contextualized setting. This suggests that surrounding linguistic context strengthens the semantic foundation for emphasis, further enhancing emotional expressiveness.

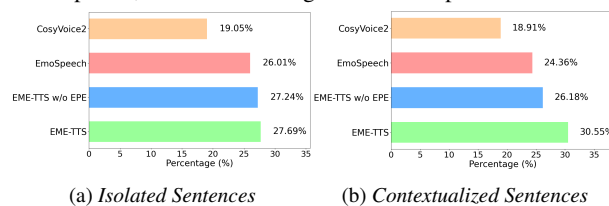


Figure 2: Effect of Emphasis on Emotional Expressiveness in TTS Models.

### 3.2.4. Speech Quality and Naturalness

We assess the quality and naturalness of the synthesized speech through both objective and subjective evaluations. Objectively, we utilize the NISQA library [31] to predict naturalness scores ratings on a 5-point scale. Subjectively, participants completed task 4 (MOS Rating), in which they rated 100 randomly shuffled speech samples for overall audio quality and naturalness, ranging from 1 (bad) to 5 (excellent). These 100 utterances were selected from the test set, ensuring an equal distribution of emotions. Table 4 indicates that EME-TTS mitigates artifacts introduced by emphasis control, enhancing synthesis quality.

Table 4: Comparison of Models for MOS and NISQA Scores.

Model	MOS ( $\uparrow$ )	NISQA ( $\uparrow$ )
Original	4.94 $\pm$ 0.03	4.17 $\pm$ 0.57
Reconstructed	4.86 $\pm$ 0.08	4.11 $\pm$ 0.58
EmoSpeech [9]	4.14 $\pm$ 0.20	3.71 $\pm$ 0.74
EME-TTS w/o EPE	3.98 $\pm$ 0.32	3.66 $\pm$ 0.68
<b>EME-TTS</b>	<b>4.22 <math>\pm</math> 0.28</b>	<b>3.76 <math>\pm</math> 0.60</b>

## 4. Conclusion

This paper presents EME-TTS, a framework that explores how emphasis enhances emotional expressiveness and how to maintain its perceptual clarity and stability across emotions. By leveraging variance-based emphasis features, weakly supervised learning, and EPE, EME-TTS demonstrate its effectiveness in generating emotionally expressive speech with clear and controllable emphasis. For future work, we aim to further investigate the role of different emphasis strategies in enhancing emotional expressiveness, particularly in achieving improvements for specific emotions where the current model’s enhancements remain limited.

## 5. Acknowledgements

This work was supported in part by the Scientific Research Staring Foundation of Hangzhou Institute for Advanced Study (2024HIASC2001), in part by Zhejiang Provincial Natural Science Foundation of China (No. LQN25F020001), and in part by the Key R&D Program of Zhejiang (2025C01104).

## 6. References

- [1] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.
- [2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [3] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.
- [4] 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.
- [5] 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.
- [6] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, "Naturalspeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [7] K. Lee. "Expressive-FastSpeech2". [Online]. Available: <https://github.com/keonlee9420/Expressive-FastSpeech2>
- [8] N. Tits, K. El Haddad, and T. Dutoit, "Exploring transfer learning for low resource emotional tts," in *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 1*. Springer, 2020, pp. 52–60.
- [9] D. Diatlova and V. Shutov, "Emospeech: guiding fastspeech2 towards emotional text to speech," in *12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 106–112.
- [10] 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.
- [11] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.
- [12] Y. Lei, S. Yang, X. Wang, and L. Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022.
- [13] J. Zaïdi, H. Seuté, B. van Niekerk, and M.-A. Carbonneau, "Daft-exprt: Cross-speaker prosody transfer on any text for expressive speech synthesis," in *Interspeech 2022*, 2022, pp. 4591–4595.
- [14] A. Eriksson and M. Heldner, "The acoustics of word stress in english as a function of stress level and speaking style," in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015), Dresden, Germany, September 6-10, 2015*, 2015, pp. 41–45.
- [15] A. Eriksson, P. M. Bertinetto, M. Heldner, R. Nodari, and G. Lenoci, "The acoustics of lexical stress in italian as a function of stress level and speaking style," in *Interspeech 2016*. The International Speech Communication Association (ISCA), 2016, pp. 1059–1063.
- [16] A. Eriksson, R. Nodari, J. Šimko, A. Suni, and M. Vainio, "Lexical stress perception as a function of acoustic properties and the native language of the listener," in *The 10th International Conference on Speech Prosody: Communicative and Interactive Prosody*. ISCA, 2020, pp. 449–453.
- [17] R. Li, Z. Wu, Y. Huang, J. Jia, H. Meng, and L. Cai, "Emphatic speech generation with conditioned input layer and bidirectional lstms for expressive speech synthesis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5129–5133.
- [18] S. Shechtman, R. Fernandez, and D. Haws, "Supervised and unsupervised approaches for controlling narrow lexical focus in sequence-to-sequence speech synthesis," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 431–437.
- [19] S. Seshadri, T. Raitio, D. Castellani, and J. Li, "Emphasis control for parallel neural tts," in *Interspeech 2022*, 2022, pp. 3378–3382.
- [20] Y. Zhong, C. Zhang, X. Liu, C. Sun, W. Deng, H. Hu, and Z. Sun, "Ee-tts: Emphatic expressive tts with linguistic information," in *Interspeech 2023*, 2023, pp. 4873–4877.
- [21] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.
- [22] M. de Seyssel, A. D'Avirro, A. Williams, and E. Dupoux, "EmphAssess : a prosodic benchmark on assessing emphasis transfer in speech-to-speech models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 495–507.
- [23] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.
- [24] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.
- [25] Y. Wu, X. Tan, B. Li, L. He, S. Zhao, R. Song, T. Qin, and T.-Y. Liu, "Adaspeech 4: Adaptive text to speech in zero-shot scenarios," in *Interspeech 2022*, 2022, pp. 2568–2572.
- [26] 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.
- [27] P. K. Diederik, "Adam: A method for stochastic optimization," *ICLR*, 2015.
- [28] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang *et al.*, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.
- [29] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [30] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 15 747–15 760.
- [31] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," in *Interspeech 2020*, 2020, pp. 1748–1752.