EE-TTS: Emphatic Expressive TTS with Linguistic Information

Anonymous submission to INTERSPEECH 2023

Abstract

While Current TTS systems perform well in synthesizing high-quality speech, producing highly expressive speech remains a challenge. Emphasis, as a critical factor in determining the expressiveness of speech, has attracted more attention nowadays. Previous works usually enhance the emphasis by adding intermediate features, but they can not guarantee the overall expressiveness of the speech. To resolve this matter, we propose Emphatic Expressive TTS (EE-TTS), which leverages multi-level linguistic information from syntax and semantics. EE-TTS contains an emphasis predictor that can identify appropriate emphasis positions from text and a conditioned acoustic model to synthesize expressive speech with emphasis and linguistic information. Experimental results indicate that EE-TTS outperforms baseline with MOS improvements of 0.49 and 0.67 in expressiveness and naturalness. EE-TTS also shows strong generalization across different datasets according to AB test results.

Index Terms: Text to speech, linguistic information, expressiveness, BERT, pre-training

1. Introduction

Over the past few years, text-to-speech (TTS) models [1, 2, 3, 4, 5, 6, 7] have made significant strides in enhancing intelligibility, quality, naturalness, and data efficiency. However, when it comes to some scenarios that require highly expressive speech, such as gaming dubbing and live streaming, original TTS systems may still generate speech with flat and mediocre prosody [8], resulting in a lack of emotional resonance. Thus, researchers are paying more attention to expressive TTS, and have attempted to enhance speech expressiveness by improving the overall prosody [9, 10] or by capturing and modeling the emotion of speech [11, 12].

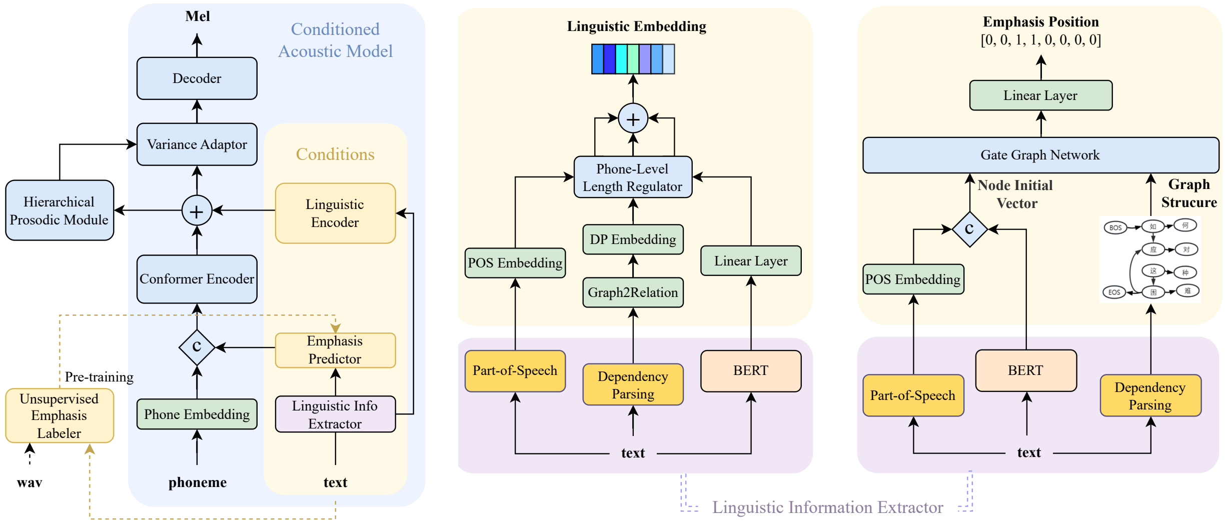
Emphasis plays a vital role in determining speech expressiveness by affecting complex variations of many aspects of speech prosody, including pitch, phoneme duration, and spectral energy [13, 14, 15]. Thus, several studies have proposed emphasizing control in TTS systems [16, 17, 18, 19, 20]. [16] enables emphasis control in an HMM-based TTS system by combining several handcrafted features. [17, 19] improve the emphasis effect by incorporating more intermediate acoustic features like pitch range [19] or variance-based features [17]. However, the overall speech expressiveness of these works is still far from the ground truth in some scenarios that need highly expressive speech. Furthermore, they are unable to synthesize emphatic expressive speech without emphasis labels, which can be expensive to obtain during inference. Therefore, [20] tried to predict emphasis position from the input text, while it still did not present a consistent controllability of emphasis and did not show a satisfied overall expressiveness of speech combined with the TTS model based on their results. According to some linguistics works [21, 22], the position and expression of emphasis highly depend on the syntax and semantics of the text. So one main reason for the above flaws from previous works is they do not consider the underlying principle or human inductive bias of emphasis like syntactic and semantic information to help the TTS model learn the distribution of expressive datasets.

In this paper, we propose Emphatic Expressive TTS (EE-TTS), a novel TTS model that utilizes linguistic information from syntax and semantics to generate emphatic expressive speech without emphasis labels. By incorporating two types of syntactic information, namely intra-word (the Part-Of-Speech (POS) of each word) and inter-word (the Dependency Parsing (DP) features) as well as semantic information extracted through the pre-trained BERT [23] model, EE-TTS fully exploits linguistic information. EE-TTS consists of 1) a linguistic information extractor to extract the syntactic and semantic information from the text; 2) an emphasis predictor to predict the positions of emphasis according to linguistic information; and 3) a conditioned acoustic model to generate expressive speech conditioned on emphasis positions and linguistic information. Besides, we use conformer [24] instead of transformer as the encoder of the acoustic model due to its relative position embedding, which is corresponding to the importance of the relative position on emphasis [22]. Given the high cost of annotating emphasis labels for text, we take advantage of massive speech-text data without emphasis labels to pre-train the emphasis predictor and acoustic model by generating emphasis labels through a signal-based method.

We conduct experiments on two mandarin TTS datasets with high expressiveness. The results show that EE-TTS can produce more expressive and natural speech with appropriate emphasis, surpassing current emphatic TTS systems by 0.49 of expressiveness and 0.67 of naturalness MOS improvements. We further carried out ablation studies to reveal the effectiveness of each aspect of linguistic information, the chosen architecture, as well as the pre-trained emphasis predictor. In a word, we conclude the main contributions of this work as follows:

- By fully exploiting linguistic information (syntax and semantics), EE-TTS can predict more reasonable emphasis positions from the text.
- Conditioned on the appropriate emphasis position and linguistic information, EE-TTS can consistently synthesize more expressive and natural speech with emphasis position.
- High robustness and great generalization ability of EE-TTS are demonstrated according to experimental results[1].

[1] Some samples of synthesized speech for reference: `https://expressive-emphatic-ttsdemo.github.io/`



(a) *EE-TTS* (b) *Linguistic Encoder* (c) *Emphasis Predictor*

Figure 1: *The entire framework of EE-TTS. The dashed lines in subfigure (a) indicate the pre-training procedure. Subfigures (b) and (c) show the detailed structure of the linguistic encoder and emphasis predictor respectively, as well as the linguistic information extractor.*

2. Proposed Method

2.1. Overview

The overall architecture of EE-TTS is shown in Figure 1a. We choose FastSpeech2 [3] as the base architecture of the acoustic model, leveraging emphasis positions and linguistic embedding as conditions. These conditions are obtained through the emphasis predictor and the linguistic encoder respectively. The linguistic information extractor generates syntactic and semantic information from the input text, which is further fed into both the emphasis predictor and the linguistic encoder. Inspired by [19], a hierarchical prosodic module is incorporated to model nuanced prosody in speech.

2.2. Linguistic Information Extractor

To involve inductive bias in linguistics, we propose linguistic information extractor, as shown in the purple blocks in Figure 1b and 1c. To extract syntactic information, we first use jieba[2] to segment the input text and then use pyltp[3] to predict Part-of-Speech (POS) tags and Dependency Parsing (DP) relations of all words at the intra-word and inter-word levels respectively. The DP result is presented graphically, with each word having only one out edge except the root word. As for semantic information, we use pre-trained BERT [23] is used to capture it at the character level. The multi-level linguistic information also reflects the hierarchical structure of the text [25].

2.3. Conditioned Acoustic Model

The acoustic model synthesized the speech conditioned on emphasis positions and linguistic embedding. By combining Convolutional neural network and Transformer, Conformer [24] can model both local and global dependencies, showing impressive performance in various tasks. Thus, we choose Conformer as the encoder for EE-TTS to benefit from its ability to model hierarchical structure. Moreover, Conformer incorporates relative positional encoding from TransformerXL [26], to handle input sequences of varying lengths and generate more accurate position information.

Both emphasis positions and linguistic embedding are generated from extracted linguistic information. The emphasis positions are given by the emphasis predictor, which is described in Section 2.4. Figure 1b illustrates the process of obtaining linguistic embedding. DP relations are first serialized by a Graph2Relation operation, which selects the type of the only one out edge of each word as the label and assigns a root label for the root word. DP relations and POS tags undergo two separate embedding layers and are then expanded with the output BERT through a phone-level length regulator to keep consistent with the size of the encoder output. These three are summed to a linguistic embedding and added to the encoder outputs.

2.4. Emphasis Predictor

Figure 1c depicts the details of the emphasis predictor. To generate the initial vector for each character node, we first embed the intra-word POS tags and expand them to the character level. After that, we concatenate them with the character-level outputs of BERT. For the DP relations, we add the BOS (Begin Of Sentence) and EOS (End of Sentence) nodes in the DP relation graph and then encode the graph with the node initial vector using a Gated Graph Neural Network (GGN) [27] to obtain the character-level features. Finally, two linear layers are employed to predict the binary classification result for each character. A value of 0 denotes the absence of emphasis on that character, while a value of 1 indicates the presence of emphasis. The character-level labels are then passed through an embedding layer and expanded to the phone level to concatenate with the phone embedding.

2.5. Pre-train with unsupervised emphasis labeling

Pre-train and finetune paradigm is quite common these days to benefit from pre-trained big models on large-scale datasets[28, 29]. However, collecting a vast amount of labeled data on emphasis is extremely challenging and expensive due to the significant subjectivity involved in determining whether a word is emphasized or not, and the ambiguity that a single sentence may

[2] https://github.com/fxsjy/jieba
[3] https://pypi.org/project/pyltp/

have more than one valid emphasis pattern. To make use of unlabeled data in a cost-effective way, we employ the Wavelet Prosody Toolkit[4] [30] to obtain pseudo emphasis labels. This toolkit is based on the continuous wavelet transform (CWT) of a weighted sum of the pitch, energy, and duration signals to calculate a prominence score for each character. These scores are quantized into two categories to indicate whether a character is emphasized or not for pre-trained datasets. Both the acoustic model and the emphasis predictor are pre-trained with these pseudo emphasis labels.

3. Experiments

3.1. Dataset

3.1.1. TTS Training Data

For pre-training, we used a mandarin dataset consisting of approximately 90,000 utterances with a total length of approximately 80 hours. The dataset includes several single-speaker datasets such as an open-source female corpus from data-baker [31] and a multi-speaker and multi-style dataset of approximately 30 hours, which includes 60 speakers and 7 different speech styles. All datasets were trimmed to remove silence at the beginning and end and were downsampled to 24 kHz. The dataset was randomly divided into a training set of 89,000 sentences and an evaluation set of 1,000 sentences.

For fine-tuning, we used a private mandarin dataset in gaming style with a total of 3,500 utterances from a female speaker (F1). About 1,800 utterances and 3,000 characters in the dataset are emphasized. To evaluate the model generalization ability to other expressive datasets, we also finetuned another dataset in live streaming style with a total of utterances 3,000 from another female speaker (F2), and about 1,200 utterances and 2,200 characters are emphasized. The emphasis positions of these two datasets are labeled by one professional annotator with sufficient training.

3.1.2. Emphasis Predictor Training Data

We use the same pre-training dataset with the TTS model for position predictor, a total of 90,000 sentences with unsupervised emphasis labels. For finetuning, we utilize the emphasis position confidence scores of the pre-trained model to filter the unsupervised emphasis labels and selected a total of 9,550 sentences with high confidence. These sentences combined with 6,500 sentences from speakers F1 and F2 to be our fine-tuning dataset of emphasis predictor.

3.2. Model Configurations

3.2.1. TTS Training Configurations

We utilize the basic configuration of the Fastspeech2 [3] for the models listed below unless otherwise explained. We choose the FastSpeech2 with an emphasis embedding and the hierarchical prosodic module [19] as the baseline to compare fairly. For our proposed model, the conformer encoder has 4 layers with both input and encoder dimensions of 256 and 2 attention heads, following the implementation and the default configurations of Espnet[5]. We use *bert-base-chinese* available on HuggingFace[6] as our pre-trained BERT and fine-tuned with our model training. We trained all models to 250,000 steps on 4 Tesla V100-16GB GPUs with batch size 64. We modified the first anneal step to 200,000 due to the size of the fine-tuning dataset. The baseline model is trained to 250,000 steps only with the fine-tuning dataset. For all the other models, we first pre-trained the model to 180,000 steps with the pre-trained dataset and then fine-tuned it to 250,000 steps. For the two models with BERT, we use an independent learning rate with exponential decay of 0.7 rather than 0.5 for FastSpeech2 to prevent the BERT module from not converging. Besides, we use a default HiFiGAN [32] trained with the fine-tuning dataset as the vocoder for all TTS models.

3.2.2. Emphasis Predictor Configurations

For the emphasis predictor, we use the same pre-trained BERT model in the TTS system and fine-tuned it with the model. The POS embedding size is set to 30. The GGN module is implemented following the default configuration with a total of 17 relations including BOS and EOS and 3 iterations, the output size is 512 following two linear layers with 128 and 2 output sizes. Our pre-training and fine-tuning processes were both trained for 50 epochs with a learning rate of 5e-5 on a Tesla V100-16GB GPU with batch size 32.

4. Results

4.1. Evaluation Metrics

We conduct several subject tests to evaluate our model comprehensively. A total of 180 utterances are randomly shuffled for MOS tests. The overall naturalness and expressiveness of speech are evaluated by a standard naturalness MOS (N-MOS) and an expressiveness MOS (E-MOS). For both MOS tests, 25 native raters read the descriptions and listened to the audio examples of each level first, then listened to all the utterances and gave 5 scaled scores from 1 to 5 based on their subjective perception of how naturalness or expressiveness of each speech. Apart from the subjective tests, we also evaluate a commonly used objective metric called Root Mean Square Error (RMSE) of Fundamental Frequency (F0) for Ablation Studies in Section 4.2. We calculate the average RMSE of the F0 in Hertz of random 100 utterances in the validation set as an auxiliary metric to evaluate which model fit the pitch prosody better. We also performed AB preference tests by 15 listeners resulting in a total of 100 utterances to verify the generalization to different datasets conveniently. All the tests listed below are done in speaker F1, except the AB preference tests are done for both F1 and F2.

4.2. Overall Performance Evaluation

Table 1: *Results of Naturalness and Expressiveness MOS for different TTS systems with 95% confidence intervals.*

Method	N-MOS	E-MOS
Ground Truth	4.66 ± 0.05	4.65 ± 0.05
Baseline (GT)	3.67 ± 0.06	3.76 ± 0.06
Proposed (GT)	4.34 ± 0.06	**4.25 ± 0.06**
Proposed (Pred)	**4.37 ± 0.06**	4.24 ± 0.06

In table 1, we compare the naturalness MOS and expressiveness MOS for Ground Truth (GT), our proposed model and baseline model with the human-labeled emphasis position of GT. We also provide the result of the proposed model with the labels predicted from our emphasis predictor. It is apparent

[4] https://github.com/asuni/wavelet_prosody_toolkit

[5] https://github.com/espnet/espnet

[6] https://huggingface.co/bert-base-chinese

that our proposed model significantly outperforms the baseline in both naturalness and expressiveness (p-values≪0.05). The MOS results of the proposed model with predicted labels even slightly outperform the one with Ground Truth labels, which may be because the position predictor tends to predict the emphasized characters that are present in the training set, and at the same time, the TTS model learns better emphasis expression on these characters.

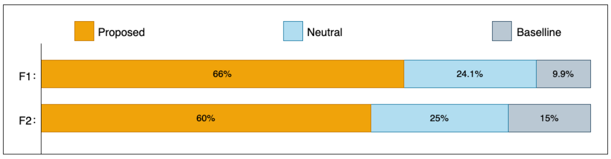


Figure 2: *AB preference results between EE-TTS and baseline of two datasets.*

Fig 2 shows the ab preference test results between EE-TTS and the baseline for two datasets, judgers prefer those speeches generated from the proposed model notably. A similar result between speaker F1 and F2 shows EE-TTS generalize to different datasets well. In the AB preference tests of two datasets, approximately 10 percent of the baseline audio samples were favored more over the proposed models, as in this portion, the positions predicted by the emphasis predictor are not reasonable enough. Unreasonable emphasis positions can lead to a decrease in the human perception of naturalness and expressiveness though the overall prosody has more variation.

4.3. Ablation Studies

We conduct ablation studies step by step to evaluate the effectiveness of each module in the proposed model. A total of 6 settings are involved for comparison: 1) EE-TTS; 2) EE-TTS equipped with a transformer encoder, not conformer, denoted as *-C+T*; Settings 3) 4) and 5) remove the BERT module, dependency parsing module, and part-of-speech module step by step, denoted as *-BERT*, *-DP*, and *-POS* respectively; 6) EE-TTS without unsupervised labels (UL) pre-training procedure.

Table 2: *Results of ablation studies of the acoustic model. C stands for Conformer and T stands for Transformer in the second setting. The last row shows the results of the EE-TTS without Unsupervised Labels(UL) during pre-training*

Method	N-MOS	E-MOS	F0-RMSE
EE-TTS	**4.34 ± 0.06**	**4.25 ± 0.06**	**53.495**
-C+T	4.19 ± 0.06	4.12 ± 0.06	56.979
-BERT	4.03 ± 0.06	4.01 ± 0.06	57.352
-DP	4.11 ± 0.06	4.05 ± 0.06	57.405
-POS	3.99 ± 0.07	4.01 ± 0.06	57.956
EE-TTS w/o UL	3.97 ± 0.06	4.04 ± 0.06	53.865

As shown in Table 2, the BERT module and the conformer encoder helped improve both the expressiveness and naturalness of speech significantly (p-values≪0.05), so did pre-training with unsupervised emphasis labels. The part-of-speech features helped only improve naturalness but did not benefit the expressiveness much, and the employment of dependency parsing seems to have no advantages. For the F0-RMSE, we can see this metric gradually decreases as each module is added step by step, and the trend matches the trend of the MOS score increase indicating the consistency of our results. However, the existence of part-of-speech and dependency parsing features do not affect the expressiveness much. It may be because these two types of information only affect where the emphasis appears but not how people express the emphasis acoustically, or due to the error propagation from the systems used to predict and extract them. We also observed a strong positive correlation between the MOS scores for expressiveness and naturalness in our results. This may owe to the fact that for audio with low naturalness, even if it has more variations, it may not resonate with the listener due to the lack of human likeness.

Ablation studies for the emphasis predictor are conducted to further reveal the effectiveness of linguistic information. The precision, recall, and F-score results of these methods are reported in Table 3. The results indicated the benefits of the performance of emphasis prediction by leveraging each kind of linguistic information. Besides, we also give a metric called reasonable precision (R-Precision) to indicate the rate of predicted positions that are reasonable for human subjective judgment. Since the location of emphasis in the speech has high variability and individuality, in real-world applications, we care more about whether the predicted emphasis locations sound appropriate to humans rather than need them exactly located at the same position as the ground truth. Similar MOS results between our proposed model with predicted labels and GT labels in Table 1 also confirm this assumption.

Table 3: *Results of ablation studies of the emphasis predictor*

Methods	Precision	Recall	F-score	R-Precision
Proposed	**0.52**	**0.63**	**0.57**	**0.87**
-DP	0.50	0.57	0.53	0.86
-POS	0.52	0.53	0.53	0.84
-BERT	0.41	0.45	0.43	0.77

5. Conclusions

In this paper, we proposed the Emphatic Expressive TTS (EE-TTS) model, a novel approach that offers a promising solution for generating both expressively and naturally emphasized speech without emphasis labels. By leveraging linguistic information such as syntax and semantics, the model predicts more reasonable emphasis positions and produces more expressive emphasized speech compared to the baseline. Furthermore, the ablation studies demonstrate the integration of the BERT model and the conformer encoder allows the model to capture semantic as well as relative position information, resulting in significant improvements in both the expressiveness and naturalness MOS. Notably, the proposed model does not require input emphasis labels and it is generalizable shown in the AB test, making it easy to apply in practice. This approach, though benefits from pre-training with unsupervised labels, still needs a certain amount of human-labeled emphasis labels for the fine-tuning dataset. In future research, our goal is to make the most of large-scale base data by enhancing the precision of unsupervised emphasis labeling tasks, which may allow us to eliminate the need for human labeling and ultimately improve the overall utilization of the data. Overall, our work fully exploits linguistic information to generate highly expressive TTS with appropriate emphasis and paves the way for future research.

6. References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 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] 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," *arXiv preprint arXiv:2205.04421*, 2022.

[6] C. Zhang, Y. Ren, X. Tan, J. Liu, K. Zhang, T. Qin, S. Zhao, and T.-Y. Liu, "Denoispeech: Denoising text to speech with frame-level noise modeling," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7063–7067.

[7] Y. Ren, C. Zhang, and Y. Shuicheng, "Bag of tricks for unsupervised text-to-speech," in *International Conference on Learning Representations*, 2023.

[8] T. Kenter, M. K. Sharma, and R. Clark, "Improving prosody of rnn-based english text-to-speech synthesis by incorporating a bert model," 2020.

[9] T. Hayashi, S. Watanabe, T. Toda, K. Takeda, S. Toshniwal, and K. Livescu, "Pre-trained text embeddings for enhanced text-to-speech synthesis." in *INTERSPEECH*, 2019, pp. 4430–4434.

[10] G. Xu, W. Song, Z. Zhang, C. Zhang, X. He, and B. Zhou, "Improving prosody modelling with cross-utterance bert embeddings for end-to-end speech synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6079–6083.

[11] Y. Wu, X. Wang, S. Zhang, L. He, R. Song, and J.-Y. Nie, "Self-supervised context-aware style representation for expressive speech synthesis," *arXiv preprint arXiv:2206.12559*, 2022.

[12] H.-W. Yoon, O. Kwon, H. Lee, R. Yamamoto, E. Song, J.-M. Kim, and M.-J. Hwang, "Language model-based emotion prediction methods for emotional speech synthesis systems," *arXiv preprint arXiv:2206.15067*, 2022.

[13] 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.

[14] A. Eriksson, A. S. Suni, M. T. Vainio, and J. Simko, "The acoustic basis of lexical stress perception," in *Proceedings of the 9th International Conference on Speech Prosody 2018*. International Speech Communications Association, 2018.

[15] 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 *Proceedings of 10th International Conference on Speech Prosody 2020, Tokyo, Japan*. ISCA, 2020.

[16] 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.

[17] S. Seshadri, T. Raitio, D. Castellani, and J. Li, "Emphasis control for parallel neural tts," *arXiv preprint arXiv:2110.03012*, 2021.

[18] L. Liu, J. Hu, Z. Wu, S. Yang, S. Yang, J. Jia, and H. Meng, "Controllable emphatic speech synthesis based on forward attention for expressive speech synthesis," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 410–414.

[19] 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.

[20] B. Stephenson, L. Besacier, L. Girin, and T. Hueber, "Bert, can he predict contrastive focus? predicting and controlling prominence in neural tts using a language model," *arXiv preprint arXiv:2207.01718*, 2022.

[21] A. Beltrama and A. Trotzke, "Conveying emphasis for intensity: Lexical and syntactic strategies," *Language and Linguistics Compass*, vol. 13, no. 7, p. e12343, 2019.

[22] Y. Xu and C. X. Xu, "Phonetic realization of focus in english declarative intonation," *Journal of Phonetics*, vol. 33, no. 2, pp. 159–197, 2005.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[24] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[25] M. Brolin, "The hierarchy of chinese grammar: A cross-sectional study of l2 chinese within processability theory," 2017.

[26] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[27] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," *arXiv preprint arXiv:1511.05493*, 2015.

[28] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie *et al.*, "Sample efficient adaptive text-to-speech," *arXiv preprint arXiv:1809.10460*, 2018.

[29] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu, "Adaspeech: Adaptive text to speech for custom voice," *arXiv preprint arXiv:2103.00993*, 2021.

[30] 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.

[31] D. Baker, "Chinese standard mandarin speech corpus," 2017.

[32] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.