Multi-granularity Semantic and Acoustic Stress Prediction for Expressive TTS

Wenjiang Chi* Xiaoqin Feng* Liumeng Xue† Yunlin Chen* Lei Xie† Zhifei Li*
* Shanghai Mobvoi Information Technology Co., Ltd, China.
E-mail: {wenjiang.chi, xiaoqin.feng, yunlinchen, zfli}@mobvoi.com
† Audio, Speech and Language Processing Group (ASLP)
E-mail: lmxue@nwpu-aslp.org, lxie@nwpu.edu.cn

***Abstract*—Stress, as the perceptual prominence within sentences, plays a key role in expressive text-to-speech (TTS). It can be either the semantic focus in text or the acoustic prominence in speech. However, stress labels are always annotated by listening to the speech, lacking semantic information in the corresponding text, which may degrade the accuracy of stress prediction and the expressivity of TTS. This paper proposes a multi-granularity stress prediction method for expressive TTS. Specifically, we first build Chinese Mandarin datasets with both coarse-grained semantic stress and fine-grained acoustic stress. Then, the proposed model progressively predicts semantic stress and acoustic stress. Finally, a TTS model is adopted to synthesize speech with the predicted stress. Experimental results on the proposed model and synthesized speech show that our proposed model achieves good accuracy in stress prediction and improves the expressiveness and naturalness of the synthesized speech.**

I. Introduction

In recent years, text-to-speech (TTS) systems have made remarkable progress in terms of speech quality and generation speed [1]–[8]. The TTS system is even capable of producing almost human-like speech. However, the expressiveness and naturalness of the synthesized speech can still be improved [9]–[11]. Prosody, which is crucial and essential for expressive speech synthesis, is mainly composed of rhythm, stress and intonation [12]–[16]. Stress, as a perceptual prominence within words or utterances, is a critical factor in prosody [17]. Therefore, accurately modeling stress is beneficial to improve the expressiveness and naturalness of TTS.

Constructing a Mandarin dataset with stress annotations is more challenging compared to constructing one in English. English is a stress-timed language, so it is relatively easy to label the prominence using a well-established English ToBI annotation system [18]. While, Mandarin is a syllable-timed language, which is characterized by syllables with the corresponding tone and pitch contour [19]. The interaction between tone, intonation, and stress makes the stress annotation of Mandarin difficult. Chinese Mandarin stress can be categorized into sentential stress and lexical stress [20], [21]. Sentential stress always reflects on certain syllables or words in need of logical semantics or emotional expressiveness. While, lexical stress, as a part of the word phonetic structure, can be used to distinguish the word sense and part of speech (POS), which is affected by acoustic features, such as pitch, duration and energy.

Generally, there are two typical approaches to acquiring stress labels from text. One is a manual approach by listening to the speech and perceiving the prominence. Li *et al*. released ASCCD, an open-source Mandarin corpus with stress annotations that are labeled by human perception [22]. The other is an automatic approach via Continuous Wavelet Transform (CWT) analysis of the speech [23]. Talman *et al.* [24] utilized the CWT to automatically extract word-level acoustic prominence of LibriSpeech [25]. Both methods are performed based on acoustic information in speech without the consideration of semantic information in the text transcription of the speech, which may result in less accuracy for stress detection and less expressivity for speech synthesis.

Based on the dataset either with manual stress annotations or automatic stress annotations, stress can be predicted from the acoustic features extracted from the speech [26], [27], such as minimum pitch, maximum pitch, duration and so on, linguistic features [19] analyzed from the text, such as tone, POS, positions of the words, or the combination of acoustic and linguistic features [20], [26]. As for TTS, it is more practical to predict stress from only text because no speech is accessible when generating speech from text. Accordingly, word embedding extracted from the text is introduced into the model of stress prediction [17], [28]. Furthermore, with the significant improvement of pre-trained language models, such as BERT [29], the learned contextualized word representation is involved in rich semantic and syntactic cues of the text, which could be beneficial to the stress prediction and the downstream TTS task. Talman *et al.* [24] employed the contextualized word representation extracted from BERT to predict stress, outperforming other models using word embedding or linguistic features. However, although it achieved an excellent result in stress prediction, the performance of TTS with the predicted stress is unknown.

In this paper, we propose a multi-granularity stress prediction model for TTS, in which stress is predicted from the contextualized word representation extracted from text via BERT and the TTS framework of VITS [8] is adopted to generate speech with stressed words. Considering the two-level nature of Mandarin stress in terms of semantic information of text and acoustic information of speech, we first build two Chinese Mandarin datasets with multi-grained stress

annotations. One is labeled with coarse-grained stress which attends to the phrase based on the semantic information of the text. The other is labeled with fine-grained stress that mainly focuses on specific words according to the prominence perception of the speech. Then, we design a multi-granularity stress prediction model to progressively predict stress from coarse-grained level to fine-grained level. Finally, we feed the predicted stress into a TTS model to synthesize speech with stressed words. We objectively and subjectively evaluate the performance of the stress prediction and the corresponding synthesized speech. Experimental results demonstrate that the proposed multi-granularity stress prediction model achieves more accurate performance and the generated speech based on the predicted stress is more expressive and natural.

II. Dataset Construction

For the standard stress corpus collection, the voice actor is often required to read well-designed utterances fluently and accurately, and then professional labelers or people from crowdsourcing projects annotate the stress according to the perceived prominence from the recordings, as shown in the left part of Figure 1. The stress expression obtained by this approach depends on the voice actor's performance and is limited by the recording environment, pre-designed script, etc. It may not suitable for real-world scenarios and is unaffordable as well. We propose a new multi-grained stress annotation method as shown in the right part of Figure 1, which is comprised of coarse-grained stress with semantic information and fine-grained stress with acoustic information. It is noted that we mainly focus on stressed and unstressed behavior instead of different stress levels [17], so we explicitly annotate the stressed words and then the remaining words are classified into unstressed words. Additionally, before labeling the stress, crowdsourcing people are trained on a subset corpus multiple times in advance, aiming to ensure intra- and inter-rater annotation consistency and reliability. The consistencies of those two-stage stress labeling after training are 76% and 85% in all annotators.

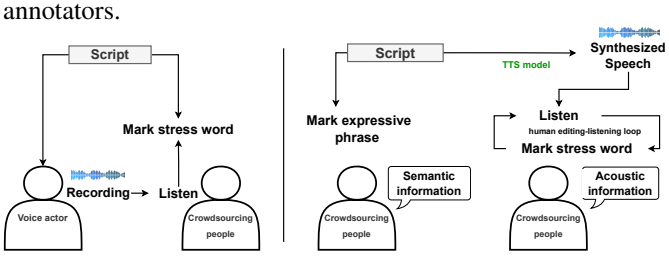

Fig. 1. The difference of stress data construction

A. Coarse-grained Semantic Stress

We first construct a coarse-grained stress corpus that centralizes semantic phrases within a sentence. To this end, people from the crowdsourcing label phrases which are potential stress according to the semantic information delivered in the text. To ensure the accuracy and consistency of the stress label, we establish semi-open annotation criteria. The specification of stress annotation are as follows:

- Mainly focus on emphasized entities, descriptive phrases, and tone conjunctions in the text. For example, the stress phrase of the sentence "年轻的母亲暴跳如雷，竟然打翻了桌子" (The young mother gets into a rage, unexpectedly knocking over the table) is "暴跳如雷" (gets into a rage) and "竟然" (unexpectedly), both of which clearly express the emotion, i.e., angry and surprise.
- To maintain the annotation consistency, we recommend annotating 1 to 3 phrases per sentence by extending or reducing the annotations based on the first rule.
- Based on our statistical results of subword length for stress character, all annotated phrases should keep a full sub-phrase length within the range of 2 to 6 characters.

B. Fine-grained Acoustic Stresss

年 轻 的 母 亲 暴 跳 如 雷， 竟 然 打 翻 了 桌 子!
年 轻 的 母 亲 暴 跳 如 雷， 竟 然 打 翻 了 桌 子!
The young mother gets into a rage, unexpectedly knocking over the table!

Fig. 2. An example of the dataset with multi-grained stress. The character in blue is semantic stress and the character in red is acoustic stress

As mentioned before, standard stress labels are annotated by listening to the recordings, in which stressed words are fixed. Consequently, such stress datasets may lack flexibility and diversity for stress prediction and speech synthesis. To solve this problem, we propose a new fine-grained stress labeling method based on acoustic information. Specifically, we first utilize a high-quality stress-controllable **TTS model** as described in Section III-C to let users set the desired stressed words in the given text to generate speech with stress. Please kindly take note that the TTS model can synthesize speech with stressed words by feeding stress tags, but it cannot predict stress. Then, we use the generated speech for fine-grained stress labeling. Additionally, we introduce a human listening-editing loop to modify the stress position and repeatedly. The human listening-editing loop consists of two circular steps: (1) Listening, where people listen to the synthesized audio generated from the above TTS model. (2) Editing, where people listen to the synthesized speech and then mark the stress word in the current transcript to make the speech perceived more natural. For example, even though "暴跳如雷" (gets into a rage) can be a semantic stress, the speech with all these 4 continuous stress characters does not natural, so finally, we're going to set the stress position at "暴". Similarly, to ensure the reliability of stress labeling, specifications of acoustic stress annotation are as follows:

- Only the locations with high stress intensity are annotated to ensure the accuracy of stress annotation.
- We recommend annotating 1 to 3 positions per sentence, and each position should be in the range of 1 to 2 characters, which maintains the annotation consistency
- To mitigate potential confusion arising from repeated listening, annotators are allowed to engage in the human

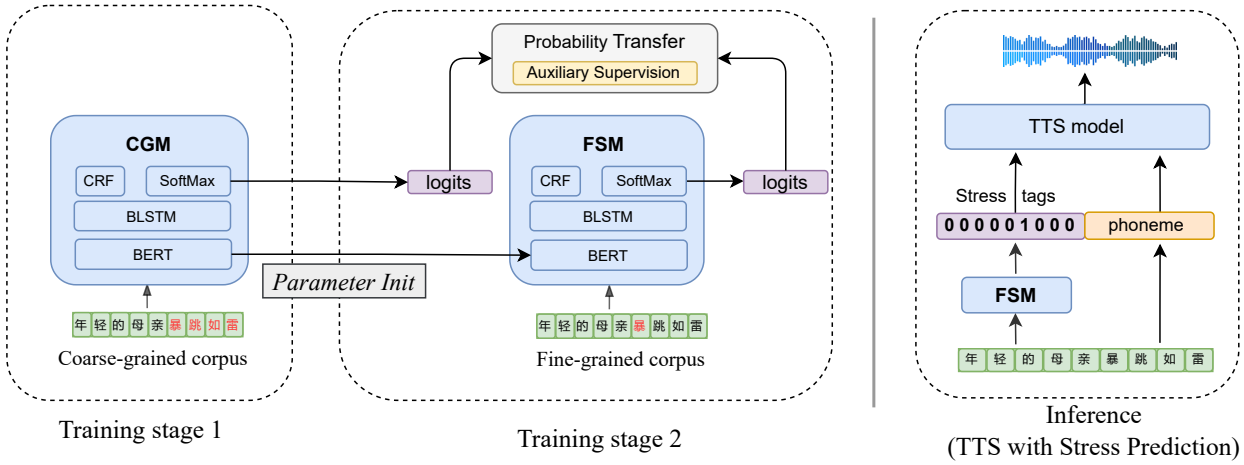

Fig. 3. The overall architecture of our proposed model, where the character in red is ground-truth stressed words. The illustration for the calculation of auxiliary supervised loss in training stage 2

listening-editing loop for each transcript a maximum of 3 times.

Figure 2 shows an example of our dataset with multi-grained stress, where the character in blue is labeled as semantic stress and the character in red is labeled as acoustic stress.

III. Multi-granularity Stress Prediction

The overall architecture of our proposed multi-granularity stress prediction model is presented in Figure 3. The coarse-grained stress and fine-grained stress provide different aspects of information within a sentence, i.e., semantic and acoustic information, respectively. To take full advantage of these two types of stress, the stress prediction model we proposed is a two-stage training model. In the first stage, we trained a coarse-grained stress model (CGM) using a coarse-grained stress corpus. In the second stage, we trained a fine-grained stress model (FSM) using a fine-grained stress corpus. We connected these two stages using two strategies: (1) initializing FSM's BERT parameter with the parameters learned from CGM, and (2) designing a coarse-grained supervision loss to maintain CGM's supervision effect and meanwhile ensure the diversity of FSM. The final stress is predicted from FSM, which is learned by transferring knowledge from CGM. Compared with the model trained on the dataset with acoustic stress only, our model incorporates stress from both semantic and acoustic information, improving the diversity of stress and also benefiting expressive speech synthesis.

A. Stage 1:Coarse-grained Semantic Stress Prediction

As shown in Figure 3, in the first stage, the coarse-grained stress model (CGM) is trained on the coarse-grained corpus as described in Section II-A to identify the salient entities within a sentence. The input of the model is Chinese Mandarin characters, which is consumed by a pre-trained Chinese BERT [30] with 12 transformer layers and 768 hidden states, followed by a Bidirectional Long Short-Term Memory (BLSTM) [31], [32] layer with 768 hidden units. In the BLSTM layer, dropout with the probability of 0.1 is adopted to prevent overfitting.

The loss function of CGM L_{CGM} includes a sequence loss of CRF [33] L_{crf} and a cross-entropy loss L_{ce}, i.e., $L_{CGM} = L_{crf} + L_{ce}$, which could better perform classification task (0 is unstressed and 1 is stressed) and meanwhile keep the contextual connection of the input character sequences. Specifically, the L_{crf} is:

$$\begin{aligned} L_{crf} &= -logP(y|x) \\ &= -(score(y) - log(\Sigma_{\hat{y}} score(\hat{y}))) \end{aligned} \tag{1}$$

where y is the predicted sequence label, x is the current input sentence, $score(y)$ is the score of the current CRF predicted path y, and $score(\hat{y})$ denotes all possible sequences of labels. Besides, L_{ce} is:

$$L_{ce} = -\sum_{i=1}^{seq\_len} \sum_{k=1}^{n\_class} y_{ik} log G1(\hat{y}_{ik}|x) \tag{2}$$

where seq_len is the length of the input sequence, n_class is the number of label classes, y_{ik} refers to the true label, $\hat{y}_{ik}$ is the k-th value of the predicted label and $G1$ refers to the output probability of the softmax layer in CGM.

B. Stage 2:Fine-grained Acoustic Stress Prediction

After acquiring the coarse-grained semantic stress in the stage 1, we further extract fine-grained acoustic stress.

In the stage 2, a fine-grained stress model (FSM) is utilized to predict stress from Chinese Mandarin characters using the fine-grained stress corpus as described in Section II-B. The structure of FSM is similar to that of CGM, but the BERT is initialized by the parameters learned from CGM. The dropout is also applied to the BLSTM layer with a probability of 0.3.

The loss function of FSM L_{FSM} is $L_{FSM} = L_{crf} + L_{cgce}$, where L_{crf} is CRF loss same as that of CGM and L_{cgce} is the auxiliary supervised loss which is designed based on the focal loss [34] . We use L_{cgce} to preserve the semantic information learned from the CGM so as to better connect the two training stages and predict stress as well. L_{cgce} is formulated as:

$$L_{cgce} = -\sum_{i=1}^{seq\_len}\sum_{k=1}^{n\_class} y_{ik} log(F2) \quad (3)$$

$$F2 = Min(\beta G2_i, 1) F1(\hat{y}_{ik}|x) \quad (4)$$

where $F1$ refers to the output probability of softmax layer in the FSM, $G2_i$ is the probability transferred from $G1_i$ for the i-th character, which is calculated by the following equation:

$$G2_i = \begin{cases} \frac{1}{\beta} & , if\ gt_i = 1 \\ G1_i & , otherwise \end{cases} \quad (5)$$

where gt_i is the ground-truth of fine-grained stress label and $G1_i$ is the predicted probability from CGM. β is the reciprocal of $1/\beta$. $1/\beta$ is the probability threshold for CGM to be 1 (if $G1_i >= 1/\beta$, then the i-th character is stressed). We use β as a regularization parameter to constrain the impact of coarse-grained supervision on the FSM.

We empirically suggest that β smaller than 10 is more effective. Moreover, the Min truncated normalization is employed to transfer the probability, preventing gradient explosion.

The process of $F2$ calculation is illustrated in Figure 4, in which $G2_i$ is transferred from $G1_i$ according to gt_i and then multiplies β and $F1$, generating $F2$. So, $F2$ is adjusted by gt_i:

- When the gt_i is 1 (stressed), the $F2_i$ will not be affected.
- When the gt_i is 0 (unstressed), the direction of $F2_i$ change is consistent with $G1_i$.

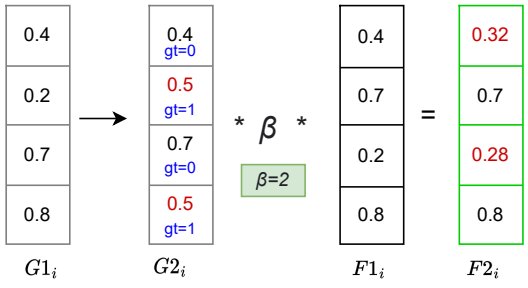

Fig. 4. Probability transfer, where **gt** means the ground-truth of fine-grained stress label and $1/\beta$ is the probability threshold for CGM to be 1.

In other words, we strengthen the L_{cgce} only when the CGM's predicted probability is larger than $1/\beta$, and vice versa, maintaining the CGM supervision effect and ensuring the diversity of the FSM itself.

C. TTS with Predicted Stress

The TTS model adopts the framework of VITS, which is able to synthesize speech with stressed words by feeding the stress tags. As shown in the right part of Figure 3, FSM takes the text as input and outputs the stress tag for each character in the text. And then, the TTS model takes the phoneme sequence extracted from the text and the predicted stress tags as input and generates speech with the corresponding stress. The TTS model is used to verify the effectiveness of the proposed model by feeding the predicted stress as the stress tag. Besides, it is also used to build the dataset with fine-grained acoustic stress (Section II-B) due to its high performance of speech generation.

IV. Experiments

A. Experimental Setups

The detailed information of the datasets, which have varying numbers of sentences, is listed in Table I. As the description of the dataset collection process in Section 2, we collected the coarse-grained stress dataset and fine-grained stress dataset separately. The coarse-grained stress dataset contains 6,817 sentences and the fine-grained stress data includes 13,615 sentences. The stress distribution of the datasets listed in the last column shows that the percentage of fine-grained stress is smaller than that of coarse-grained stress, which is consistent with our annotation specifications as described in Section II, indicating that the annotation results are valid. To train the stress prediction model, we divide each dataset into three subsets with a ratio of 8:1:1, including the training set, validation set and testing set. Moreover, the TTS model is trained on the voice actor's recordings with pre-defined stressed words. Only in this way, we can obtain a relatively reliable TTS system that can generate speech with stressed words.

TABLE I
STATISTICAL RESULTS OF THE STRESS DATASET.

Data Type	#Sentence	Stress Distribution		
		#Character	#Stress	Percentage(%)
Coarse-grained	6,817	125,645	37,063	29.50
Fine-grained	13,615	205,132	22,714	11.07

In the two stages of model training, the batch size is set as 4 for two GPUs and the learning rates of BERT and CRF are 1e-3 and 5e-5, respectively. β in the auxiliary supervised loss L_{cgce} is set to 2.

B. Experimental Results

Objective Evaluations We objectively evaluate the performance of CGM and FSM, comparing with different stress predictor models investigated in Talman *et al.* [24] using our fine-grained acoustic stress corpus. The objective evaluation results, including precision (micro), recall (micro) and F1 (micro) score, are listed in Table II. As we can see, the results show that the performance of our proposed model, i.e., FSM, is better than BERT-base, 3-layer BLSTM and CRF. We suggest that pre-trained language models, such as BERT, are more suitable for our small dataset to capture contextual semantic information. Moreover, FSM and FSM (*one-stage) has a lower F1 score than CGM, while it tends to generate more natural speech (as shown in Table III and Figure 5). The results demonstrate the challenge of directly modeling fine-grained stress, indicating the effectiveness of our proposed multi-granularity stress prediction approach.

Subjective Evaluations We carry out 5-point MOS [35] subjectively listening tests in terms of naturalness (NMOS)

TABLE II
OBJECTIVE EVALUATION RESULTS.

Model	Precision	Recall	F1
CGM	0.8471	0.9061	0.8756
FSM	0.7593	0.6762	0.7153
FSM (*one-stage)	0.6218	0.6284	0.6251
BERT-base	0.6050	0.5737	0.5890
3-layer BLSTM	0.0100	0.8471	0.0197
CRF	0.5278	0.0252	0.0482

and expressiveness (EMOS). Twenty listeners participate in the listening test, rating on 20 samples randomly selected from the testing set [1]. The subjective results are shown in Table III. Except for the first-stage stress prediction model CGM and our final stress prediction model FSM, there are other three types of generated speech for comparison: (a) **Original** represents the synthesized speech without stress; (b) **ManualSet** stands for the synthesized speech with manual annotated acoustic stress; (c) **CGM**$_{random}$ is the generated speech with a random selection of stress positions based on the results of the CGM.

TABLE III
SUBJECTIVE EVALUATION RESULTS.

Type	Original	ManualSet	CGM	CGM$_{random}$	FSM
NMOS	3.86	3.98	3.75	3.83	**3.95**
EMOS	3.9	4.08	3.83	3.88	**4.07**

From the subjective results, we can see that CGM performs poorly, resulting in unnatural and abnormal speech. This can be attributed to the fact that CGM solely focuses on semantic stress (sentential level) annotated from the text without considering of acoustic features of the speech. The final stress should be at the fine-grained level (lexical level) based on acoustic information, which is why CGM$_{random}$ outperforms CGM. The result of our proposed model FSM is close to that of the ground-truth result from ManualSet. We speculate that FSM improves the naturalness and expressiveness of TTS by leveraging both semantic and acoustic information.

C. Ablation Studies

In this section, we conduct ablation studies on the auxiliary supervised loss L_{cgce} and two-stage training strategy. Specifically, we replace L_{cgce} with L_{ce}, i.e., w/o L_{cgce}, and further use the parameters of pre-trained Chinese BERT instead of CGM to initialize FSM, i.e., w/o CGM. Please kindly take note that 'w/o CGM' denotes the utilization of a one-stage training approach solely with the fine-grained stress corpus (acoustic stress information).

The ablation results of objective evaluations are shown in Table IV. It can be seen that the initialization of CGM can enhance the effect of the FSM a little bit. However, the auxiliary supervision loss of CGM significantly improves the performance of the FSM, achieving the highest F1 score of 71.53% with 7% to 8% improvement. We guess that while BERT can provide abundant semantic information, the CGM model still offers remarkable stress-related information in the sentence, thereby facilitating the FSM in detecting fine-grained stress keywords. The subjective preference test results, as depicted in Figure 5, demonstrate that the speech produced by FSM trained on a two-stage approach with coarse-grained information exhibits greater naturalness and expressiveness. Overall, the results demonstrate that the incorporation of coarse-grained information supervision is an effective approach to preventing stress weight dispersion during training.

[1]Samples can be found at https://xqfeng-josie.github.io/stress/

TABLE IV
ABLATION RESULTS OF OBJECTIVE EVALUATION

Model	Loss	Model$_{\text{init}}$	Precision	Recall	F1
FSM	L_{cgce}	CGM	0.7593	0.6762	0.7153
FSM	w/o L_{cgce}	CGM	0.6981	0.5780	0.6335
FSM (*one-stage)	w/o L_{cgce}	w/o CGM	0.6218	0.6284	0.6251

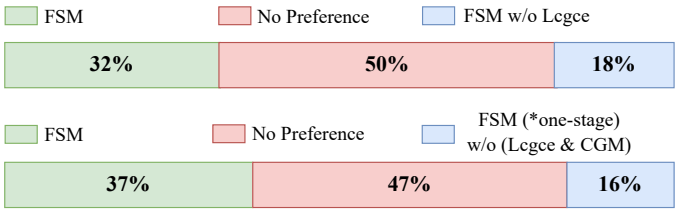

Fig. 5. Preference test results for ablation studies.

V. CONCLUSIONS

In this work, we propose a multi-granularity stress prediction model to improve the naturalness and expressiveness of TTS. In consideration of the nature of multi-grained stress, we construct the dataset with coarse-grained semantic stress and fine-grained acoustic stress to incorporate both the semantic knowledge in the text and the acoustic information in the speech. Then, the two-stage stress prediction model progressively predicts the coarse-grained stress and fine-grained stress from text, where the pre-trained language model is adopted to extract contextualized word representation from the text. Experimental results objectively and subjectively show that our proposed stress prediction model gets a higher F1 score and achieves more natural and expressive synthetic speech. In the future, we will further improve the multi-stage stress prediction model to close the performance gap between human speech and synthetic speech.

REFERENCES

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4779–4783.

[3] 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.

[4] 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.

[5] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," *arXiv preprint arXiv:2006.03575*, 2020.

[6] C. Yu, H. Lu, N. Hu, *et al.*, "Durian: Duration informed attention network for multimodal synthesis," *arXiv preprint arXiv:1909.01700*, 2019.

[7] Z.-H. Ling, S.-Y. Kang, H. Zen, *et al.*, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.

[8] 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.

[9] X. Li, C. Song, J. Li, Z. Wu, J. Jia, and H. Meng, "Towards multi-scale style control for expressive speech synthesis," *arXiv preprint arXiv:2104.03521*, 2021.

[10] 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.

[11] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, "Controllable cross-speaker emotion transfer for end-to-end speech synthesis," *arXiv preprint arXiv:2109.06733*, 2021.

[12] R. Skerry-Ryan, E. Battenberg, Y. Xiao, *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*, PMLR, 2018, pp. 4693–4702.

[13] 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.

[14] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6945–6949.

[15] Y. Wang, R. Skerry-Ryan, Y. Xiao, *et al.*, "Uncovering latent style factors for expressive speech synthesis," *arXiv preprint arXiv:1711.00520*, 2017.

[16] Y. Bian, C. Chen, Y. Kang, and Z. Pan, "Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis," *arXiv preprint arXiv:1904.02373*, 2019.

[17] Y. Zheng, Y. Li, Z. Wen, B. Liu, and J. Tao, "Text-based sentential stress prediction using continuous lexical embedding for mandarin speech synthesis," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2016, pp. 1–5.

[18] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, *et al.*, "Tobi: A standard for labeling english prosody.," in *ICSLP*, vol. 2, 1992, pp. 867–870.

[19] Y. Li and J. Tao, "Mandarin stress analysis and prediction for speech synthesis," in *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, Springer, 2015, pp. 83–95.

[20] Shao, Yanqiu, Han, *et al.*, "Study on automatic prediction of sentential stress for chinese putonghua text-to-speech system with natural style," 2007.

[21] M. Chu and M. Bao, "Comparison of sentential-stress allocation within base phrases among different reading styles," in *Speech Prosody 2004, International Conference*, 2004.

[22] L. Aijun, C. Xiaoxia, S. Guohua, *et al.*, "The phonetic labeling on read and spontaneous discourse corpora," in *proceedings of International Conference on Spoken Language Processing (ICSLP)*, Citeseer, 2000.

[23] 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.

[24] A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, and M. Vainio, "Predicting prosodic prominence from text with pre-trained contextualized word representations," *arXiv preprint arXiv:1908.02262*, 2019.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[26] C. Ni, W. Liu, and B. Xu, "Mandarin pitch accent prediction using hierarchical model based ensemble machine learning," in *2009 IEEE Youth Conference on Information, Computing and Telecommunication*, IEEE, 2009, pp. 327–330.

[27] S. Shechtman and M. Mordechay, "Emphatic speech prosody prediction with deep lstm networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5119–5123. DOI: 10.1109/ICASSP.2018.8462473.

[28] Y. Mass, S. Shechtman, M. Mordechay, *et al.*, "Word Emphasis Prediction for Expressive Text to Speech," in *Proc. Interspeech 2018*, 2018, pp. 2868–2872. DOI: 10.21437/Interspeech.2018-1159.

[29] 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.

[30] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for chinese bert," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021.

[31] P. Zhou, W. Shi, J. Tian, *et al.*, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2016, pp. 207–212.

[32] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[33] T. Müller, H. Schmid, and H. Schütze, "Efficient higher-order crfs for morphological tagging," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 322–332.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[35] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for storytelling applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1137–1144, 2006.