A CHARACTER-LEVEL SPAN-BASED MODEL FOR MANDARIN PROSODIC STRUCTURE PREDICTION

Xueyuan Chen[1,†], Changhe Song[1,†], Yixuan Zhou[1], Zhiyong Wu[1,2,],*
Changbin Chen[3], Zhongqin Wu[3], Helen Meng[1,2]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
[3] TAL Education Group, Beijing, China

{chenxuey20, sch19, zhouyx20}@mails.tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk, {chenchangbin, wuzhongqin}@tal.com

ABSTRACT

The accuracy of prosodic structure prediction is crucial to the naturalness of synthesized speech in Mandarin text-to-speech system, but now is limited by widely-used sequence-to-sequence framework and error accumulation from previous word segmentation results. In this paper, we propose a span-based Mandarin prosodic structure prediction model to obtain an optimal prosodic structure tree, which can be converted to corresponding prosodic label sequence. Instead of the prerequisite for word segmentation, rich linguistic features are provided by Chinese character-level BERT and sent to encoder with self-attention architecture. On top of this, span representation and label scoring are used to describe all possible prosodic structure trees, of which each tree has its corresponding score. To find the optimal tree with the highest score for a given sentence, a bottom-up CKY-style algorithm is further used. The proposed method can predict prosodic labels of different levels at the same time and accomplish the process directly from Chinese characters in an end-to-end manner. Experiment results on two real-world datasets demonstrate the excellent performance of our span-based method over all sequence-to-sequence baseline approaches.

Index Terms— prosodic structure prediction, text-to-speech, character-level, tree structure, span-based decoder

1. INTRODUCTION

Prosodic structure prediction (PSP) is crucial to the naturalness of synthesized speech in text-to-speech (TTS) system [1]. In general, prosodic structure is defined as a three-level tree structure including prosodic word (PW), prosodic phrase (PPH) and intonational phrase (IPH) [2]. For a typical Mandarin TTS system, word segmentation (WS) is performed before PSP to convert continuous characters into lexical words with part-of-speech (POS) [3]. As a result, error accumulation from word segmentation may affect the performance of PSP.

With the development of statistical machine learning methods on PSP, like maximum entropy [4], conditional random field (CRF) [5, 6] and deep recurrent neural network (RNN) [7, 8, 9, 10], the best reported results are achieved with bidirectional long short-term memory (BLSTM) [7, 8, 9] and CRF [11] respectively. Further an end-to-end model BLSTM-CRF [12] is proposed to integrate these two methods, which adopts BLSTM as encoder to model linguistic features from raw text and uses CRF as decoder to predict prosodic labels for each character. As PSP task is usually decomposed into three independent two-category classification sub-tasks for PW, PPH and IPH respectively, multi-task learning (MTL) framework is utilized to unify these three sub-tasks for overall optimization [13]. Word segmentation task is also set as an auxiliary sub-task within the MTL framework [14] to avoid error accumulating when conducted separately. Further, RNNs are replaced by multi-head self-attention, which is highly parallel and connects two arbitrary characters directly regardless of their distance [15]. Moreover, using pre-trained embedding mapping to replace traditional linguistic features (like POS, word length) can enhance the prediction performance [16], and bidirectional encoder representations from transformers (BERT) can provide more linguistic information [14].

However, existing methods generally transform PSP task into a sequence-to-sequence (Seq2Seq) problem and ignore the tree dependency of prosodic structure among PW, PPH and IPH. These Seq2Seq models require complex feature engineering to guarantee the effectiveness and the practicality of inference. Moreover, the inevitable class imbalance seriously damages the performance of Seq2Seq models, where the positive labels are much less than the negative labels for all these three two-category classification sub-tasks. On the other hand, researchers have consistently found that the best performance is achieved by systems that explicitly require the decoder to generate well-formed tree structures [17, 18].

Inspired by the success of span-based parser in syntactic tree prediction task [19, 20, 21] and the high correlation between syntactic structure and prosodic structure [17], we propose a span-based Mandarin PSP model to obtain an optimal prosodic structure tree, which can be converted to relevant prosodic label sequence. Instead of the prerequisite word segmentation process, rich linguistic features are extracted from Chinese character-level BERT and sent to encoder with self-attention architecture. On top of this base, span representation and label scoring are used to describe all possible prosodic structure trees, of which each tree has its corresponding score. A bottom-up CKY-style algorithm is further used to find the optimal tree with the highest score for a given sentence.

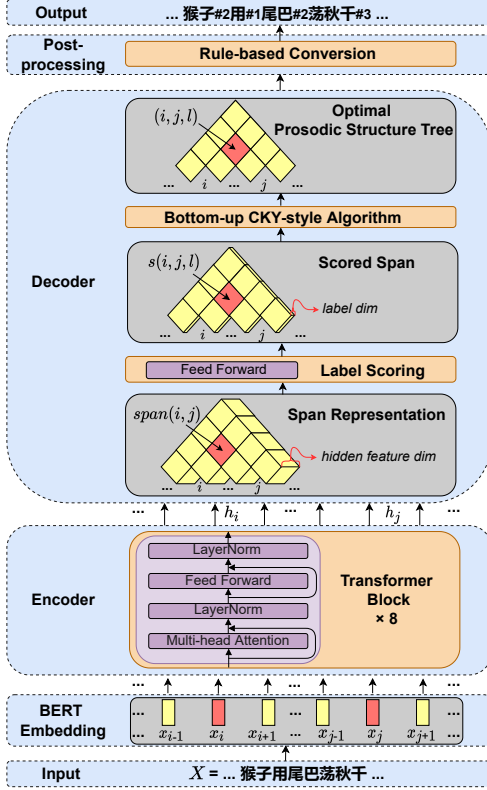† Equal contributions.
* Corresponding author.

Fig. 1. Proposed model structure with span-based decoder.

Our proposed end-to-end model adopts prosodic structure tree to unify the three sub-tasks of PW, PPH and IPH, which can predict prosodic labels of different levels at the same time and accomplish the process by directly accepting Chinese characters as input. Experiment results on two real-world datasets demonstrate the excellent performance of our proposed character-input model over all Seq2Seq baseline approaches that exploit additional WS and POS input, even on the limited amount of TTS text data. We further investigate the influence of linguistic span representation in our proposed method, and experiments indicate that better task-specific linguistic representation is also essential to enhance the performance of PSP.

2. METHODOLOGY

As illustrated in Fig.1, our proposed model is composed of four components: BERT embedding, Transformer-based encoder, span-based decoder and post-processing.

2.1. Transformer-based encoder

The input text X is first embedded with a well-pretrained character-level BERT model [22] that is composed of a stack of Transformer blocks and pretrained with a huge amount of Mandarin text data.

Then, we use several Transformer blocks [23] to obtain a sequence of context-sensitive hidden vectors (denoted as h_i and h_j for x_i and x_j, respectively) as the linguistic representation. Each block consists of two stacked sublayers: a multi-head attention mechanism and a position-wise feed-forward sublayer. And each sublayer is followed by a residual connection and a Layer Normalization.

2.2. Span-based decoder

Inspired by the the success of span-based constituency parser [19, 20, 21], we define all the potential prosodic structure trees and their corresponding scores with span representation and label scoring for each given sentence. A CKY-style dynamic programming is utilized during inference to find the optimal tree with the highest score. The model is trained with margin constraints under the inference scheme.

2.2.1. Prosodic structure tree

A prosodic structure tree can be regarded as a collection of labeled spans over a sentence. A labeled span is denoted by a triplet (i, j, l) with i and j referring to the beginning and ending positions of a span (i, j) whose label $l \in \mathscr{L}$. Here, $\mathscr{L}$ is the generalized prosodic label set of size L. One thing to note is that we assign additional atomic labels to the nested prosodic constituents, such as the case that span (i, j) with a label '#2-#1' is both PW and PPH. Thereafter, a prosodic structure tree can be represented as follows:

$$T := \{(i_t, j_t, l_t) : t = 1, ..., |T|\}. \quad (1)$$

As shown in Fig.1, the representation of a span (i, j) is defined as the substraction of hidden vectors: $h_{ij} = h_i - h_j$. In our work, a label scoring procedure is further adopted to evaluate the score $s(i, j, l)$ of assigning the label l to span (i, j), which is implemented by feeding the span representation through a feedforward network whose output dimension equals to the number of possible labels:

$$s(i, j, l) = [W_2 g(W_1 h_{ij} + z_1) + z_2]_l, \quad (2)$$

where g is ReLU activation function. On this basis, we assign a real-valued score $s(T)$ to each possible prosodic structure tree T as:

$$s(T) = \sum_{(i,j,l) \in T} s(i, j, l). \quad (3)$$

The tree with the highest above score can be then regarded as the optimal prosodic structure tree.

2.2.2. Dynamic programming for finding optimal tree

A bottom-up CKY-style algorithm is adopted to find the optimal tree. Fig.2(a) gives an illustration of the process, which is described in detail below.

As the example illustrates in Fig.2(b), a prosodic structure tree is an n-ary tree. The bottom-up CKY-style algorithm handles n-ary trees by binarizing them. For the nodes that are created during the course of binarization but not themselves full prosodic constituents, a dummy label $\varnothing$ is introduced to them with the property that $\forall i, j : s(i, j, \varnothing) = 0$. Under this scheme, let $s_{best}(i, j)$ denote the best score of the subtree for span (i, j). For a general given span (i, j), we independently assign it a best label $l \in \mathscr{L}$ and pick a best split point k $(i < k < j)$, then repeat this recursion process for the left and right subspans:

$$s_{best}(i, j) = \max_l s(i, j, l) + \max_k [s_{best}(i, k) + s_{best}(k, j)]. \quad (4)$$

While for $j = i + 1$, the best score only relies on the label l:

$$s_{best}(i, i + 1) = \max_l s(i, i + 1, l). \quad (5)$$

We can finally compute $s_{best}(0, n)$ for a sentence of length n, and reconstruct the prosodic structure tree through traversing split points of that optimal computation process. Nodes assigned $\varnothing$ are omitted during the reconstruction process to obtain the optimal full n-ary tree as Fig.2(b) shows. The overall complexity is $O(n^3 + Ln^2)$.
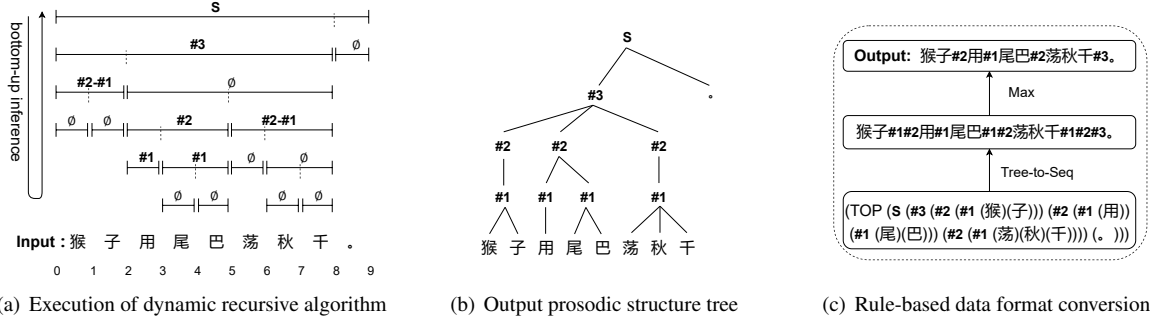
(a) Execution of dynamic recursive algorithm   (b) Output prosodic structure tree   (c) Rule-based data format conversion

Fig. 2. The execution of bottom-up algorithm and post-processing. The input Mandarin text means 'the monkey swings with its tail.'

2.2.3. Margin Training

As is common for structured prediction problems [24], given the gold tree T^*, the model is trained under the above inference scheme to make all trees T satisfy the margin constraints:

$$s(T^*) \geq s(T) + \Delta(T, T^*), \tag{6}$$

by minimizing the hinge loss:

$$\max(0, \max_T[s(T) + \Delta(T, T^*)] - s(T^*)), \tag{7}$$

where Δ is the Hamming loss on labeled spans. It can be performed efficiently using a slight modification of the dynamic programming, which replaces $s(i, j, l)$ with $s(i, j, l) + 1\,[l \neq l_{ij}^*]$, where l_{ij}^* is the label of span(i, j) in the gold tree T^*.

2.3. Post-processing

After the span-based decoder, a simple but important rule-based data format conversion module is connected to convert the intermediate representation of tree into the prosodic label sequence.

As shown in Fig.2(c), for each level, we first change the positions of the prosodic labels to the end of the corresponding prosodic level components. In this way, we can obtain a sequence that contains multi-level prosodic labels. Then, we merge the multi-level labels after each PW, and only retain the highest-level labels, considering these labels contain the information of the lower-level labels. This coercive method ensures that we can finally get a well-formed prosodic label sequence in spite of the local label confusion caused by insufficient training in some cases.

3. EXPERIMENT

3.1. Datasets

There is no common public dataset for the PSP task. We select two datasets with great difference in label distribution, sentence length and dataset size to evaluate the performance and generalization ability of our proposed method. Prosodic boundaries of all the sentences are manually labeled by annotators through reading the transcriptions and listening to the utterances. One of them is a large home-grown dataset from People's Daily Corpus mainly composed of long sentences with WS and POS, and the other is a limited public dataset from Databaker [25] mainly composed of short sentences without

WS and POS. Both datasets are divided into training, validation, and test with the ratio of 8:1:1. The details are shown in Table 1.

Table 1. Basic statistical information of datasets. 'S_{num}', 'S_{max}', 'S_{ave}' means the number, max length and average length of sentences, while 'PW_{num}', 'PPH_{num}', 'IPH_{num}' means the total number of PW, PPH and IPH within each dataset.

	S_{num}	S_{max}	S_{ave}	PW_{num}	PPH_{num}	IPH_{num}
Home-grown	100,000	365	38	1,627,371	586,847	290,714
Databaker	10,000	37	17	74,093	33,963	19,746

3.2. Compared methods

We have selected the most representative methods in recent years as baselines for comparison. They are all Seq2Seq methods and treat PW, PPH and IPH as three subtasks. Additional WS and POS are further added in all baseline methods on home-grown dataset. All models use pre-trained Chinese character-level BERT embedding[1] with frozen parameters.

- **BLSTM-CRF**: BLSTM-CRF based model achieving widely accepted results is used as one of our baselines [12].
- **Transformer-Softmax**: It uses Transformer as encoder, followed by a two-dimensional Softmax as decoder [15]. PW, PPH, IPH labels are predicted as three independent sub-tasks.
- **Transformer-Softmax-MTL**: Based on above *Transformer-Softmax*, it introduces MTL to connect three independent sub-tasks by concatenating the output of the previous task onto the input of the next task for overall optimization [13].
- **Transformer-CRF**: It uses Transformer as encoder, and followed by CRF to predict prosodic labels.
- **Transformer-Tree**: Our proposed method, which uses Transformer-based encoder and span-based decoder. Code and trained models are publicly available.[2]

For all Transformer related models, the encoders are the same as that of **Transformer-Tree** as described in section 2.1. Taking punctuation into account, the 5-category WS (i.e. B, M, E, S, P) is embedded to 5-dim WS embedding and 100-category POS is embedded to 50-dim POS embedding, which are both concatenated onto the vector of each character as input.

---

[1]https://huggingface.co/bert-base-chinese
[2]https://github.com/thuhcsi/SpanPSP

Table 2. Results on Home-grown People's Daily dataset.

	WS & POS	PW Pre	PW Rec	PW F1	PPH Pre	PPH Rec	PPH F1	IPH Pre	IPH Rec	IPH F1
BLSTM-CRF	w/o	94.46%	95.18%	94.82%	82.14%	76.06%	78.99%	94.43%	95.90%	95.16%
	w/	96.54%	97.80%	97.17%	82.30%	77.99%	80.09%	95.16%	96.42%	95.78%
Transformer-Softmax	w/o	95.32%	95.64%	95.48%	82.83%	78.33%	80.52%	95.18%	96.69%	95.93%
	w/	97.39%	98.37%	97.88%	81.95%	81.19%	81.57%	95.03%	97.07%	96.09%
Transformer-Softmax-MTL	w/o	95.18%	96.69%	95.93%	81.49%	80.34%	80.91%	95.09%	96.93%	96.00%
	w/	97.57%	98.21%	97.89%	82.79%	81.27%	82.02%	95.16%	97.08%	96.11%
Transformer-CRF	w/o	96.25%	96.39%	96.32%	83.30%	82.04%	82.66%	95.57%	97.06%	96.31%
	w/	98.22%	98.64%	98.43%	85.59%	84.21%	84.89%	95.71%	97.52%	96.60%
Transformer-Tree	w/o	**99.29%**	**99.12%**	**99.21%**	**91.23%**	**90.50%**	**90.86%**	**97.98%**	**99.03%**	**98.50%**
△: Ours - Best Previous	–	+1.07%	+0.48%	+0.78%	+5.64%	+6.29%	+5.97%	+2.27%	+1.51%	+1.90%

Table 3. Results on Databaker dataset.

	PW Pre	PW Rec	PW F1	PPH Pre	PPH Rec	PPH F1	IPH Pre	IPH Rec	IPH F1
BLSTM-CRF	87.58%	87.19%	87.38%	76.62%	73.73%	75.15%	94.72%	85.86%	90.07%
Transformer-Softmax	84.46%	86.60%	85.52%	77.84%	73.09%	75.39%	94.59%	84.68%	89.36%
Transformer-Softmax-MTL	84.73%	85.30%	85.01%	77.36%	73.82%	75.55%	94.88%	85.65%	90.03%
Transformer-CRF	89.34%	88.26%	88.80%	78.60%	74.66%	76.58%	94.86%	86.24%	90.34%
Transformer-Tree	**97.52%**	**96.96%**	**97.24%**	**88.06%**	**87.36%**	**87.71%**	**95.13%**	**87.26%**	**91.03%**
△: Ours - Best Previous	+8.18%	+8.70%	+8.44%	+9.46%	+12.70%	+11.13%	+0.25%	+1.02%	+0.69%

3.3. Results and analysis

To compare the methods, we use Precision, Recall and F1 score as metrics to evaluate the performance of prosodic label prediction. Experiment results on both datasets are shown in Table 2 and Table 3 respectively. It can be observed that **Transformer-CRF** has achieved the best results among all the Seq2Seq baseline methods, and adding the ground truth of WS and POS can effectively improve the performance. **Transformer-Tree** significantly outperforms all baselines on all metrics, especially for PPH on both datasets.

Compared with the independent two-category classification methods, MTL method simply combines the three sub-tasks of PW, PPH and IPH, and is still a Seq2Seq method with limited effect. Our proposed end-to-end model adopts prosodic structure tree to unify the three sub-tasks, which can predict prosodic labels of different levels at the same time and accomplish the process by directly accepting Chinese characters as input in an end-to-end manner. The proposed method achieves the best performance over all Seq2Seq baseline approaches that exploit additional WS and POS input. For the limited TTS text data, our proposed method also achieves the best results. All these results demonstrate that our character-level span-based model does not rely on complex feature engineering, and eliminates the possibility of error accumulation caused by prerequisite word segmentation module, which is more conducive to practical applications.

3.4. Investigation on linguistic representation

To further investigate the influence of linguistic span representation in our proposed **Transformer-Tree**, we have tried two other settings based on the proposed method: i) **Proposed - Encoder**: the Transformer-based encoder is removed, and BERT embedding is passed to span-based decoder directly; ii) **Proposed + Fine-tuning BERT**: BERT parameters are not frozen and embedding can be fine-tuned during training.

As shown in Table 4, we observe that the performance has dropped significantly on all metrics when the Transformer-based encoder is removed, and the better result can be achieved if BERT embedding is further fine-tuned according to our target task on the basis of **Transformer-Tree**. Both introducing Transformer-based encoder and fine-tuning BERT embedding are helpful for learning more suitable character embedding, which indicates that better task-specific linguistic span representation is also essential to enhance the performance of prosodic structure prediction.

Table 4. Results of investigation on linguistic representation.

	Home-grown PW-F1	Home-grown PPH-F1	Home-grown IPH-F1	Databaker PW-F1	Databaker PPH-F1	Databaker IPH-F1
Proposed	99.21%	90.86%	98.50%	97.24%	87.71%	91.03%
- Encoder	98.41%	82.61%	95.05%	95.10%	84.30%	88.07%
+ Fine-tuning BERT	**99.47%**	**92.16%**	**98.72%**	**97.55%**	**89.11%**	**92.78%**

4. CONCLUSION

In this paper, we propose a character-level span-based Mandarin prosodic structure prediction model to obtain an optimal prosodic structure tree, which can be converted to corresponding prosodic label sequence. The proposed method can predict prosodic labels of different levels at the same time and accomplish the process directly from Chinese characters in an end-to-end manner. Experiment results on two real-world datasets demonstrate the excellent performance over all Seq2Seq baseline approaches that exploit additional WS and POS input. And we further find that better task-specific linguistic representation is also essential to enhance the performance of prosodic structure prediction.

5. REFERENCES

[1] Zhigang Chen, Guoping Hu, and Wei Jiang, "Improving prosodic phrase prediction by unsupervised adaptation and syntactic features extraction," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[2] Min Chu and Yao Qian, "Locating boundaries for prosodic constituents in unrestricted mandarin texts," in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 6, Number 1, February 2001: Special Issue on Natural Language Processing Researches in MSRA*, 2001, pp. 61–82.

[3] Yao Du, Zhiyong Wu, Shiyin Kang, Dan Su, Dong Yu, and Helen Meng, "Prosodic structure prediction using deep self-attention neural network," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 320–324.

[4] Jian-Feng Li, Guo-ping Hu, and Renhua Wang, "Chinese prosody phrase break prediction based on maximum entropy model," in *Eighth International Conference on Spoken Language Processing*, 2004.

[5] Gina-Anne Levow, "Automatic prosodic labeling with conditional random fields and rich acoustic features," in *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008.

[6] Yao Qian, Zhizheng Wu, Xuezhe Ma, and Frank Soong, "Automatic prosody prediction and detection with conditional random field (crf) models," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 135–138.

[7] Chuang Ding, Lei Xie, Jie Yan, Weini Zhang, and Yang Liu, "Automatic prosody prediction for chinese speech synthesis using blstm-rnn and embedding features," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 98–102.

[8] Asaf Rendel, Raul Fernandez, Ron Hoory, and Bhuvana Ramabhadran, "Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5655–5659.

[9] Yibin Zheng, Ya Li, Zhengqi Wen, Xingguang Ding, and Jianhua Tao, "Improving prosodic boundaries prediction for mandarin speech synthesis by using enhanced embedding feature and model fusion approach.," in *INTERSPEECH*, 2016, pp. 3201–3205.

[10] Yuchen Huang, Zhiyong Wu, Runnan Li, Helen Meng, and Lianhong Cai, "Multi-task learning for prosodic structure generation using blstm rnn with structured output layer.," in *INTERSPEECH*, 2017, pp. 779–783.

[11] Andrew Rosenberg, Raul Fernandez, and Bhuvana Ramabhadran, "Phrase boundary assignment from text in multiple domains," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[12] Yibin Zheng, Jianhua Tao, Zhengqi Wen, and Ya Li, "Blstm-crf based end-to-end prosodic boundary prediction with context sensitive embeddings in a text-to-speech front-end.," in *Interspeech*, 2018, pp. 47–51.

[13] Huashan Pan, Xiulin Li, and Zhiqiang Huang, "A mandarin prosodic boundary prediction model based on multi-task learning.," in *Interspeech*, 2019, pp. 4485–4488.

[14] Yang Zhang, Liqun Deng, and Yasheng Wang, "Unified mandarin tts front-end based on distilled bert model," *arXiv preprint arXiv:2012.15404*, 2020.

[15] Chunhui Lu, Pengyuan Zhang, and Yonghong Yan, "Self-attention based prosodic boundary prediction for chinese speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7035–7039.

[16] Junjie Pan, Xiang Yin, Zhiling Zhang, Shichao Liu, Yang Zhang, Zejun Ma, and Yuxuan Wang, "A unified sequence-to-sequence front-end model for mandarin text-to-speech synthesis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6689–6693.

[17] Zhengchen Zhang, Fuxiang Wu, Chenyu Yang, Minghui Dong, and Fugen Zhou, "Mandarin prosodic phrase prediction based on syntactic trees.," in *SSW*, 2016, pp. 160–165.

[18] Danqi Chen and Christopher D Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 740–750.

[19] Mitchell Stern, Jacob Andreas, and Dan Klein, "A minimal span-based neural constituency parser," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 818–827.

[20] Nikita Kitaev and Dan Klein, "Constituency parsing with a self-attentive encoder," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2676–2686.

[21] David Gaddy, Mitchell Stern, and Dan Klein, "What's going on in neural constituency parsers? an analysis," *arXiv preprint arXiv:1804.07853*, 2018.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[24] Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin, "Learning structured prediction models: A large margin approach," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 896–903.

[25] Databaker technology Inc. (Beijing), "Open source Chinese female voice database," 2019.