SELF-ATTENTION BASED PROSODIC BOUNDARY PREDICTION FOR CHINESE SPEECH SYNTHESIS

Chunhui Lu[1,2], *Pengyuan Zhang**[1,2], *Yonghong Yan*[1,2,3]

[1] Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics,China
[2] University of Chinese Academy of Sciences, China
[3]Xinjiang Key Laboratory of Minority Speech and Language Information Processing,
Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, China

ABSTRACT

Predicting prosodic boundaries from input text plays an important role in Chinese text-to-speech (TTS) system, which directly influences the naturalness and intelligibility of synthesized speech. In this paper, we propose to combine self-attention with multitask learning for prosodic boundary prediction. Self-attention is used to capture the dependency between two arbitrary characters in the input sentence, while multitask learning models the relationships between prosodic boundaries and lexicon words by setting word segmentation as an auxiliary task. The proposed method can generate prosodic boundary labels directly from Chinese characters and achieve the whole process end-to-end. Experimental results show the effectiveness of our proposed model and prove that the performance can be further improved by pretraining the model with extra word segmentation data.

Index Terms— prosodic boundary prediction, self-attention, multitask learning, speech synthesis

1. INTRODUCTION

Predicting prosodic structure from text is an essential step for statistical parametric speech synthesis (SPSS), whose results combined with other linguistic information are further utilized to predict acoustic parameters such as duration, pause, pitch and spectrum. This indicates that the accuracy of prosodic structure prediction largely determines the naturalness and even the intelligibility of synthesized speech.

In Chinese text-to-speech (TTS) systems, a hierarchical prosodic structure, including prosodic word (PW), prosodic phrase (PPH) and intonational phrase (IPH), is widely employed to distinguish different levels of pauses between words [1]. Predicting these prosodic structure is to identify whether each word boundary in a sentence is a prosodic boundary and decide its prosodic hierarchy. To achieve this, many statistical machine learning methods have been investigated in earlier times, including classification and regression tree (CART) [2], hidden Markov model (HMM) [3], maximum entropy model (ME) [4], and conditional random fields (CRF) [5, 6]. Among these methods, CRF achieved the best reported performance.

More recently, deep recurrent neutral network (RNN) based architecture along with embedding features have been investigated [7–11]. These works have proved that using bidirectional long short-term memory (BLSTM) recurrent networks achieves superior performance over the CRF based method as it captures long-range context information. Moreover, replacing traditional linguistic features with embedding features learned from raw text can further enhance the performance. Despite its successes, these RNN based models have limitations. For one thing, RNNs compute the hidden state of each time-step sequentially, which precludes its parallelization and leads to the $O(n)$ path length between two arbitrary characters. Another one is concerned with input features. Using word embeddings as input means that automatic word segmentation should be performed before prosodic boundary prediction, so the word segmentation errors will be propagated and accumulated.

In this paper, we propose to use self-attention to replace RNNs for the task of prosodic boundary prediction. In contrast to RNNs, self-attention is highly parallel and connects two arbitrary characters with each other directly regardless of their distance. Along with self-attention, multitask learning framework is implemented to capture the correlations between prosodic units and lexicon words by setting word segmentation as a secondary task. Our model is end-to-end and predicts prosodic boundaries directly from Chinese characters to avoid the negative effects of word segmentation errors. Moreover, we investigate the effectiveness of using extra word segmentation data to pretrain the model as we believe improving word segmentation accuracy is beneficial for prosodic boundary prediction. The rest of this paper is organized as follows. Section 2 describes our proposed method.

*Pengyuan Zhang is the corresponding author.

This work is partially supported by the National Natural Science Foundation of China (Nos. 11590773, 11590770), the Pre-research Project for Equipment of General Information System (No.JZX2017-0994/Y306).

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

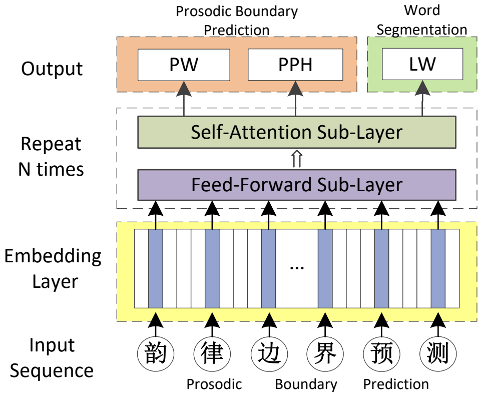

Fig. 1. The architecture of proposed model

Section 3 discusses the experiments, followed by the conclusions in section 4.

2. PROPOSED METHOD

As illustrated in Fig.1, our proposed model is composed of three components. The embedding layer maps each Chinese character of input sentence into corresponding embeddings as basic feature vectors and these features are updated during training. The main component of our model consists of N identical layers. Each layer contains a feed-forward sub-layer followed by a self-attention sub-layer. These layers receive a sequence of character embeddings as input, and pass the output to different task related softmax classification layers to generate final prediction results.

2.1. Self-attention model

Inspired by the successful use of self-attention in many natural language processing (NLP) tasks [12–14], we investigate it in prosodic boundary prediction which can be seen as a sequence labeling problem.

We adopt the multi-head self-attention [12], whose computation graph is depicted in Fig.2. It consists of h attention heads, each of which learns a distinct attention function from different representation subspaces to attend at different positions in the sequence.

Specifically, given an input matrix $\mathbf{X} \in \mathbb{R}^{t\times d}$, the multi-head attention mechanism first maps this matrix to h different queries, keys and values matrices by using linear projection. Formally, for the i-th head, we denote the queries, keys and values by $\mathbf{Q} \in \mathbb{R}^{t\times d/h}$, $\mathbf{K} \in \mathbb{R}^{t\times d/h}$ and $\mathbf{V} \in \mathbb{R}^{t\times d/h}$ respectively. Then the scaled dot-product attention [12] is used to compute the context vectors as:

$$\mathbf{M}_i = \text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V} \quad (1)$$

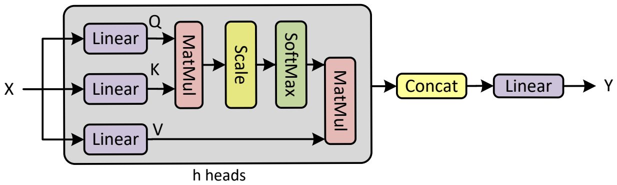

Fig. 2. The computation graph of multi-head self-attention mechanism

Finally, the output $\mathbf{Y}$ is computed as below:

$$\mathbf{M} = \text{Concat}(\mathbf{M}_1, ..., \mathbf{M}_h) \quad (2)$$

$$\mathbf{Y} = \mathbf{M}\mathbf{W} \quad (3)$$

where $\mathbf{M} \in \mathbb{R}^{t\times d}$ and $\mathbf{W} \in \mathbb{R}^{d\times d}$.

Using the weighted sum to produce output vectors limits the representational power of our model. To further improve the modeling power of the network, we add a feed-forward sub-layer before self-attention sub-layer. It consists of two linear projections with a ReLU activation [15] in the middle:

$$\text{FFN}(\mathbf{X}) = \text{ReLU}(\mathbf{X}\mathbf{W}_1 + b_1)\mathbf{W}_2 + b_2 \quad (4)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times d_f}$, $\mathbf{W}_2 \in \mathbb{R}^{d_f\times d}$ and we set $d_f = 400$ in all our experiments.

To ease the training process, we employ a residual connection [16] around each of the two sub-layers mentioned above. The output $\mathbf{Y}$ of each sub-layer is computed by the following equation:

$$\mathbf{Y} = \mathbf{X} + \text{Sub-Layer}(\mathbf{X}) \quad (5)$$

After the residual connection, layer normalization [17] is applied to stabilize the activations of the network.

Since self-attention cannot distinguish different positions of the sequence, we add positional encodings to the input embeddings. In this work, we use sine and cosine functions of different frequencies [12] to encode the position:

$$\text{PE}(t, 2i) = \sin(t/10000^{2i/d}) \quad (6)$$

$$\text{PE}(t, 2i+1) = \cos(t/10000^{2i/d}) \quad (7)$$

Where t is the position and i is the dimension. The positional encodings have the same dimension as the input character embeddings, so that they can be added directly without introducing additional parameters.

2.2. Multitask learning

For Chinese speech synthesis, IPH is often separated by punctuations and easy to be recognized, so we focus on PW and PPH prediction in this work. Unlike previous works that predicted different levels of prosodic boundary by different models. We treat PW and PPH prediction as two related tasks and predict them in one model by using multitask learning framework.

It's hard to model prosodic boundaries directly from Chinese characters because the lack of word level representations. We add word segmentation as an auxiliary task so that prosodic boundary prediction task can acquire word boundaries information by sharing high-level features from the shared hidden layers. Same to the conventional multitask learning network, our model can be trained by minimizing the global loss computed from the sum of costs from three tasks.

3. EXPERIMENTS

3.1. Datasets

We evaluate the proposed method on a neutral Chinese speech synthesis corpus recorded by a professional female speaker which contains 9,000 sentences. Prosodic boundaries of all the sentences are manually labeled by an annotator through listening to the utterances and reading the transcriptions. 90% of the utterances are used for training, 5% are used for validation, and the remaining 5% are used for testing. A large set of raw texts are collected to pretrain embeddings using word2vec [18]. The word embeddings dimension is set to 300 while the character one is set to 100. Besides, the MSR training set of SIGHAN Bakeoff 2005 [19] which contains more than 80,000 sentences is used as extra corpus for model pretraining in system SA-MTL-P.

3.2. System built

The output for PW and PPH have three dimensions, each corresponds to one of three boundary tags: B, NB and O. B for boundary, NB for non-boundary, and O for other symbols. For word segmentation, we use 5 tags to represent different character positions in the word: S for single word, B for the first character in a word, M for middle characters, E for the last one and O for other symbols.

To investigate the relationships between prosodic boundaries and lexicon words, it's worth comparing different systems with or without word level information as input. Based on these, the following systems are built with the help of TensorFlow [20].

1. **BLSTM-CRF**: BLSTM-CRF based model which uses word embeddings as input and achieves state-of-the-art results in [11] is used as our baseline. The input words are manually labeled.

2. **SA-Char**: The model based on self-attention with Chinese characters as the only input.

3. **SA-Char-W**: Adding a five-dimension one-hot vector that represents the position of the character in its corresponding word to the input of system **SA-Char**.

4. **SA-MTL**: Adding word segmentation as an extra task to system **SA-Char**.

Table 1. Results for different systems

Systems	PW-F1	PPH-F1	W-ACC
BLSTM-CRF	92.99	81.69	-
SA-Char	89.30	80.81	-
SA-Char-W	**93.66**	83.39	-
SA-MTL	90.58	81.93	0.8556
SA-MTL-P	93.65	**83.89**	**0.9109**

5. **SA-MTL-P**: Using the extra word segmentation data to pretrain the system **SA-MTL**.

For self-attention related models, the number of hidden layers and heads h are both set to 4. To mitigate overfitting, dropout [21] layers are added before the residual connections, the attention softmax layer and the feed-forward ReLU hidden layer, and the keep probabilities are set to 0.8, 0.9, 0.9 respectively. Parameter optimization is performed using Adam [22] with default learning rate at 0.001.

3.3. Results

To evaluate our systems, we use F1 score for PW (PW-F1) and PPH (PPH-F1), and word segmentation accuracy (W-ACC) as our measurements. W-ACC is calculated as the number of characters with correct word position tag divided by the number of all the characters. Experimental results are shown in Table 1.

Comparing system SA-Char to other three self-attention based systems, it's obvious that introducing word level information is effective and necessary for prosodic boundaries modeling. By using multitask learning framework, system SA-MTL captures the correlations between prosodic boundaries and lexicon words to some degree and consequently improves the results on system SA-Char. Using extra word segmentation data to pretrain the model greatly improves the performance on system SA-MTL and outperforms the baseline system BLSTM-CRF on both PW and PPH boundaries. For PW, SA-MTL-P achieves comparable result to system SA-Char-W which uses the "golden" character position tags during training and evaluation and is supposed to be the upper bound. It further reveals that PW boundaries and lexicon words are highly relevant and improving W-ACC is crucial to the performance of PW. In terms of PPH, SA-MTL-P even achieves better result than SA-Char-W, which might because by the utilization of large-scale data, the network better models the long-distance dependencies in the text which are more important for PPH.

3.4. Analysis

As self-attention is first used to prosodic boundary prediction, we analyze the main factors that influence the results on the system SA-MTL-P. Experimental results are shown in Table 2. Rows 1-4 show the effects of different number of lay-

Table 2. Detailed results on system SA-MTL-P

	FFN	PE	Depth	PW-F1	PPH-F1	W-ACC
1	√	√	4	93.65	**83.89**	**0.9109**
2	√	√	2	92.47	83.34	0.9014
3	√	√	3	93.06	83.46	0.9071
4	√	√	5	**93.91**	83.47	0.9075
5	×	√	4	91.92	82.20	0.9031
6	√	×	4	75.72	50.33	0.7951

Table 3. Comparision between argmax decoding and constrained decoding

Decoding	PW-F1	PPH-F1	W-ACC
Argmax Decoding	**93.65**	**83.89**	0.9109
Constrained Decoding	93.53	83.22	0.9109

ers. Increasing the model depth gradually improves the performance, but there are no longer consistent improvements on both PW and PPH when the depth coming to 5. So we believe 4 layers are sufficient for prosodic boundary modeling. Row 5 show the results of 4 layered model without feed-forward sub-layers. Its performance is even not as good as the performance of 2 layered model with FFN . It indicates that the feed-forward sub-layers are the essential components to enhance the model expressive ability. Furthermore, comparing rows 1 and 6 we can draw the conclusion that positional encodings are indispensable for our model.

Previous work [11] showed that adding a CRF layer at the output of the model allows the network to measure the probability of transition between labels and to generate the most optimal tag sequences. So it's necessary to compare our decoding method which uses the tag with highest probability for each character as the final output with the CRF constrained decoding method. Experimental results are listed in Table 3, from which we can observe a slightly performance drop when using constrained decoding. It indicates that our self-attention based model is powerful enough to capture the transition relationship among labels.

To figure out what information the model attends to, we visualize the attention weights for different heads in layer 3 of 4, which are showed in Fig.3. We observe that each head has different attention weights given a certain input sentence, which shows that multi-head attention can learn information from different representation subspaces. For each character in the sentence, the heads mainly focus on the two characters adjacent to it and this range is sufficient to decide PW boundaries at most times as PW usually consists of two or three characters. Moreover, the first character also attends to the last one in head 1 and 4, which proves that self-attention mechanism is able to model the long-distance dependencies of two arbitrary characters.

To further evaluate the prosody modeling performance of system BLSTM-CRF and SA-MTL-P, we conduct an AB preference test of the synthesized speech. 20 sentences are randomly selected from the test set with different prosodic boundary prediction results and corresponding speech are generated using our LSTM based TTS engine[1]. A group of 10 subjects were asked to give their preference in terms of the naturalness for each speech pair. The percentage preference is shown in Fig.4. We can clearly see that the proposed method (system SA-MTL-P) is significantly better than the baseline system BLSTM-CRF in terms of the naturalness of synthesized speech ($p < 0.0001$).

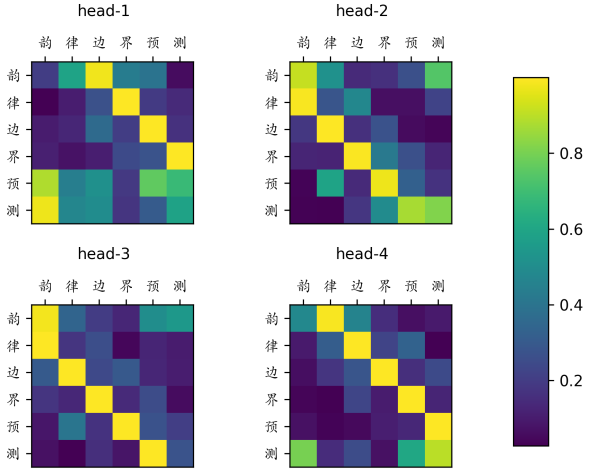

Fig. 3. The visualization of attention weights in layer 3 of 4. The input is prosodic boundary prediction (in Chinese)

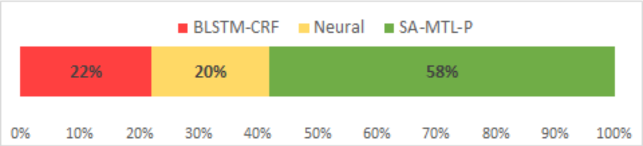

Fig. 4. The preference of AB test.

4. CONCLUSIONS

In this paper, we present a self-attention based multitask learning architecture which achieves prosodic boundary prediction and word segmentation at the same time. Self-attention is used to capture the contextual dependencies of the input sentence. While multitask learning further helps to model correlations between prosodic structure and lexicon words. By using extra word segmentation data to pretrain the model, the performance can be further improved and outperforms previous state-of-the-art method (BLSTM-CRF). In the future, we will investigate the possibility of using self-attention to other text analysis tasks of speech synthesis.

[1] https://chunhui-lu.github.io/ICASSP2019/index.html

5. REFERENCES

[1] M. Chu and Y. Qian, "Locating boundaries for prosodic constituents in unrestricted mandarin texts," *Computational Linguistics and Chinese Language Processing*, vol. 6, no. 1, pp. 61–82, 2001.

[2] M. Wang and J. Hirschberg, "predicting intonational boundaries automatically from text: the atis domain," in *Proceedings of DARPA Speech and Natural Language Workshop,*, 1991, pp. 378–383.

[3] X. Nie and Z. Wang, "Automatic phrase break prediction in chinese sentences," *Journal of Chinese information Processing*, pp. 39–44, 2003.

[4] J. Li, G. Hu, and R. Wang, "Chinese prosody phrase break prediction based on maximum entropy model," in *Proceedings of INTERSPEECH*, 2004, pp. 729–732.

[5] G. Levow, "Automatic prosodic labeling with conditional random fields and rich acoustic features," in *Proceedings of IJCNLP*, 2008, pp. 217–224.

[6] Y. Qian, Z. Wu, X. Ma, and F. Soong, "Automatic prosody prediction and detection with conditional random field models," in *Proceedings of ISCSLP*, 2010, pp. 135–138.

[7] C. Ding, L. Xie, J. Yan, W. Zhang, and Y. Liu, "Automatic prosody prediction for chinese speech synthesis using blstm-rnn and embedding features," in *Proceedings of ASRU*, 2015, pp. 98–102.

[8] A. Rendel, R. Fernandez, R. Hoory, and B. Ramabhadran, "Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end,," in *Proceedings of ICASSP*, 2016, pp. 5655–5659.

[9] Y. Zheng, Y. Li, Z. Wen, X. Ding, and J. Tao, "Improving prosodic boundaries prediction for mandarin speech synthesis by using enhanced embedding feature and model fusion approach," in *Proceedings of INTERSPEECH*, 2016, pp. 3201–3205.

[10] Y. Huang, Z. Wu, R. Li, H. Meng, and L. Cai, "Multitask learning for prosodic structure generation using blstm rnn with structured output layer," in *Proceedings of INTERSPEECH*, 2017, pp. 779–783.

[11] Y. Zheng, J. Tao, Z. Wen, and Y. Li, "Blstm-crf based end-to-end prosodic boundary prediction with context sensitive embeddings in a text-to-speech front-end," in *Proceedings of INTERSPEECH*, 2018, pp. 47–51.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and Polosukhin, "Attention is all you need," in *Proceedings of NIPS*, 2017, pp. 5998–6008.

[13] Z. Tan, M. Wang, J. Xie, Y. Chen, and X. Shi, "Deep semantic role labeling with self-attention," in *Proceedings of AAAI*, 2018.

[14] E. Strubell, P. Verga, D. Andor, D. Weiss, and A. Mccallum, "Linguistically-informed self-attention for semantic role labeling," in *Proceedings of EMNLP*, 2018.

[15] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of ICML10*, 2010, pp. 807–814.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of CVPR*, 2016, pp. 770–778.

[17] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of NIPS*, 2013, pp. 3111–3119.

[19] T. Emerson, "The second international chinese word segmentation bakeoff," in *the fourth SIGHAN workshop on Chinese language Processing*, 2005.

[20] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, and et al, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *OSDI*, vol. 16, pp. 265–283, 2016.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of ICLR*, 2015.