Prosodic Structure Prediction using Deep Self-attention Neural Network

Yao Du*, Zhiyong Wu*, Shiyin Kang†, Dan Su†, Dong Yu†, Helen Meng‡
* Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
E-mail: mathewdu@gmail.com, zywu@se.cuhk.edu.hk
† Tencent AI Lab, Tencent, Shenzhen, China
E-mail: {shiyinkang, dansu, dyu}@tencent.com
‡ Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
E-mail: hmmeng@se.cuhk.edu.hk

Abstract—**Prosodic structure prediction is a key part of the text analysis front-end of the text-to-speech (TTS) system. It predicts prosodic boundary tags given the input text context, which is essential to the naturalness of synthesized speech. Conventional methods such as conditional random fields (CRF) and recurrent neural network (RNN) have been successfully applied to this task. However, the lack of modeling temporal dependencies at different scopes (the short-term dependency as well as the long-span dependency across the entire sentence) limits their performance. In this paper, we propose a self-attention network with semantic features extracted by a pre-trained bidirectional encoder representations from Transformers (BERT) model to predict the prosodic structure. Experimental results show that the proposed approach outperforms the strong baseline CRF model with an absolute improvement of 3.4% in total accuracy.**

I. Introduction

For a typical Chinese TTS system as shown in Fig.1. Chinese word segmentation and POS tagging are first undertaken to tokenize the input Chinese sentence into lexical words with part-of-speech (POS) information that are further fed into prosodic structure prediction module to predict prosodic boundaries including prosodic word (PW), prosodic phrase (PPH), intonational phrase (IPH). Fig.2 depicts an example of prosodic structure hierarchy commonly used in Chinese. Grapheme-to-phoneme conversion is performed after the prosodic structure prediction to determine the pronunciation of each word. Those front-end processing results are used to predict the acoustic parameters which can be fed to vocoder to synthesis the speech signals. Prosodic structure prediction plays an important role in TTS system. It affects the naturalness of the synthesized speech.

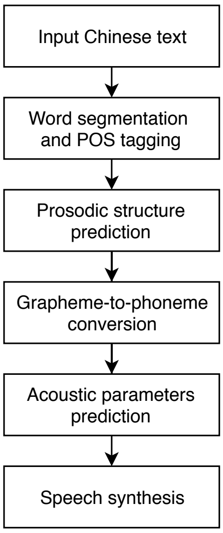
Input Chinese text
Word segmentation and POS tagging
Prosodic structure prediction
Grapheme-to-phoneme conversion
Acoustic parameters prediction
Speech synthesis

Fig. 1. A typical Chinese TTS system.

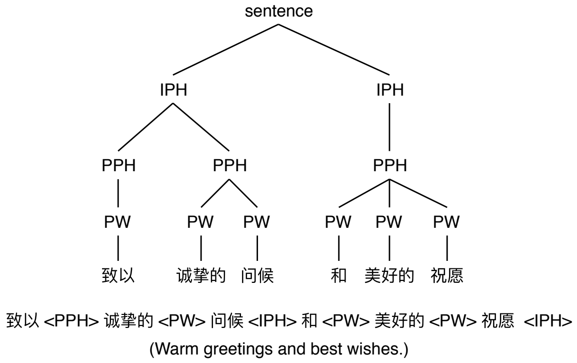
sentence
IPH
IPH
PPH
PPH
PPH
PW
PW
PW
PW
PW
PW
致以
诚挚的
问候
和
美好的
祝愿
致以 <PPH> 诚挚的 <PW> 问候 <IPH> 和 <PW> 美好的 <PW> 祝愿 <IPH>
(Warm greetings and best wishes.)

Fig. 2. An example of Chinese prosodic structure hierarchy.

A variety of methods have been proposed to predict prosodic boundaries. In the early time, simple rule-based methods are proposed. With the development of machine learning technologies, many statistical methods have been applied to predict prosodic boundaries, including classification and regression tree (CART) [1], hidden Markov model (HMM) [2], maximum entropy model (ME) [3] and conditional random fields (CRF) [4]. Among these models, CRF has been reported to achieve the best performance on the task of prosodic structure generation [4]. Recently, recurrent neural network (RNN) has been successfully applied to prosodic structure prediction [5]. However, there still remains a major challenge for both CRF and RNN to address the interaction between short-term and long-term contextual information. Furthermore, the CRF needs hand-craft indicator features based on expert knowledge.

In this paper, we apply self-attention mechanism to the task of prosodic structure prediction. It can directly make the connections between the input features of two arbitrary words in a sentence, regardless of their distance. Besides, it can capture the structure information of the sentence due to the self-attention mechanism that computes a word representation based on the features of all the words in the sentence.

II. Features

In previous work, word embedding is used to predict prosodic boundaries through RNN [6]. We adopt pre-trained contextualized word embeddings obtained from bidirectional encoder representations from Transformers (BERT) [7] which has achieved state-of-the-art performance for a wide range of natural language processing tasks. Previous work has reported the correlation between the part-of-speech (POS) tag and the prosodic boundary type [2] [8]. Hence POS information is incorporated into the input features for prosodic structure prediction in our work. Obviously, punctuation symbols always indicate prosodic boundaries. Besides, prosodic boundary is affected by the number of syllables in a lexical word (word length). A long lexical word often corresponds to a single prosodic word [9].

Based on the works mentioned above, the input word-level features are listed as followings:

- Word embedding
- POS
- Word length
- Punctuation symbol after the word

For convenience, we refer to the combination of POS, word length and punctuation after the word as rich word-level textual features hereafter.

III. Proposed Method

A. Self-Attention

Self-attention has been successfully used in many tasks including natural language inference [10], neural machine translation [11] and sequence labeling [12]. Our work follows these applications and applies self-attention mechanism to the task of prosodic structure prediction.

Attention mechanism can be described as computing the similarity of a query and each key to get a set of weights which are used to obtain a weighted sum of corresponding values. The output weighted sum can be seen as a representation of that query. There are different functions to calculate the similarity. Additive attention and dot-product attention are commonly used functions. In this work, we adopt a scaled dot-product attention, a variant of dot-product attention [13], as illustrated in Fig.3.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d}})V \qquad (1)$$

where Q, K and V are attention queries, keys and values respectively. d is the dimension of the query.

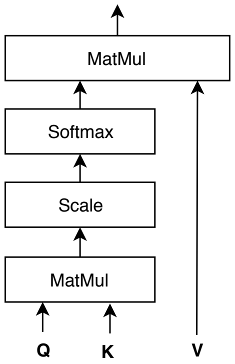

Fig. 3. The computation graph of scaled dot-product attention.

Self-attention is a attention mechanism [11] that only needs a single input sequence X to compute its representation, i.e. $Q = K = V = X$. Multi-head attention projects the queries, keys and values to different subspaces through linear transformation, and then performs the scaled dot-product attention at each subspace in parallel. Each output of dot-product attention is d_v-dimensional value. All the outputs are concatenated to form a $h \times d_v$ dimensional value and projected linearly to yield the output Y, as depicted in Fig.4.

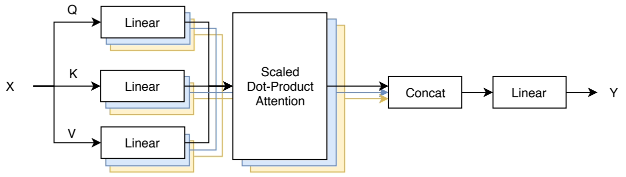

Fig. 4. Multi-head self-attention mechanism.

The mathematical representation of multi-head attention can be depicted as follows:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \qquad (2)$$

$$MultiHead(Q,K,V) = Concat(head_1, ..., head_h)W \qquad (3)$$

where $W_i^Q \in \mathbb{R}^{d \times d_k}, W_i^K \in \mathbb{R}^{d \times d_k}, W_i^V \in \mathbb{R}^{d \times d_v}$ are the transformations parameters of queries, keys and values respectively for i-th head. $W \in \mathbb{R}^{hd_v \times d}$ is the transformation matrix for the last linear transition. In our work, we set the head number as $h = 8$. For each head, we use $d = 256$, $d_k = d_v = d/h = 64$.

B. Nonlinear sub-layer

1) Feed-forward sub-layer: Feed-forward network (FFN) sub-layer can be used along with self-attention sub-layer to transform the input nonlinearly from the bottom layers. It consists of two linear transformations with a ReLU activation [14] in the middle.

$$FFN(X) = ReLU(XW_1)W_2 \qquad (4)$$

where $W_1 \in \mathbb{R}^{d \times d}$, $W_2 \in \mathbb{R}^{d \times d}$ are the parameters to learn.

2) RNN sub-layer: Prosodic structure prediction can be viewed as a sequence prediction task. RNN is appropriate for sequence modeling. Long Short-Term Memory (LSTM) [15] is a RNN unit with introduced gate mechanism. Gated recurrent unit (GRU) [16], a variant of LSTM, has simpler structure. They are widely applied to sequence modeling tasks. We explore three kinds of RNN sub-layers, including unidirectional GRU sub-layer, bidirectional LSTM (BLSTM) sub-layer and bidirectional GRU (BGRU) sub-layer in our work.

C. Residual connection

Residual connection has been reported an effective way to train very deep neural network. We apply residual connection to each sub-layer as follows:

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

where X, Y is the input and output of each sub-layer respectively. It is implemented by a shortcut connection and element-wise addition. Layer normalization [17] is also performed after the shortcut connection to obtain more stable hidden-to-hidden dynamics in deep neural network.

D. Position encoding

Although the self-attention mechanism can learn the dependencies between the words in a sentence at any distance, it can't exploit the relative position information of the words. There are many approaches to encode the position information. Timing signal [11] is a convenient way that can be generated through the following formulation:

$$timing(t, 2i) = sin(t/10000^{2i/d}) \tag{6a}$$

$$timing(t, 2i+1) = cos(t/10000^{2i/d}) \tag{6b}$$

where t is the time-step index, $2i$ and $2i+1$ is the channel index. Each dimension of positional encoding corresponds to a sinusoid.

E. Model architecture

As Fig.5 shows, the input to our model is the word-level features sequence. The word embedding e_i added with positional encoding is concatenated with rich word-level textual features r_i, at i-th time step, to form the features fed into front dense layer. That dense layer outputs the fused features which are further fed into N identical blocks to learn the deep representation and dependencies between different time-steps. Each block consists of a nonlinear sub-layer and a self-attention sub-layer. The last softmax layer outputs the probabilities of all prosodic boundary types at each time step.

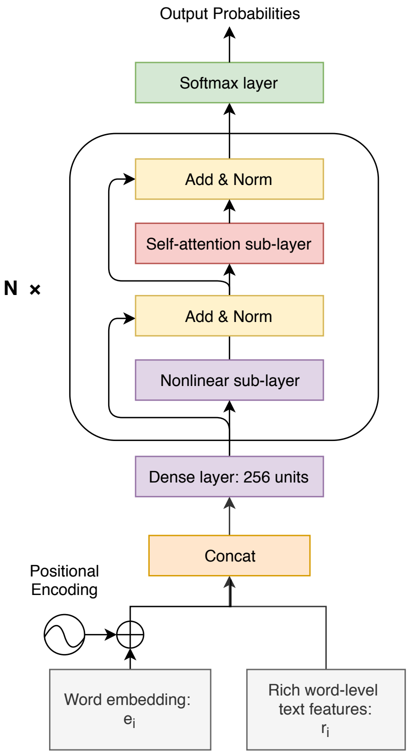

Fig. 5. The architecture of deep self-attention neural network.

IV. Experiments

A. Experimental setup

1) Dataset and preprocessing: Our corpus contains 41,483 sentences with prosodic structure boundaries labeled by a professional annotator. We randomly select 37,483 sentences for training, another 2,000 sentences for validation, the rest 2,000 sentences for test.

Word segmentation and POS tagging are carried out by a front-end preprocessing tool. We can easily get the punctuation symbol after each lexical word and the number of syllables within each lexical word by means of text analysis.

2) Settings and regularization: The size of each hidden layer is set to 256. The number of heads h is set to 8. Adam is adopted as the optimizer of our model. The initial learning rate is set to 0.0001. The training batch size is set to 256. Dropout [18] layer with keep probability of 0.8 are applied before each sub-layer. Label smoothing is a regularization mechanism that makes the model learn to be more unsure but improves performance [19]. In our work, it is applied with a smoothing value of 0.1 during training.

B. Evaluation Metrics

In our work, we mainly use total accuracy (T-ACC) to compare the performance of different models, F1 scores of PW, PPH, IPH are also recorded to ensure the models performance on these measurements. T-ACC can be calculated as:

$$T\text{-}ACC = \frac{N_{correctly\_predicted\_samples}}{N_{total\_samples}} \tag{7}$$

where $N_{correctly\_predicted\_samples}$ is the number of correctly predicted samples, $N_{total\_samples}$ is the number of total samples. The F1 measure considers both precision and recall. For example, the F1 score for predicting intonational phrase boundaries can be calculated as followings:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

$$Precision = \frac{N_{correctly\_predicted\_IPH}}{N_{predicted\_IPH}} \tag{9}$$

$$Recall = \frac{N_{correctly\_predicted\_IPH}}{N_{ground\_truth\_IPH}} \tag{10}$$

where $N_{correctly\_predicted\_IPH}$ is the number of correctly predicted IPH samples, $N_{predicted\_IPH}$ is the number of predicted IPH samples, $N_{ground\_true\_IPH}$ is the number of ground truth IPH samples.

V. Experimental Results and Analysis

A. *Results of different model configurations*

In this section, we discuss the main factors that influence our model performance.

Nonlinear sub-layer After we set the model depth N to 6, we have tried different nonlinear sub-layers, including FFN sub-layer, GRU sub-layer, BLSTM sub-layer and BGRU sub-layer. The results are presented in Table I, from which we can see that the choice of BGRU sub-layer achieves best performance. The model with FFN sub-layer shows less total accuracy, but it trains faster than the model with recurrent sub-layers.

TABLE I
The results of different nonlinear sub-layer.

Nonlinear sub-layer	FFN	GRU	BLSTM	BGRU
T-ACC	0.8089	0.8158	0.8185	0.8241

Model Depth In deep self-attention network, we have stacked N identical blocks (each block contains a nonlinear sub-layer and a self-attention sub-layer). After we set the nonlinear sub-layer to BGRU, we have tried different number of identical blocks in order to evaluate how the depth of model affects the performance. As Table II shows, our model achieves best performance with N=6.

TABLE II
The results of different settings of model depth.

N Blocks	2	3	4	5	6	7
T-ACC	0.8188	0.8205	0.8237	0.8221	0.8241	0.8208

B. *Confusion matrix for prediction results*

The prosodic structure tags, which are labeled by a professional annotator, serve as ground truth. The prediction results of our model can be evaluated by confusion matrix. The matrices for the baseline CRF and our model are presented in Table III and Table IV. Compared with CRF, the correctly predicted PW, PPH samples of our model are significantly increased. Our model are likely to confuse PPH with PW. One possible reason is that PPH has the least samples among those four prosodic boundary types in our data set, which makes it more subtle than other boundary types. In total, those results indicate that the deep self-attention neural network provides improvement over the baseline CRF.

TABLE III
Confusion matrix for CRF.

Predicted Actual	NB	PW	PPH	IPH
NB	**4142** **(83.39%)**	702 (14.13%)	123 (2.48%)	0 (0.00%)
PW	526 (7.41%)	**5714** **(80.49%)**	843 (11.87%)	16 (0.23%)
PPH	122 (4.07%)	1240 (41.32%)	**1553** **(51.75%)**	86 (2.87%)
IPH	17 (0.46%)	111 (3.02%)	147 (4.00%)	**3397** **(92.51%)**

TABLE IV
Confusion matrix for deep self-attention network.

Predicted Actual	NB	PW	PPH	IPH
NB	**4333** **(87.24%)**	590 (11.88%)	43 (0.87%)	1 (0.02%)
PW	406 (5.72%)	**6035** **(85.01%)**	649 (9.14%)	9 (0.13%)
PPH	56 (1.87%)	1193 (39.75%)	**1679** **(55.95%)**	73 (2.43%)
IPH	8 (0.22%)	85 (2.31%)	182 (4.96%)	**3397** **(92.51%)**

C. *Comparison with related models*

We have implemented the following models for comparison.

CRF Conventional CRF model with lexical words, POS tagging labels, word length and post-word punctuations as input.

BLSTM-EMB The model consists of a BLSTM layer and a output layer. It predicts the prosodic structure boundary type using word embedding. The word embedding is same as that of our model.

BGRU-RICH The model architecture consists of a dense layer followed by six BGRU-RNN layers, the last softmax layer outputs the probabilities of all prosodic boundary types. The input features are same as our model's input features as described in Section II.

SELF-ATT Our proposed approach for prosodic structure prediction as illustrated in Fig.5. The nonlinear sub-layer is set to BGRU and the number of identical blocks is set to 6.

TABLE V
Experimental results of related models.

Model	PW F1	PPH F1	IPH F1	T-ACC
CRF	0.7687	0.5481	0.9474	0.7901
BLSTM-EMB	0.7751	0.5254	0.8477	0.7767
BGRU-RICH	0.7988	0.5984	0.9496	0.8193
SELF-ATT	0.8046	0.6046	0.9499	0.8241

Experimental results are presented in Table V. Comparing CRF with SELF-ATT, the latter achieves better performance with an absolute improvement of 3.4% in total accuracy. Compared with BLSTM-EMB, our model achieves much better performance on all measurements. It also shows that word embedding is not a sufficient feature for the task of three-level prosodic structures prediction. With the self-attention mechanism, our model provides a minor improvement by 0.48% compared with BGRU-RICH, indicating self-attention sub-layers along with BGRU sub-layers help to learn the latent dependencies of words in the sentence, resulting in best performance compared with related models.

VI. Conclusions

In this paper, we have applied self-attention mechanism to the task of prosodic structure prediction. Experimental results show that the proposed approach outperforms the comparison systems. The application of self-attention mechanism to prosodic structure prediction task can help learn the latent dependencies information between the words in the sentence. In the future, we will investigate a hierarchical neural network that can help to learn the hierarchical relations between prosodic structures.

VII. Acknowledgements

This work is supported by joint research fund of National Natural Science Foundation of China - Research Grant Council of Hong Kong (NSFC-RGC) (61531166002, N CUHK404/15), National Natural Science Foundation of China (61433018, 61375027) and National Social Science Foundation of China (13&ZD189). We would also like to thank Tencent AI Lab Rhino-Bird Focused Research Program (JR201803, JR201942) and Tsinghua University - Tencent Joint Laboratory for the support.

References

[1] M. Q. Wang and J. Hirschberg, "Predicting intonational boundaries automatically from text: the atis domain," in *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991.

[2] P. Taylor and A. W. Black, "Assigning phrase breaks from part-of-speech sequences," *Computer Speech & Language*, vol. 12, no. 2, pp. 99–117, 1998.

[3] J.-F. Li, G.-p. Hu, and R. Wang, "Chinese prosody phrase break prediction based on maximum entropy model," in *Eighth International Conference on Spoken Language Processing*, 2004.

[4] Y. Qian, Z. Wu, X. Ma, and F. 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.

[5] A. Vadapalli and S. V. Gangashetty, "An investigation of recurrent neural network architectures using word embeddings for phrase break prediction." in *Interspeech*, 2016, pp. 2308–2312.

[6] G. Mesnil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding." in *Interspeech*, 2013, pp. 3771–3775.

[7] 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.

[8] E. Sanders and P. A. Taylor, "Using statistical models to predict phrase boundaries for speech synthesis." 1995.

[9] J. Vaissière, "Rhythm, accentuation and final lengthening in french," in *Music, language, speech and brain*. Springer, 1991, pp. 108–120.

[10] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," *arXiv preprint arXiv:1606.01933*, 2016.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[12] Z. Tan, M. Wang, J. Xie, Y. Chen, and X. Shi, "Deep semantic role labeling with self-attention," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[14] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

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

[18] 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.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.