AUTOMATIC PROSODY PREDICTION FOR CHINESE SPEECH SYNTHESIS USING BLSTM-RNN AND EMBEDDING FEATURES

Chuang Ding[1], *Lei Xie*[1,2], *Jie Yan*[2], *Weini Zhang*[2], *Yang Liu*[2]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]School of Software and Microelectronics, Northwestern Polytechnical University, Xi'an, China
{cding, lxie, jyan, wnzhang, yangliu}@nwpu-aslp.org

ABSTRACT

Prosody affects the naturalness and intelligibility of speech. However, automatic prosody prediction from text for Chinese speech synthesis is still a great challenge and the traditional conditional random fields (CRF) based method always heavily relies on feature engineering. In this paper, we propose to use neural networks to predict prosodic boundary labels directly from Chinese characters without any feature engineering. Experimental results show that stacking feed-forward and bidirectional long short-term memory (BLSTM) recurrent network layers achieves superior performance over the CRF-based method. The embedding features learned from raw text further enhance the performance.

Index Terms— automatic prosody prediction, speech synthesis, neural network, BLSTM, embedding features

1. INTRODUCTION

Prosody refers to the rhythm, stress and intonation of speech, including variations in duration, loudness and pitch. It is well known that speech prosody plays an important perceptual role in human speech communication [1]. Specifically, perception of prosodic boundaries is essential for listeners. In Chinese speech synthesis systems, typical prosody boundary labels consist of prosodic word (PW), prosodic phrase (PPH) and intonational phrase (IPH), which construct a three-layer prosody structure tree [2], as shown in Fig. 1. The leaf nodes of tree structure are lexical words that can be derived from a lexical-based word segmentation module. Whether the prosody labels are properly predicted will directly affect the naturalness and intelligibility of the synthesized speech.

Previous studies have investigated a great number of features, their relevance to prosody generation in speech production and various prosodic modeling methods. Some syntactic cues like part-of-speech (POS), syllable identity, syllable stress and their contextual counterparts are commonly used for prosody boundary prediction [3, 4, 5]. Many statistical methods have been investigated to model speech prosody, including classification and regression tree [6], hidden Markov model [7], maximum entropy model [8] and conditional random fields (CRF) [9]. To our knowledge, the best reported results were achieved with CRF due to its ability of relaxing strong model independence assumption and solving the label bias problem [1, 10].

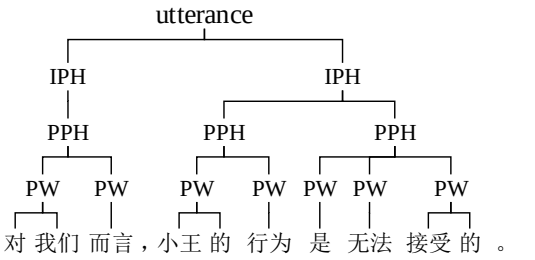

Fig. 1. Three-layer prosody structure tree in Chinese.

Despite years of research, it is still a great challenge to predict correct prosodic labels from unrestricted text for a text-to-speech (TTS) system. Obviously, there are two major drawbacks of the CRF-based prosody prediction in Chinese speech synthesis. First, it heavily relies on the performances of Chinese word segmentation (CWS) and POS tagging [11]. Second, the particle size and the inevitable segmentation errors in CWS have negative effects on the subsequent prosodic boundary prediction task. Moreover, the choice of effective features, from a broad set of feature templates, is critical to the success of such systems [12]. Much of the effort goes into feature engineering, which is notoriously labor-intensive, mainly based on the experience of an annotator.

Recently, deep neural networks (DNN) have been increasingly investigated in order to minimize the effort of feature engineering in sequential labeling tasks. Zheng et al. [12] applied neural networks to CWS and POS tagging and proposed a perceptron-style algorithm to speed up the training process with negligible loss in performance. Pei et al. [13] proposed a max-margin tensor neural network for CWS to model interactions between tags and context characters by exploiting tag embeddings and tensor-based transformation. These researches have proved that DNN is able to achieve similar or even superior performance over CRF-based method with minimal feature engineering in sequential labeling tasks. Therefore, it is promising to apply DNN architectures to automatic prosody prediction. However, we notice that the neural net-

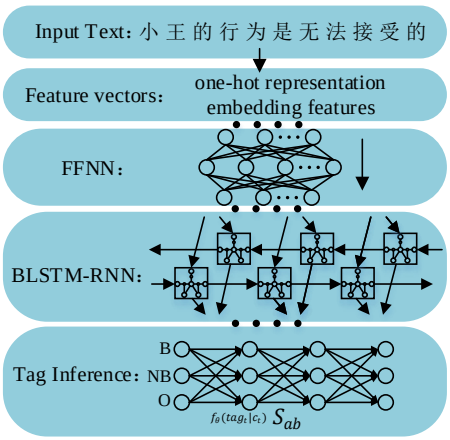

Fig. 2. The neural network architecture for prosodic boundary prediction. In tag inference, B, NB and O denote boundary, non-boundary and others (e.g., punctuation), respectively.

works used in previous researches are feed-forward structures that keep the assumption of sample independence and provide only limited context modeling ability by operating on a fixed-size window of input samples. Instead, bidirectional recurrent neural networks (BRNN) are able to incorporate contextual information from both past and future inputs [14]. Specifically, BRNN with long short-term memory (LSTM) cells, namely BLSTM-RNN, has become a popular model [15].

In this paper, we address the prosodic boundary prediction problem using neural networks. There are three main contributions. (1) We propose a neural network approach to predict prosody labels directly from Chinese characters without any feature engineering. (2) We show that superior performance is achieved by stacking feed-forward and bidirectional long short-term memory (BLSTM) recurrent layers. (3) We leverage a large raw text corpus to obtain useful character embedding features. Both objective and subjective evaluations show that the proposed architecture achieves superior performance over the CRF-based method and the embedding features further enhance the performance.

2. THE PROPOSED APPROACH

Just like CWS and POS tagging, automatic prosody prediction can be treated as a sequential labeling task that assigns boundary labels to characters of an input sentence. In order to make the prediction models less dependent on the feature engineering, we choose to use a variant of the neural network architecture proposed by [16] for probabilistic language model. This architecture was subsequently used for CWS and POS tagging [12]. As shown in Fig. 2, the architecture takes raw text as input and maps each Chinese character into a basic feature vector. The following layers are two types of neural networks, FFNN and BLSTM-RNN, used to discover multiple levels of feature representations from the basic feature vectors. The output layer is a graph over which tag inference is achieved by the Viterbi algorithm.

2.1. Feature Vectors

The characters fed into network are transformed into feature vectors by a mapping operation. Typically, a character dictionary D of size $|D|$ is extracted from the training set and unknown characters are mapped to a special symbol that is not used elsewhere. Each Chinese character can be typically represented by a one-hot vector, the size of which is $|D|$, and all dimensions are marked as 0 except the location of the character in D, which is marked as 1. However, the one-hot representation, with high dimensions, fails to model the semantic similarity between the ideographic characters. In contrast, the *distributed* representation or *embedding* feature, in form of a low dimensional continuous-valued vector learned using neural networks from raw text in a fully unsupervised way, is assumed to carry important syntactic and semantic information [18] [19]. Recently, Mansur et al. [20] have shown superior performance in Chinese word segmentation by the use of embedding features based on a neural language model [16]. Besides [16], Mikolov et al. [18] proposed a faster skip-gram model called *word2vec*[1]. As our preliminary experiments do not show much performance difference among various embedding features, we simply choose *word2vec* in this study because it can be trained much faster.

2.2. Network Structures and Training

Two types of neural networks are investigated in this paper: FFNN and BLSTM-RNN. FFNN, trained with a back-propagation learning algorithm [21], is widely used in many practical applications. In a typical FFNN, every unit in a layer is connected with all the units in the previous layer, which takes in the output of the previous layer and computes a new set of non-linear activations for next layer. However, the assumption of sample independence brings in only limited context modeling ability.

Researchers have proposed RNN to solve the limitation of FFNN. However, conventional RNN is only able to make use of previous context information. This is not accurate in modeling speech prosody that is highly related with both past and future contexts. Instead, bidirectional RNN can access both the preceding and succeeding input contexts with two separate hidden layers, which are then fed to the same output layer. The activation function $\mathcal{H}$ of RNN is usually a sigmoid or hyperbolic tangent function, which often causes the gradient vanishing problem that prevents RNN from modeling the long-span relations in sequence features. An LSTM architecture, which uses purpose-built memory cells to store information, can overcome this problem and model longer contexts. Fig. 3 illustrates a single LSTM memory cell. For LSTM, $\mathcal{H}$ is implemented by the following functions:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$

[1] https://code.google.com/p/word2vec/

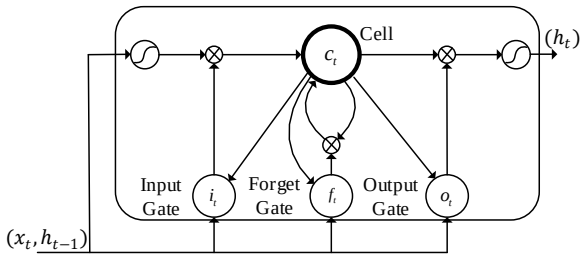

Fig. 3. Long short-term memory cell.

$$
\begin{aligned}
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
c_t &= f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
h_t &= o_t tanh(c_t)
\end{aligned}
$$

where $x = (x_1, x_2, ...x_t..., x_T)$ is the input feature sequence, σ is the logistic function, and i, f, o and c are the input gate, forget gate, output gate and cell memory, respectively. W is the weight matrix and the subscript indicates it is the matrix between two different gates.

BLSTM-RNN is a combination of LSTM and BRNN. Deep bidirectional LSTM-RNN can be established by stacking multiple BLSTM-RNN hidden layers on top of each other. The output sequence of one layer is used as the input sequence of the next layer. The hidden state sequences, h^n, consist of forward and backward sequences $\overrightarrow{h}^n$ and $\overleftarrow{h}^n$, iteratively computed from $n = 1$ to N and $t = 1$ to T as follows:

$$
\begin{aligned}
\overrightarrow{h}_t^n &= \mathcal{H}(W_{\overrightarrow{h}^{n-1}\overrightarrow{h}^n}\overrightarrow{h}_t^{n-1} + W_{\overrightarrow{h}^n\overrightarrow{h}^n}\overrightarrow{h}_{t-1}^n + b_{\overrightarrow{h}}^n), \\
\overleftarrow{h}_t^n &= \mathcal{H}(W_{\overleftarrow{h}^{n-1}\overleftarrow{h}^n}\overleftarrow{h}_t^{n-1} + W_{\overleftarrow{h}^n\overleftarrow{h}^n}\overleftarrow{h}_{t-1}^n + b_{\overleftarrow{h}}^n), \\
y_t &= W_{\overrightarrow{h}^N y}\overrightarrow{h}_t^N + W_{\overleftarrow{h}^N y}\overleftarrow{h}_t^N + b_y.
\end{aligned}
$$

where $y = (y_1, y_2, ...y_t..., y_T)$ is the output prosodic boundary sequence.

In our study, the feed-forward layers are trained with typical backpropagation (BP) algorithm and the back-propagation through time (BPTT) method is used for training of BLSTM layers. BPTT is applied to both forward and backward hidden nodes and back-propagates layer by layer. The weight gradients are computed over the entire utterance [22]. The neural networks can be trained effectively in a layer-wised training manner, which makes it convenient to stack different types of neural network layers on top of each other to form a deep architecture. The deep architecture is able to build up progressively higher level representations of the input data, which is a crucial factor of the recent success of hybrid systems [17].

2.3. Tag Inference

To model the tag dependency and infer the tag sequence globally, given a set of tags $G = \{B, NB, O\}$, a transition score S_{ab} is introduced for jumping from tag $a \in G$ to tag $b \in G$. For the input character sequence of a sentence $c_{[1:T]}$ with a tag sequence $tag_{[1:T]}$, a sentence-level score is then given by the sum of transition and network scores [12, 23]:

$$
l(c_{[1:T]}, tag_{[1:T]}, \theta) = \sum_{t=1}^{T}(S_{tag_{t-1}tag_t} + f_\theta(tag_t|c_t))
$$

where $f_\theta(tag_t|c_t))$ indicates the score output for tag_t at the t-th character by the networks. Given a sentence $c_{[1:T]}$, we can find the best tag path $tag^*_{[1:T]}$ by maximizing the sentence score:

$$
tag^*_{[1:T]} = \arg\max_{\forall l_{[1:T]}} l(c_{[1:T]}, tag_{[1:T]}, \theta).
$$

The Viterbi algorithm can be used for tag inference. The description above shows that it is easy to stack feature vectors, neural networks and tag inference together. Thus, the proposed architecture can be trained in a layer-wised fashion.

3. EXPERIMENTS

Totally 48210 sentences randomly selected from People's Daily were used in our experiments. Prosodic boundaries (PW, PPH and IPH) were labelled by professional annotators with corresponding speech and labeling consistency is ensured. Word segmentation and POS tagging were carried out by a front-end preprocessing tool. The accuracy of word segmentation is 97% and the accuracy of POS tagging is 96%. The corpus was partitioned into three parts: a training set with 43390 utterances, a validation set with 2410 utterances for parameter tuning and a testing set with another 2410 utterances. A character dictionary D of size 4030 was extracted from the training set. A large set of raw texts was also collected from People's Daily for unsupervised embedding feature learning. All texts were preprocessed with text normalization.

In the experiments, PW, PPH and IPH were predicted separately. That is to say, three separate neural network models were trained independently for PW, PPH and IPH using the CURRENNT toolkit [24]. Each character in a sentence was assigned to one of the following three boundary tags: B for a prosodic boundary, NB for a non-boundary, and O for other symbols such as punctuation. Precision (P), recall (R) and F-score (F) were calculated as standard objective evaluation criteria.

A CRF-based prosodic boundary prediction approach was used as baseline and boundary prediction (B, NB and O) was operated at word level. Atomic features in the CRF approach include word identity, POS tags, the length of word and the predicted tag from the previous boundary level. A linear statistical model was applied to optimize the feature templates. Parameters grid search was adopted to achieve the best performance of the CRF model. The CRF++ toolkit[2] was used for the CRF-based prosodic boundary prediction. The baseline results are shown in Table 1.

Boundary	P (%)	R (%)	F (%)
PW	95.34	96.73	96.03
PPH	83.41	83.68	83.06
IPH	84.85	73.39	78.71

Table 1. The results of CRF-based prosody prediction.

Topology	B, BB, BBB, BBBB FFB, FBF, BFF, FBB, BFB, BBF
Num of nodes	32, 64, 128, 256

Table 2. Different network configurations in the experiments.

We investigated the performance of neural network architecture with different topologies, as described in Table 2, where F and B denote a feed-forward layer and a BLSTM layer, respectively. The number of the nodes were kept the same for all hidden layers in every tested network architecture. Specifically, the network input is an M-dimensional feature vector, where M=4030 for the PW prediction and M=4031 for the PPH and IPH prediction [3]. The network output corresponds to the three boundary tags (B, NB and O). All networks were trained with a momentum of 0.9, a learning rate of 1e-3 for PW and 1e-4 for PPH and IPH. BPTT was performed using stochastic gradient descent (SGD) with 32 parallel sentences. The training stops if no lower error on the validation set can be achieved within the last 10 epochs. The best performances for different prosodic boundary levels are shown in Table 3. We interestingly discover that the best performances at different levels are all obtained with a topology of FBB. When we compare Table 3 with the CRF-baseline Table 1, we find that the proposed neural network approach achieves competitive performance at the PW level and significant improvements at the PPH and IPH levels.

We also studied the effectiveness of the character embedding features. Different sizes of unsupervised training data (400M, 800M, 1200M, 1600M and 2000M text) and embedding feature sizes (100, 200, 300 and 400) were tested. The best network architectures, as shown in Table 3, were used in the experiments. Please note that the dimension of feature vector is greatly reduced as compared with the one-hot representation. The results shown in Table 4 indicate that the embedding features can further improve the performance of automatic prosodic boundary prediction.

We further conducted an A/B preference test on the naturalness of the synthesized speech. A set of 100 sentences were randomly selected from the test set and the prosodic boundary labels were achieved by:

(1) CRF-based model in Table 1;

(2) NN with one-hot representation in Table 3;

(3) NN with embedding features in Table 4.

[2]http://taku910.github.io/crfpp/

[3]The predicted tag from the previous level was used as a feature.

Boundary	P (%)	R (%)	F (%)	TP / Num of nodes
PW	96.02	96.69	96.35	FBB / 32
PPH	82.50	86.75	84.57	FBB / 128
IPH	84.06	79.33	81.63	FBB / 64

Table 3. The best performance of each level and the corresponding network topology (TP).

Boundary	P (%)	R (%)	F (%)	Embedding feature size
PW	96.27	96.91	96.59	300
PPH	82.89	87.13	84.96	400
IPH	84.81	79.88	82.27	100

Table 4. The results of neural network architecture with embedding features and the corresponding feature size.

We carried out two sessions of comparative evaluations: (1) vs (2) and (2) vs (3). A set of 20 sentence pairs of each session was randomly selected from the 100 pairs with different prosody prediction results and speech was generated through a typical HMM-based TTS system. A group of 10 subjects were asked to choose which one was better in terms of the naturalness of synthesis speech. The percentage preference is shown in Figure 4. We can clearly see that the NN architecture with one-hot representation can achieve better naturalness of synthesized speech as compared with CRF, while the use of embedding features further improves the natrualness.

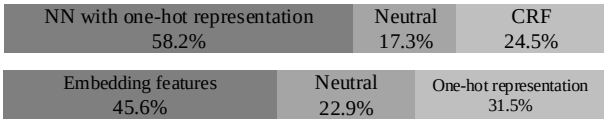

Fig. 4. The percentage preference of A/B test.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose to use neural network architectures to predict prosodic boundary labels directly from Chinese characters without feature engineering. We show that superior performance is achieved by stacking feed-forward and bidirectional long short-term memory (BLSTM) recurrent layers. We obtain useful character embedding features from raw text. Both objective and subjective evaluations show that the proposed neural network approch achieves superior performance over the CRF-based approach and the use of embedding features can further boost the performance. For future work, it is promising to predict PW, PPH and IPH labels in a unified neural network and n-gram character embedding features can be further investigated.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (61175018 and 61571363).

6. REFERENCES

[1] Yao Qian, Zhizheng Wu, Xuezhe Ma, and Frank Soong, "Automatic prosody prediction and detection with conditional random field (crf) models," in *Proceedings of ISCSLP*, 2010, pp. 135–138.

[2] Jingwei Sun, Jing Yang, Jianping Zhang, and Yonghong Yan, "Chinese prosody structure prediction based on conditional random fields," in *Proceedings of ICNC*, 2009, vol. 3, pp. 602–606.

[3] Je Hun Jeon and Yang Liu, "Automatic prosodic events detection using syllable-based acoustic and syntactic features," in *Proceedings of ICASSP*, 2009, pp. 4565–4568.

[4] Vivek Rangarajan, Shrikanth Narayanan, and Srinivas Bangalore, "Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework," in *Proceedings of NAACL HLT*, 2007, pp. 1–8.

[5] Philipp Koehn, Steven Abney, Julia Hirschberg, and Michael Collins, "Improving intonational phrasing with syntactic information," in *Proceedings of ICASSP*, 2000, pp. 1289–1290.

[6] Min Chu and Yao Qian, "Locating boundaries for prosodic constituents in unrestricted mandarin texts," *Computational linguistics and Chinese language processing*, pp. 61–82, 2001.

[7] Xin Nie and Zuo-ying Wang, "Automatic phrase break prediction in chinese sentences," *Journal of Chinese information Processing*, pp. 39–44, 2003.

[8] Jian-Feng Li, Guoping Hu, and Ren-hua Wang, "Chinese prosody phrase break prediction based on maximum entropy model," in *Proceedings of INTERSPEECH*, 2004, pp. 729–732.

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

[10] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of 18th ICML*, 2001, pp. 282–289.

[11] Zhao Sheng, Tao Jianhua, and Cai Lianhong, "Learning rules for chinese prosodic phrase prediction," in *Proceedings of ACL*, 2002, pp. 1–7.

[12] Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu, "Deep learning for chinese word segmentation and pos tagging.," in *Proceedings of EMNLP*, 2013, pp. 647–657.

[13] Wenzhe Pei, Tao Ge, and Chang Baobao, "Maxmargin tensor neural network for chinese word segmentation," in *Proceedings of ACL*, 2014, pp. 293–303.

[14] Mike Schuster and Kuldip K Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

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

[16] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin, "A neural probabilistic language model," *The Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[17] Alex Graves, Navdeep Jaitly, and A-R Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Proceedings of ASRU*, 2013, pp. 273–278.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of NIPS*, 2013, pp. 3111–3119.

[19] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, "Linguistic regularities in continuous space word representations.," in *Proceedings of HLT-NAACL*, 2013, pp. 746–751.

[20] Mairgup Mansur, Wenzhe Pei, and Baobao Chang, "Feature-based neural language model and chinese word segmentation," *Proceedings of 6th IJCNLP*, vol. 1, no. 2.3, pp. 2–3, 2013.

[21] Shin-ichi Horikawa, Takeshi Furuhashi, and Yoshiki Uchikawa, "On fuzzy modeling using fuzzy neural networks with the back-propagation algorithm," *IEEE transactions on Neural Networks*, vol. 3, no. 5, pp. 801–806, 1992.

[22] Ronald J Williams and David Zipser, "Gradient-based learning algorithms for recurrent networks and their computational complexity," *Back-propagation: Theory, architectures and applications*, pp. 433–486, 1995.

[23] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa, "Natural language processing (almost) from scratch," *The Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[24] Felix Weninger, Johannes Bergmann, and Björn Schuller, "Introducing currennt–the munich open-source cuda recurrent neural network toolkit," *Journal of Machine Learning Research*, vol. 15, 2014.