Automatic Prosodic Structure Labeling using DNN-BGRU-CRF Hybrid 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*—The speech corpus with labeled prosodic structure information is crucial for text-to-speech (TTS) synthesis to train a reliable model that can generate high quality natural synthetic speech. Traditional manual prosodic structure labeling is laborious and time-consuming and may encounter an inconsistency problem caused by different annotators. Automatic prosodic labeling is thus desirable, which can not only speed up the labeling process, but also protect the labeling results from the inconsistency problem. This paper presents a DNN-BGRU-CRF hybrid neural network, which aggregates the advantages of deep neural network, bidirectional gated recurrent units and conditional random fields, to label three-level prosodic structure boundaries. It exploits both text and acoustic cues in a neural network framework. Experimental results demonstrate the effectiveness of the proposed model.**

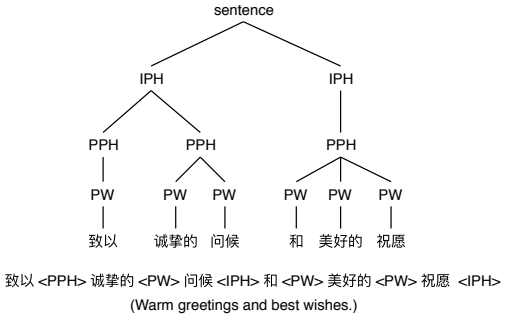

Fig. 1. Example of the prosodic hierarchy.

I. Introduction

A speech corpus with precisely labeled prosodic structure is important to build a high-quality text-to-speech (TTS) synthesis system. Traditionally, the labeling procedure is carried out by professional annotators. Obviously, it is laborious and time-consuming. Furthermore, sometimes there may be inconsistent labeling results between different annotators due to different interpretations of the sentences. An automatic prosodic structure boundary labeling system is essential to speed up the process for preparing the TTS corpus, and can also solve the inconsistency problem.

A typical Chinese prosodic hierachy is illustrated in Fig.1. We adopt a three-levels prosodic hierarchical structure which is commonly used by other researchers [1]. Each lexical word in a sentence is assigned to one of the following four distinct boundary tags: NB for non-boundary, PW for prosodic word boundary, PPH for prosodic phrase boundary, IPH for intonational boundary.

Since the prosodic structure labeling task can be interpreted as a classification problem, various machine learning methods have been proposed for this task. In the early time, decision tree was utilized to model the relation between acoustic features and prosodic boundaries [2].

With the development of statistical learning methods, hidden Markov model (HMM) [3][4], maximum entropy (ME) model [5][6] and conditional random fields (CRF) [7][8] have been proposed to model prosodic structure boundaries.However, local classifiers such as decision tree and sequence models such as HMM and CRF can't well model long-time dependency.

We propose a DNN-BRGU-CRF hybrid neural network which combines dense layers, gated recurrent units based recurrent neural network with CRF layer to label prosodic boundaries. It can utilize both text and acoustic cues, and capture bidirectional context information. Besides, the CRF layer can take sequence level labeling into consideration.

II. Method

A. Features

1) Text features: Much research has indicated there are close correlations between the part-of-speech (POS) tag and the types of prosodic boundaries [9][10]. Therefore, POS information is incorporated into the text-based features for our task. Punctuation symbols have long been accepted as a reliable indicator of prosodic boundaries [11][12]. Besides, prosodic boundaries are affected by the number of syllables in a lexical word (word length). Long lexical word often corresponds to a single prosodic word [13].

Based on the work mentioned above, the textual features adopted in our model are listed as follows:

- POS of the word.

- Punctuation symbol after the word.
- Word length: the numbers of syllables in Chinese lexical word.
- Word identity: represents each word.
- Word-final phoneme identity: represents the Pinyin final of the last syllable of the word.

As depicted in Fig.2, the embedding layer with word ID as input is used to learn task-specific 200-dimensional word embeddings. Similarly, the embedding layer with word-final phoneme ID as input is used to learn task-specific phoneme embeddings.

2) Acoustic features: Besides textual information, we also want to exploit acoustic cues to improve the labeling performance. Many researchers have analysed the acoustic parameters of prosodic structure boundaries. Pre-boundary lengthening is a strong cue to prosodic phrase boundaries [2]. Duration of pause is also relevant to prosodic boundary levels [14]. According to Yang and Wang[15], both prosodic phrase boundaries and intonational phrase boundaries have significant pitch resets.

It is reported that degree of pitch contour change varies with boundary levels based on the statistical analysis on phrase-final F0 slopes [16]. Therefore, we incorporate the pitch and energy contours statistics of word-final syllable to the acoustic features for our task. In order to capture the pitch reset phenomenon, log F0 difference and log energy difference between the last voiced frame of current word and the first voiced frame of succeeding word are considered.

The acoustic features we have explored are listed as follows:

- Duration level of post-word pause. A bucketing scheme is used to transform the continuous pause duration values into several discrete levels: Pause-0 ($0 \leq p < 50$ ms), Pause-1 ($50 \leq p < 150$ ms), Pause-2 ($150 \leq p < 350$ ms), Pause-3 ($350 \leq p < 450$ ms), Pause-4 ($p \geq 450$ ms).
- Duration level of the last syllable of a word (word-final syllable). A bucketing scheme, similar to coping with pause duration, is used to classify the duration of word-final syllable into 9 bins. A 9-dimentional one-hot vector can represent duration level of word-final syllable.
- Statistics on log F0 and log energy contours of word-final syllable: maximum, minimum, range, mean and standard deviation.
- Log F0 difference and log energy difference between last voiced frame of current word and the first voiced frame of succeeding word.

B. Bidirectional Gated Recurrent Units based Neural Network

Prosodic structure labeling is a sequence labeling task, the factors affecting IPH boundary may be adjacent acoustic features or prosodic boundaries far away from the current position. Recurrent neural networks (RNNs) can't capture long time context information due to the vanishing gradient and exploding gradient problems, long short-term memory (LSTM) [17] is designed to solve long time lag problems. Gated recurrent units (GRU) [18], a variant of LSTM, is adopted in our task for its simpler structure.

Automatic prosodic labeling task may need both the past and the future information. Bidirectional RNN with gated recurrent units (BGRU-RNN) [19] is suitable for sequence modeling. It gathers the two directional information by merging the forward and backward GRU outputs. Therefore, BGRU-RNN is adopted in our model to learn the context information.

C. Conditional Random Fields

In prosodic boundary automatic labeling task, the boundary type of the current word is also dependent on the boundary types of adjacent words. CRF focuses on sentence level instead of individual positions or time-steps. To model the label transition from time-step i-1 to time-step i, CRF is parameterized by a state transition matrix of size $K*K$, where K is the label set size. For example in our task, $K = 4$.

Recently, the integrated LSTM-CRF model has been successfully applied in Part-of-speech tagging task [20]. Since our task is also a sequence tagging problem, we employ the CRF layer as the output layer of our model. With such a layer, we can efficiently use past and future labels to predict the current label. The score of a input sequence $W = (w_1, w_2, ..., w_T)$ along with a path of labels $Y = (y_1, y_2, ..., y_T)$ is given by the sum of transition scores and neural scores:

$$s(W, Y, \Theta) = \sum_{i=1}^{T} (f_1(y_{t-1}, y_t, \theta_1) + f_2(w_t, y_t, \theta_2)) \quad (1)$$

where Θ is the parameters of all layers depicted in Fig.2, θ_1 and $f_1(y_{t-1}, y_t, \theta_1)$ are respectively the parameters and score function of CRF part, θ_2 and $f_2(w_t, y_t, \theta_2)$ are the parameters and score function of the rest neural network part, $\Theta = \theta_1 \bigcup \theta_2$. At training step, the model is optimized by maximizing the score of the correct label sequence. At testing step, the Viterbi algorithm is used to obtain the optimal sequence in our task. It can be depicted as follows:

$$Y^{\star} = \arg\max_{Y \in \mathcal{Y}} p(Y|W, \Theta) \quad (2)$$

where $\mathcal{Y}$ is the set of all possible label sequences.

D. Proposed Model

As Fig.2 shows, our proposed model is mainly composed of four parts: the embedding layer, DNN with three dense layers, BGRU layer and the CRF layer. It is termed a DNN-BGRU-CRF model for its hybrid structure.

We use the front DNN to learn a high level representation of text-based features including POS, word length, post-word punctuation. Dropout with 0.5 probability is applied to each dense layer as a regulariser. The trainable embedding layer is applied to learn task-specific embeddings for prosodic boundary labeling task. Then we concatenate the output of DNN with word embedding, word-final phoneme embedding and word acoustic features described in Section II-A2 to form the input to the succeeding BGRU-RNN at timestep t, the last CRF layer are used to get the optimal label sequence. Those four parts are trained altogether as a unified network.

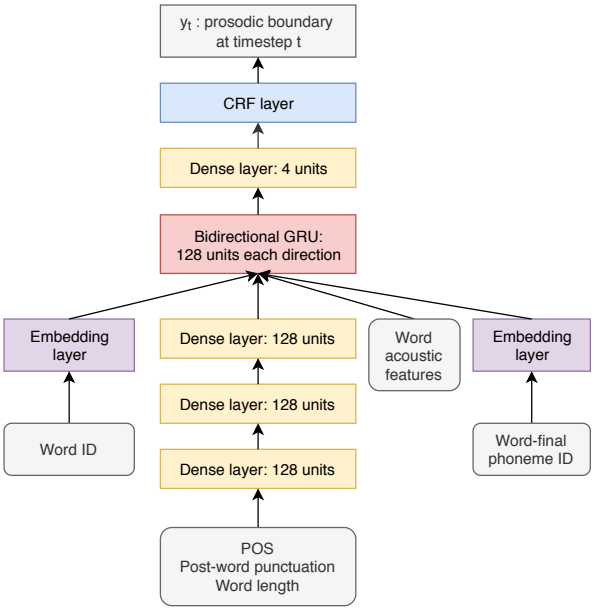

Fig. 2. The neural network architecture for automatic prosodic structure labeling.

III. EXPERIMENT

A. Corpus and Preprocessing

The corpus contains 41,483 utterances, read by a female native Mandarin speaker. Speech is sampled at 16 kHz. Prosodic structure boundaries have been labeled by professional annotator. We randomly select 37,483 utterances for training, another 2,000 utterances for validation, the rest 2,000 utterances for test.

Word segmentation and POS tagging were 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. These features form the input to the dense layers. The input to embedding layer is the index of word in the vocabulary.

From the phonetic time alignments generated from HMM framework, the boundary times of word-final syllable can be obtained. We extract pitch and energy at 5-ms intervals. The acoustic features described in Section II-A2 can be calculated from the alignment result, the pitch and energy contours.

The word-level textual features fed into the front three dense layers are represented by one-hot vectors. Within word-level acoustic features described in II-A2, duration level of word-final syllable and duration level of post-word pause are represented by one-hot vectors, other acoustic features are min-max normalized scalars.

B. Experiment Setup

After testing a set of neural network configuration such as the number of layers and size of each layer, we adopt a framework illustrated in Fig.2. For optimization, we use Adam with a minibatch size of 1024. The initial learning rate is set to 0.0001. Our model is trained for up to 40 epochs. We use TensorFlow [21] to implement the model.

CRF has been previously reported to provide state-of-the-art performance on sequential labeling. It has been used by many researchers for both prosodic boundaries prediction task and auto-labeling task [7] [22]. As it's a fairly strong baseline, we have implemented a CRF-based baseline system using both text and acoustic features. CRF++ toolkit is used for CRF model training.

C. Metrics

In our experiments, four measurements are used to compare our model with the baseline models, including total accuracy for 4-class classification (T-ACC), F1 score for PW (PW F1), F1 score for PPH (PPH F1), F1 score for IPH (IPH F1). T-ACC is calculated as:

$$T\text{-}ACC = \frac{N_{correctly\_labeled\_samples}}{N_{total\_samples}} \tag{3}$$

where $N_{correctly\_labeled\_samples}$ is the number of correctly auto-labeled samples, $N_{total\_samples}$ is the number of total samples.

Manual labeled results serve as ground truth in our task. To evaluate the performance of our model on each prosodic structure category, precision, recall and F1 score can be calculated for each category. For example, the F1 score for IPH can be calculated as follows:

$$precision = \frac{N_{correctly\_labeled\_IPH}}{N_{labeled\_IPH}} \tag{4a}$$

$$recall = \frac{N_{correctly\_labeled\_IPH}}{N_{ground\_truth\_IPH}} \tag{4b}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{4c}$$

where $N_{correctly\_labeled\_IPH}$ is the number of correctly auto-labeled IPH samples, $N_{labeled\_IPH}$ is the number of auto-labeled IPH samples, $N_{ground\_truth\_IPH}$ is the number of manual labeled IPH samples.

D. Results and Discussion

1) Compared with manual labeled result: The auto-labeled results are compared with manual labeled ground truth. The confusion matrix for our model is shown in Table I. The one for baseline CRF is shown in Table II.

From Table I, we can see that our proposed model has good performance on labeling PW and IPH. While almost 38% of PPHs are automatically labelled as PWs. However, as previous research [1] has also reported, this is acceptable and should not always be regarded as error. Confusion matrix describes the conformity between the manual label and automatic label. In the cases of disagreement between automatic label and manual label, the prosodic boundary type on the higher level (e.g. PPH) labeled as a lower level (e.g. PW) are more acceptable than the case that lower one is labeled as a higher one. In TTS, when lower level of prosodic structure is predicted as higher level, the improper inserted breaks would degrade the naturalness of the synthesized speech greatly. On the contrary,

the missing break caused by predicting higher level prosodic structure as lower level is acceptable. By comparing the part above main diagonal in Table I with the part in Table II, our model outperforms the baseline CRF method for its reduced amount of mislabeling the lower prosodic levels as the higher ones. For example, our model labeled result has 98 PPH samples automatically labeled as IPH, whereas CRF model labeled result has 129 cases.

TABLE I
CONFUSION MATRIX FOR PROSODIC STRUCTURE LABELING USING DNN-BGRU-CRF.

Manual \ Automatic	NB	PW	PPH	IPH
NB	**4227 (85.10%)**	691 (13.91%)	47 (0.95%)	2 (0.04%)
PW	505 (7.11%)	**6044 (85.14%)**	542 (7.63%)	8 (0.11%)
PPH	104 (3.47%)	1138 (37.92%)	**1661 (55.35%)**	98 (3.27%)
IPH	14 (0.38%)	14 (0.38%)	177 (4.82%)	**3467 (94.42%)**

TABLE II
CONFUSION MATRIX FOR PROSODIC STRUCTURE LABELING USING CRF.

Manual \ Automatic	NB	PW	PPH	IPH
NB	**4173 (84.01%)**	706 (14.21%)	86 (1.73%)	2 (0.04%)
PW	527 (7.42%)	**5914 (83.31%)**	644 (9.07%)	14 (0.20%)
PPH	101 (3.37%)	1093 (36.42%)	**1678 (55.91%)**	129 (4.30%)
IPH	4 (0.11%)	14 (0.38%)	183 (4.98%)	**3471 (94.53%)**

2) Compared with related models: The following different models are compared for automatic prosodic structure labeling:

1) **CRF**: Conventional CRF model using both text and acoustic features including lexical word, POS tagging labels, word length, post-word punctuation, post-word pause and word-final syllable duration level.

2) **D-BLSTM-CRF**: The input features, the front dense layers and output CRF layer are the same as model illustrated in Fig.2. The difference is that bidirectional LSTM layer is used to model the context dependency.

3) **D-BGRU-S**: The only difference from our proposed model illustrated in Fig.2 is that the softmax layer is employed as the output layer.

4) **D-BGRU-CRF***: The model structure is quite similar to our proposed model except that the word-final phoneme embedding is not included in the input features.
5) **D-BGRU-CRF**: Our proposed automatic prosodic structure labeling model with CRF layer as the output layer.

T-ACC is adopted to compare the performance of different models, F1 scores of PW, PPH, IPH are also recorded to ensure the models performance on these measurements. Experimental results of all models are shown in Table III.

Compared with CRF, our proposed D-BGRU-CRF acheives superior performance on both F1 scores of all prosodic boundary types and total accuracy. More specifically, it improves T-ACC from 0.8131 to 0.8218, IPH F1 score from 0.9525 to 0.9646 and PPH F1 score from 0.6001 to 0.6120. The D-BGRU-CRF outperforms D-BLSTM-CRF, it indicates that the BGRU layer performs better than the BLSTM layer when integrated to our model. Our proposed model D-BGRU-CRF outperforms D-BGRU-S, suggesting that CRF layer considering the sentence level information can improve the labeling performance. Comparing D-BGRU-CRF with D-BGRU-CRF*, the word-final phoneme embedding added to the input improve the T-ACC from 0.8163 to 0.8218. Due to the fact that the word-final phoneme type affects the word-final acoustic feature, integrating both of them as the input features can improve the word-final feature representation than just using the word-final acoustic feature. Among all the models, our proposed model (D-BGRU-CRF) achieves the best performance.

TABLE III
EXPERIMENT RESULTS OF RELATED MODELS.

Model	PW F1	PPH F1	IPH F1	T-ACC
CRF	0.7978	0.6001	0.9525	0.8131
D-BLSTM-CRF	0.8018	0.5984	0.9584	0.8169
D-BGRU-S	0.8026	0.5964	0.9599	0.8199
D-BGRU-CRF*	0.7993	0.5965	0.9586	0.8163
D-BGRU-CRF	0.8066	0.6120	0.9646	0.8218

IV. CONCLUSIONS

In this paper, we investigate the textual features and acoustic features for automatic prosodic structure labeling task and propose a model with a cascade of dense layers, BGRU layer and CRF layer, which can exploit both text and acoustic cues to label three-level prosodic structure boundaries in a unified neural network. We also find that word-final features like word-final phoneme embedding can improve the labeling performance in the neural network method. Experimental results demonstrate effectiveness of the proposed hybrid neural network and also justify the hybrid structure of our model. In future, we wish to explore speaker's habitual movement patterns near the prosodic boundaries based on the videos data to find the possibility of improving the labeling performance with that data.

V. ACKNOWLEDGMENTS

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] X. Ma, W. Zhang, Q. Shi, W. Zhu, "Automatic prosody labeling using both text and acoustic information," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003.

[2] C. W. Wightman, M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on speech and audio processing*, vol. 2, no. 4, pp. 469–481, 1994.

[3] S. Ananthakrishnan, S. S. Narayanan, "An Automatic Prosody Recognizer using a Coupled Multi-Stream Acoustic Model and a Syntactic-Prosodic Language Model," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005, pp. 269–272.

[4] C. Yang, Z. Ling, H. Lu, W. Guo, L. Dai, "Automatic phrase boundary labeling for Mandarin TTS corpus using context-dependent HMM," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*. IEEE, 2010, pp. 374–377.

[5] V. Rangarajan, S. Narayanan, S. Bangalore, "Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 2007, pp. 1–8.

[6] J. Li, G. Hu, W. Zhang, R. Wang, "Chinese prosody phrase break prediction based on maximum entropy model," in *Eighth International Conference on Spoken Language Processing*, 2004.

[7] Y. Qian, Z. Wu, X. Ma, F. Soong, "Automatic prosody prediction and detection with Conditional Random Field (CRF) models," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*. IEEE, 2010, pp. 135–138.

[8] G. 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.

[9] Z.Ying, X. Shi, "An RNN-based algorithm to detect prosodic phrase for Chinese TTS," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 809–812.

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

[11] I. Read, S. Cox, "Using part-of-speech for predicting phrase breaks," in *Eighth International Conference on Spoken Language Processing*, 2004.

[12] P. Taylor, A. W. Black, "Assigning phrase breaks from part-of-speech sequences," 1997.

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

[14] Y. Liu, A. Li, "Cues of prosodic boundaries in Chinese spontaneous speech," *Proc. ICPHS2003, Barcelona*, 2003.

[15] Y. Yang, B. Wang, "Acoustic correlates of hierarchical prosodic boundary in Mandarin," in *Speech Prosody 2002, International Conference*, 2002.

[16] L. Lai, S. Gooden, "Acoustic cues to prosodic boundaries in Yami: A first look," *Speech Prosody 2016*, pp. 624–628, 2016.

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

[18] 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.

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

[20] Z. Huang, X. Wei, and Y. Kai, "Bidirectional lstm-crf models for sequence tagging," *Computer Science*, 2015.

[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R.t Monga, S. Moore, D, G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, "Tensorflow: a system for large-scale machine learning," in *OSDI*, vol. 16, 2016, pp. 265–283.

[22] Z. Zhao and X. Ma, "Active learning for the prediction of prosodic phrase boundaries in chinese speech synthesis systems using conditional random fields," in *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on*. IEEE, 2015, pp. 1–5.