INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Multi-Task Learning for Prosodic Structure Generation using BLSTM RNN with Structured Output Layer

Yuchen Huang[1], *Zhiyong Wu*[1,2], *Runnan Li*[1], *Helen Meng*[1,2], *Lianhong Cai*[1]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China

{huang-yc15, lirn15}@mails.tsinghua.edu.cn,
{zywu, hmmeng}@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

Abstract

Prosodic structure generation from text plays an important role in Chinese text-to-speech (TTS) synthesis, which greatly influences the naturalness and intelligibility of the synthesized speech. This paper proposes a multi-task learning method for prosodic structure generation using bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) and structured output layer (SOL). Unlike traditional methods where prerequisites such as lexicon word or even syntactic tree are usually required as the input, the proposed method predicts prosodic boundary labels directly from Chinese characters. BLSTM RNN is used to capture the bidirectional contextual dependencies of prosodic boundary labels. SOL further models correlations between prosodic structures, lexicon words as well as part-of-speech (POS), where the prediction of prosodic boundary labels are conditioned upon word tokenization and POS tagging results. Experimental results demonstrate the effectiveness of the proposed method.
Index Terms: prosodic structure generation, structured output layer (SOL), bidirectional long short-term memory (BLSTM)

1. Introduction

For a typical Chinese text-to-speech (TTS) system [1][2], as shown in Fig. 1, the input text is first tokenized into lexicon words (LW) with part-of-speech (POS) tagging information, which are sent to the prosodic structure generation module to predict the prosodic boundary labels including prosodic word (PW) and prosodic phrase (PPH). The generated prosodic structure information is then used in grapheme-to-phoneme conversion to derive the proper pronunciations, and also further utilized to predict acoustic parameters such as pitch, duration, pause, spectrum [1][2][3] for further usage in the speech synthesis module. As can be seen, prosodic structure generation from text plays a very important role in Chinese TTS synthesis, which will greatly affect the naturalness and intelligibility of the synthesized speech.

Being aware of the importance of prosodic structure generation in practical TTS system [4][5], lots of methods have been proposed for addressing the problem. In the early time, rule-based methods were usually adopted [6][7][8]. The main idea of these work is to find some explicit rules that could build prosodic structure of a sentence from syntax information. With the development of statistical learning and the availability of prosody annotated corpora, more and more stochastic-based approaches have been proposed for prosodic boundary prediction, including classification and regression tree (CART) [9], hidden Markov model (HMM) [10][11], maximum entropy (ME) model [12], conditional random fields (CRF) [13][14]. Among all the mentioned models, CRF has been reported to achieve best performance in prosodic structure generation [13]. More recently, with the increasing popular of neural networks, recurrent neural network (RNN) is also employed in prosodic structure generation [15] for its outperformance in sequence processing.

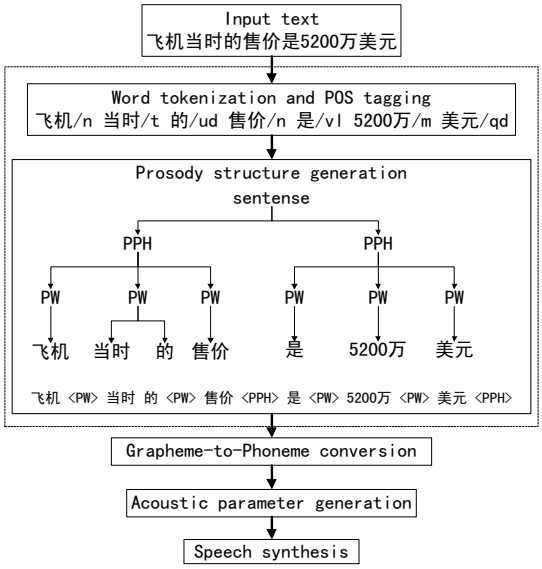

Fig. 1: *A typical Chinese text-to-speech (TTS) system*

As is well known, there are inherent correlations between lexicon word (LW), part-of-speech (POS) and prosodic word (PW), prosodic phrase (PPH). Hence the above approaches try to predict prosodic structures based on the features (e.g. length and POS of adjacent LW) that are derived from the output of the word tokenization and POS tagging module. However, some of the approaches suffer from the feature engineering problem [16]. For example, in CRF, the choice of effective features from a broad set of feature templates is critical to the success of the system. Much efforts are required to design good feature template set based on expert knowledge, which is usually quite label-intensive. Moreover, the existing approaches still lack the ability in capturing the bidirectional context information that are important for prosodic structure generation. For example, some of the monosyllabic LW may be combined with preceding or succeeding LW to form the PW according to the intonation balance requirement.

This paper proposes the use of bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) [17]

http://dx.doi.org/10.21437/Interspeech.2017-949

for prosodic structure generation. Multi-task learning framework with structured output layer (SOL) is further employed to capture the dependency of prosodic structures on POS. The proposed method possesses 3 advantages. (1) The approach can capture the bidirectional context information for prosodic structure generation by virtue of BLSTM. (2) The introduction of SOL can capture the correlations between LW, POS and PW, PPH in a unified framework without the necessity of an additional model for LW tokenization and POS tagging. (3) The proposed method can generate prosodic structures directly from raw Chinese characters without the requirement for feature engineering.

2. Method

The proposed model is implemented based on BLSTM RNN and SOL. With SOL, the proposed model inherits the stronger generalization performance and robustness of multi-task learning by sharing hidden layers and jointed training across different tasks. The use of SOL further allows the proposed model being capable to exploit the dependencies between the prosodic structure generation and the POS tagging prediction.

2.1. Feature Vectors

The proposed approach is designed to use Chinese characters as input. A simple way to represent Chinese character is one-hot vector. However, it will result in high dimensionality and cannot represent the relevance between characters. Character embedding layer [18] can be employed to map each kind of character to a vector with a given dimension. Specifically, characters with similar meanings will get closer representation after processing. Preliminary experiments indicate character embedding vector achieves better performance than one-hot vector in the prosodic structure generation task.

2.2. Bidirectional Long Short-Term Memory (BLSTM)

Prosodic structure generation is a context dependent task that may span short or long time lags. For example, the factors affecting PPH boundary may be adjacent POS tags or prosodic boundaries far away from the current position. Conventional methods such as CRF, HMM and traditional NN cannot well leverage the context information. RNN can deal with this by feeding the hidden layer output activations of the last time step to the hidden layer at current time step. But traditional RNN can only retain short term memory because of the vanishing gradient problem. LSTM is designed to tackle with long term contextual dependencies.

A single LSTM cell can be represent as follows:

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_{cf}\mathbf{C}_{t-1} + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{xf}\mathbf{x}_t + \mathbf{b}_f\right) \quad (1)$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_{ci}\mathbf{C}_{t-1} + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{xi}\mathbf{x}_t + \mathbf{b}_i) \quad (2)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{co}\mathbf{C}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{xo}\mathbf{x}_t + \mathbf{b}_o) \quad (3)$$

$$\mathbf{C}_t = \mathbf{f}_t\mathbf{C}_{t-1} + \mathbf{i}_t\tanh(\mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{W}_{xc}\mathbf{x}_t + \mathbf{b}_c) \quad (4)$$

$$\mathbf{h}_t = \mathbf{o}_t\tanh(\mathbf{C}_t) \quad (5)$$

C_t represents current LSTM cell's memory and h_t is the output of the LSTM cell. By using forget gate f_t, LSTM cell will choose which memory to forget and which to remember, thus it is able to remember some key information for a long time. Input gate i_t and output gate o_t restrict model's input and output, and also make model pay more attention to the key information related to the task.

Furthermore, the prediction of prosodic boundary label may need both preceding and succeeding context information. Bidirectional LSTM (BLSTM) can deal with this problem by using two LSTM layers with different directions and merging the results of two LSTM s to output to upper layers [19].

2.3. Structure Output Layer (SOL)

Prosodic structure generation can be hard to model from raw input features, but a related and simpler auxiliary task like POS tagging can be benefit to the modeling [20]. By sharing easier-to-understand high-level features from shared hidden layer, the prosodic structure generation can acquire general information from the related auxiliary task.

To exploit the dependency between POS tags and prosodic structure, a multi-task learning framework shown as Fig. 2 is employed by setting the POS tagging task as an auxiliary task and the prosodic structure generation task as the primary task. Structure output layer (SOL) [21] is employed to set the prosodic structure generation to be conditioned on the POS tagging task to explicitly exploit the dependency.

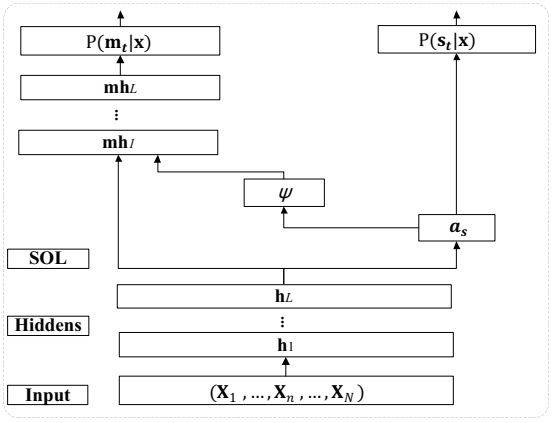

Fig. 2: *The structure of the proposed model,* $\boldsymbol{X}$ *are the input,* $\{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_L\}$ *presents the shared BLSTM-RNN hidden layers before SOL,* $\{\boldsymbol{mh}_1, \ldots, \boldsymbol{mh}_L\}$ *presents private BLSTM-RNN hidden layers after SOL of prosodic structure generation task*

$\mathrm{P}(\mathbf{m_t}|\mathbf{x})$ represents the probability of PW, PPH and not-prosodic-boundary (NPB) label after *t*-th character in current sentence and $\mathrm{P}(\mathbf{s_t}|\mathbf{x})$ represents the probability of each POS tagging label after t-th character. The prosodic structure generation and POS tagging task are computed as:

$$\mathbf{a}_s = (\mathbf{W}_{ha}\mathbf{h}_L + \mathbf{b}_a) \quad (6)$$

$$\mathrm{P}(\mathbf{s_t}|\mathbf{x}) = softmax(\mathbf{a}_s) \quad (7)$$

$$\mathbf{mh}_1 = (\mathbf{W}_{hm}\mathbf{h}_L + \mathbf{b}_m + \mathbf{W}_{am}\Psi(\mathbf{a}_s)) \quad (8)$$

$$\mathrm{P}(\mathbf{m_t}|\mathbf{x}) = softmax(\mathbf{mh}_L) \quad (9)$$

where $\{\boldsymbol{W}_{ha}, \boldsymbol{b}_a\}$ and $\{\boldsymbol{W}_{hm}, \boldsymbol{b}_m\}$ are the weight matrices and bias vectors connecting the shared hidden layer with the outputs associated with the two tasks. Ψ is a non-linear layer used with weight matrix $\mathbf{W}_{am}$ to transmit correlation information from auxiliary output layer $\boldsymbol{a_s}$ to main task's hidden layer. In typical SOL, Ψ and $\mathbf{h_L}$ will directly connect to main task's output layer. However, prosodic structure generation is a complex problem that needs a deeper network for wrapping, private hidden layers $\{\boldsymbol{mh}_1, \ldots, \boldsymbol{mh}_L\}$ are thus added before the main task's output layer.

Same to the conventional multi-task learning network, the proposed model can be trained by minimizing the global loss computed from the weighted sum of costs from two tasks, which is given by:

$$F_g = \alpha F_s + (1 - \alpha) F_m \quad (10)$$

where α is the weight at range [0, 1], F_m and F_s are the costs generated by the main task (prosodic structure generation) and the auxiliary task (POS tagging) respectively.

Compared to other multi-task learning methods, SOL is capable of exploiting dependency between tasks in an explicit way. By training with matched prosodic structure label and POS tag, the proposed model is able to gain related POS tagging information from early POS tagging prediction task and thus to improve the performance of prosodic structure generation task with more accurate prosodic boundary labels.

2.4. Weighted Categorical Cross-entropy

In our work, the main task of prosodic structure generation is a 3-class classification task, which aims to predict prosodic boundary label PW, PPH or NPB (i.e. not-prosodic-boundary) after each Chinese character. The auxiliary POS tagging is a 98-class classification task. As for multiclass classification problem, it is common to use Categorical Cross-entropy as the loss function:

$$\mathrm{F}_m = -\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\sum_{j=0}^{2}\mathbf{T\_m}_{ij}\log(\mathbf{P\_m}_{ij}) \quad (11)$$

$$\mathrm{F}_s = -\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\sum_{j=0}^{97}\mathbf{T\_s}_{ij}\log(\mathbf{P\_s}_{ij}) \quad (12)$$

where **T_m** and **T_s** represent true class of main task and auxiliary task, while **P_m** and **P_s** represent corresponding probability of each predicted class. Each class is treated as of equal importance in Categorical Cross-entropy loss function.

However, in Chinese TTS system, an improper insertion of a prosodic boundary label (PW or PPH) usually cause synthesized speech sound more unnatural than missing a specific prosodic boundary label. Motivated by this, we propose the Weighted Categorical Cross-entropy by adding weight for each class. The class with lower weight θ_i tends to have higher precision and lower recall. The new loss function is given as follows:

$$\mathrm{F}_m = -\frac{1}{\mathrm{N}}\sum_{i=1}^{N}\sum_{j=0}^{2}\theta_j\mathbf{T\_m}_{ij}\log(\mathbf{P\_m}_{ij}) \quad (13)$$

$$\theta_0 = 1 + 2\beta \quad (14)$$

$$\theta_1 = \theta_2 = 1 - \beta \quad (15)$$

where β is a tunable weighting parameter adjusting the targeting of model in the training. When $\beta = 0$, it equals to normal Categorical Cross-entropy. When $\beta > 0$, the new loss function will give class 1 and 2 (for PW and PPH respectively) lower weight during training. Hence, higher precision for PW and PPH prediction will be preferred while retaining the overall accuracy in prosody structure prediction.

3. Experiments

3.1. Experimental Setup

The dataset used for experiments contains 98,211 sentences, with 5,066 distinct Chinese characters, totally 1,054,276 PW and 585,284 PPH boundaries. All the sentences are selected from People's Daily. To make the model more robust and representative to Chinese characters, the punctuations that may directly related to prosodic boundaries (such as comma, period, etc.) are removed. The lengths of the sentences range from 5 to 100 characters. Lexicon word tokenization and POS tagging was processed with the specification defined in [22], amounting to 98 POS tagging labels in total. The prosodic boundary labels (PW and PPH) are labeled by professional annotators, and the labeling consistency between different annotators are checked in a common validation dataset. For prosodic structure generation, 80% of the aforementioned data are used as the training set, 10% data as the validation set, and the remaining 10% data as the test set. Besides, another large set of texts including 455,273 sentences is also collected from People's Daily for unsupervised character embedding feature learning and POS tagging pre-training in the "Enhance" model.

In the experiments, PW, PPH and NPB (i.e. not-prosodic-boundary) labels are predicted simultaneously as the 3-class classification task. As for character embedding, word2vec [23] is adopted for training with the embedding feature size of 300. Such 300-dimensional character embedding vectors are then used as the input of our proposed models. FNN layers of our models have 256 nodes, and BLSTM layers have 256 nodes in both forward and backward layers. All hidden layers apply dropout (dropout rate = 0.4) [24] to prevent over-fitting. Keras [25] with Theano [26] as backend is used to implement the neutral network models. CRF++ toolkit [27] is used to implement the CRF baseline model.

3.2. Evaluation Metrics

In our experiments, we use four measurements to evaluate our model, including F-0.5 score for PPH (PPH F-0.5), F-0.5 score for PW (PW F-0.5), total accuracy for 3-class classification (T-ACC) and POS accuracy (P-ACC). T-ACC is calculated as the number of samples with correct PW, PPH, NPB prediction results divided by the number of all samples. As for POS tagging, P-ACC is the number of samples with correct POS tag divided by the number of all samples related to POS. F-0.5 score is calculate as:

$$\text{F-}0.5 = \frac{(1+0.5^2)*Precision*Recall}{0.5^2*Precision+Recall} \quad (16)$$

F-0.5 score weights precision higher than recall. The reason we choose F-0.5 rather than F1 score is based on the finding the improper insertion of prosodic boundary (especially for PPH) will greatly degrade the naturalness of the synthesized speech. Hence, precision is more important for the prosodic structure generation task. In our work, we mainly use PPH F-0.5 score to compare the performance of different models. At the same time, PW F-0.5, T-ACC and P-ACC are recorded to ensure the model's performance on these measurements.

3.3. Hyper-parameters of the Proposed Model

There are a lot of hyper-parameters in our proposed model, such as the number of BLSTM layers before SOL *lb*, number of BLSTM layers after SOL *la*, parameter of weighted loss function β, learning weight of auxiliary task α and activation function type Ψ. A set of experiments need to be conducted to determine the proper settings for these parameters.

3.3.1. Model Structure

We first try to determine the number of layers of our model. Different combination of *lb* (=1,2,3) and *la* (=1,2,3) values are evaluated, with prefixed value of α=0.3, β=0.3 and Softmax for Ψ. Experimental results of PPH F-0.5 are shown in Table 1. As can be seen, the model with *lb*=2 and *la*=2 have the best performance in PPH F-0.5. Further analysis shows PW F-0.5 and T-ACC under this configuration have similar performance to the highest PW F-0.5 and T-ACC. Hence in the following experiments, we use two BLSTM layers before SOL and two BLSTM layers after SOL as the model structure.

Table 1: *PPH F-0.5 of models with different lb and la*

	lb = 1	*lb* = 2	*lb* = 3
la = 1	0.7630	0.7537	0.7670
la = 2	0.7730	**0.7770**	0.7665
la = 3	0.7657	0.7668	0.7681

3.3.2. Weighted Categorical Cross-Entropy

To validate the effectiveness of the newly proposed weighted categorical cross-entropy loss function, different values of β (=

-0.1,0,0.1,0.2,0.3) are tested, with prefixed *lb*=2, *la*=2, α=0.3 and Softmax for Ψ. The results of PPH F-0.5, PW F-0.5 and T-ACC are shown in Table 2, where the column with β=0.0 lists the results with normal loss function. The results indicate the new loss function can improve PPH F-0.5, PW F-0.5 and T-ACC when β>0 as expected. β=0.3 is chosen for following experiments.

Table 2: *Result of models with different β*

β	-0.1	0.0	0.1	0.2	0.3
PPH F-0.5	0.7593	0.7626	0.7647	0.7676	**0.7770**
PW F-0.5	0.8157	0.8221	0.8316	0.8280	**0.8340**
T-ACC	0.8921	0.8952	0.8983	0.8961	**0.8995**

Table 3: *Results of different activation function Ψ*

	PPH F-0.5	PW F-0.5	T-ACC	P-ACC
Linear	0.8543	0.8927	0.9366	0.9562
Softmax	**0.8584**	0.8940	0.9378	**0.9563**
Sigmoid	0.8563	**0.8964**	**0.9382**	0.9558
ReLU	0.8571	0.8940	0.9375	0.9560
Tanh	0.8551	0.8914	0.9365	0.9559

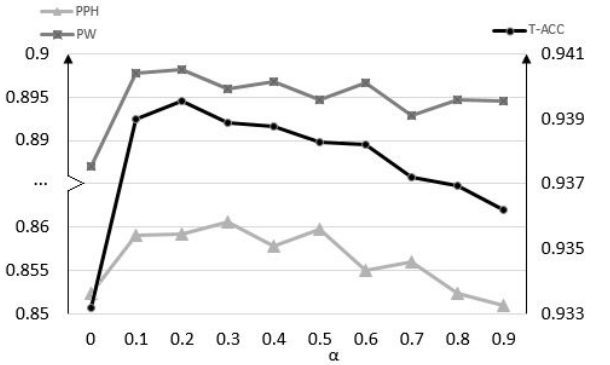

Fig. 3: *Results of different learning weight of auxiliary task α*

3.3.3. Activation Function and Learning Weight of SOL

Table 3 presents the results with different activation function Ψ in SOL structure with prefixed *lb*=2, *la*=2, α=0.3, β=0.3. Softmax works best in PPH F-0.5 and also performs well in other metrics, and is hence used in the following experiments. Fig. 3 illustrates the prosodic structure generation performance using different learning weight of auxiliary task α. The results at α=0.0 represent models without using SOL, whose results are obviously worse than the models with SOL. The model tends to have better performance when α=0.2~0.5. The results also indicate that P-ACC can reach the high value of 95.4% at α=0.3, which proves the effectiveness of our SOL architecture on the auxiliary POS tagging task. Finally α=0.3 is chosen as the optimal weight value.

3.4. Comparison with Related Models

As discussed in introduction session, POS information plays important role in prosodic structure generation. It is valuable to compare different models with or without POS information as input. Moreover, there are huge amount of data with POS tag. If the model related to auxiliary POS tagging task can be pre-trained with such data, the accuracy improvement in POS tagging is believed to be able to boost performance in prosodic structure generation. Hence, an enhanced BLSTM-SOL-E is also proposed. 6 different models are compared for prosodic structure generation:

- **CRF**: Conventional CRF model with lexicon words and POS tagging labels as input.
- **BLSTM-Word**: BLSTM model with lexicon words and POS tags as input.
- **BLSTM-Char**: BLSTM model with Chinese characters and POS tags as input.
- **BLSTM-MTL**: BLSTM model which accepts Chinese character as input (no POS input), with POS tagging as the second independent output layer.
- **BLSTM-SOL**: The proposed BLSTM model with SOL which accepts Chinese characters as the only input.
- **BLSTM-SOL-E**: The enhanced BLSTM-SOL model by adding POS pre-training and preprocessing steps.

The results are shown in Table 4, from which we can see that the BLSTM derived models perform far better than the conventional CRF model. It indicates that, unlike CRF where feature engineering is indispensable, BLSTM can model more complex context features. The BLSTM-Char model achieves the best performance as it can capture the valuable information brought by the "golden" POS tags during both training and evaluation. Comparing BLSTM-SOL with BLSTM-MTL, our proposed BLSTM-SOL perform better, which indicates the necessity and effectiveness in modeling the correlations between PW, PPH and POS; and the SOL provide a good solution to model such dependencies on PW and PPH on the POS information. Furthermore, the BLSTM-SOL-E model achieves performance comparable to the upper bound BLSTM-Char model. This indicates the pre-training of POS auxiliary task really can bring performance improvement to prosodic boundary label prediction, which might be caused by sufficient training of shared hidden layers.

Table 4: *Results comparing different models*

Method	PPH F-0.5	PW F-0.5	T-ACC	P-ACC
CRF	0.8159	0.8752	0.9313	/
BLSTM-Word	0.8599	0.8983	0.9415	/
BLSTM-Char	**0.8704**	**0.9084**	**0.9478**	/
BLSTM-MTL	0.8524	0.8870	0.9332	0.9525
BLSTM-SOL	0.8606	0.8959	0.9389	0.9543
BLSTM-SOL-E	*0.8648*	*0.9000*	*0.9421*	*0.9614*

4. Conclusions

This paper proposes a multi-task learning method for prosodic structure generation using BLSTM and SOL by predicting prosodic boundary labels directly from the Chinese characters. BLSTM RNN is used to capture the bidirectional contextual dependencies of prosodic boundary labels. While SOL further models correlations between prosodic structures, lexicon words as well as POS information. By using weighted categorical cross-entropy as loss function, the performance of the model can be further improved. Experiment results prove that the performance of the proposed method is close to that of the model which need extra POS tagging as input. The further advantage of our model is we can get prosody structure and POS tag at the same time. In the future, we will investigate the possibility of build a unified model for all text analysis tasks for Chinese TTS synthesis.

5. Acknowledgement

This work is supported by National High Technology Research and Development Program of China (2015AA016305), National Natural Science Foundation of China (NSFC) / Research Grants Council of Hong Kong (RGC) joint research scheme (61531166002, N CUHK404/15) as well as NSFC (61375027, 61433018) and National Social Science Foundation of China (13&ZD189).

6. References

[1] Chou F, Tseng C, Chen K, Lee L. A Chinese text-to-speech system based on part-of-speech analysis, prosodic modeling and non-uniform units. [in] Proc. ICASSP, 923-926, 1997.

[2] Wu Z, Cao G, Meng H, Cai L. A unified framework for multilingual text-to-speech synthesis with SSML specification as interface. Tsinghua Science and Technology, 14(5): 623-630, 2009.

[3] Bartkova K, Jouvet D. Automatic detection of the prosodic structures of speech utterances. [in] Proc. ICSC, 1-8, 2013.

[4] Zhao S, Tao J, Cai L. Rule-learning based prosodic structure prediction. Journal of Chinese Information Processing, 16(5): 30-37, 2002.

[5] Chu M, Peng H, Zhao Y, Niu Z, Chang E. Microsoft Mulan - a bilingual TTS system. [in] Proc. ICASSP, 2003.

[6] Gee J, Grosjean F. Performance structures: A psycholinguistic and linguistic appraisal. Cognitive Psychology, 15: 411-458, 1983.

[7] Cao J, Zhu W. Syntactic and lexical constraint in prosodic segmentation and grouping. [in] Proc. Speech Prosody, 2002.

[8] Wang H. Prosodic words and prosodic phrases in Chinese. Chinese Language, 6: 525-536, 2000.

[9] Wang M, Hirschberg J, Predicting intonational boundaries automatically from text. [in] DARPA Speech and Natural Language Workshop, 378-383, 1991.

[10] Taylor P, Black A. Assigning phrase breaks from part-of speech sequences. Computer Speech and Language, 12(4): 99-117, 1998.

[11] Nie X, Wang Z. Automatic phrase break prediction in Chinese sentences. Journal of Chinese information Processing, 17(4): 39-44, 2003.

[12] Li J, Hu G, Wang R. Chinese prosody phrase break prediction based on maximum entropy model. [in] Proc. Interspeech, 729-732, 2004.

[13] Qian Y, Wu Z, Ma X, Soong F. Automatic prosody prediction and detection with conditional random field models. [in] Proc. ISCSLP, 135-138, 2010.

[14] Levow G. Automatic prosodic labeling with conditional random fields and rich acoustic features. [in] Proc. IJCNLP, 217-224, 2008.

[15] Ding C, Xie L, Yan J, Zhang W, Liu Y. Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features. [in] Proc. ASRU, 98-102, 2015.

[16] Zheng X, Chwen H, Xu T. Deep learning for Chinese word segmentation and POS tagging. [in] Proc. EMNLP, 647-657, 2013.

[17] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation, 9(8): 1735-1780, 1997.

[18] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 3111-3119, 2013.

[19] Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11): 2673-2681, 1997.

[20] Caruana R. Multitask learning. Learning to learn, Springer US, 95-133, 1998.

[21] Swietojanski P, Bell P, Renals S. Structured output layer with auxiliary targets for context-dependent acoustic modelling. [in] Proc. InterSpeech, 2015.

[22] Yu S, Duan H, Swen B, Chang B. Specification for corpus processing at Peking University: Word segmentation, POS tagging and phonetic notation. Journal of Chinese Language and Computing, 13(2): 121-158, 2003.

[23] Word2vec [OL]. [2017-3-14]. https://code.google.com/p/word2vec/

[24] Hinton G, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.

[25] Chollet F. Keras: Deep learning library for theano and tensorflow. 2015.

[26] Bergstra J, Breuleux O, Bastien F, Lamblin P, Pascanu R, Desjardins G, Turian J, Warde-Farley D, Bengio Y. Theano: A CPU and GPU math compiler in Python. [in] Proc. SciPy, 1-7, 2010.

[27] Kudo T. CRF++: Yet another CRF toolkit [OL]. [2017-3-14]. http://crfpp. sourceforge. net.