9th ISCA Speech Synthesis Workshop
13-15 Sep 2016, Sunnyvale, USA

Mandarin Prosodic Phrase Prediction based on Syntactic Trees

Zhengchen Zhang⋆*, Fuxiang Wu*†*, Chenyu Yang*⋆*, Minghui Dong*⋆*, and Fugen Zhou*†

⋆Human Language Technology Department,
Institute for Infocomm Research, A*STAR, Singapore, 138632
†BeiHang University, Beijing, China, 100191

Abstract

Prosodic phrases (PPs) are important for Mandarin Text-To-Speech systems. Most of the existing PP detection methods need large manually annotated corpora to learn the models. In this paper, we propose a rule based method to predict the PP boundaries employing the syntactic information of a sentence. The method is based on the observation that a prosodic phrase is a meaningful segment of a sentence with length restrictions. A syntactic structure allows to segment a sentence according to grammars. We add some length restrictions to the segmentations to predict the PP boundaries. An F-Score of 0.693 was obtained in the experiments, which is about 0.02 higher than the one got by a Conditional Random Field based method.
Index Terms: Prosodic Phrase, Text-To-Speech, Prosody Break

1. Introduction

Intonation phrase, prosodic phrase (PP), and prosodic word (PW) are three levels of the hierarchical prosodic structure that describe the rhythm of mandarin speech [1]. The research of prosodic word boundary prediction has been widely conducted, and the results have achieved 0.9 in terms of F-score [2, 3, 4]. However, the accuracies of prosodic phrase prediction [5, 6, 7] are not as good as those of prosodic word prediction. A reason is that the number of characters in a prosodic phrase is much more than that in a prosodic word. If one takes the PP prediction as a classification problem, the positive items (breaks) are much less than the negative items (non-breaks). This imbalance will cause problems to traditional classifiers. The second reason is that many methods need prosodic word prediction results to predict the prosodic phrases. Although state-of-the-art methods of PW prediction can achieve about 90% accuracy, the error of prosodic word prediction may affect the PP prediction. Third, unlike the PW breaks, the PP break detection is a quite subjective problem that different people may pause at different positions of a sentence. For example, "该产品为总公司创造了可喜的经济效益" (The product created fascinating benefit for the parent company.) can be read as "该产品 | 为总公司创造了 | 可喜的经济效益". It also can be read as "该产品为总公司 | 创造了可喜的经济效益". Hence, the predicted results by a program may be not same with the manually labelled one, but it is still reasonable.

To address the above problems, we propose a rule based method that does not take the PP prediction as a classification problem. It does not need the prosodic word information neither. The method relies on the syntactic parsing result of a sentence, and length restrictions are added to predict the PP breaks.

Before describing the detail of the system, we introduce some related work that use the syntactic structure and length information to do prosodic modelling. In [6], syntactic structure and dependency structure are employed to improve the mandarin prosodic boundary prediction. They found that the level information of a word in the syntactic tree is one of the best five features for predicting prosodic phrase boundaries. The authors of [8] demonstrated that the syntactic structure information was able to significantly improve the duration generation in a speech synthesis system. Zhang et al noticed that the prosodic structure of a sentence is also a tree, and they proposed a method of converting a syntactic tree to a prosodic tree [5]. Synchronous Tree Substitution Grammar (STSG) was generated to describe the probabilistic mapping rules between the two types of trees. In [9], the lengths of paths between two adjacent leaf nodes, as well as between two leaf nodes and their latest common ancestor node of a syntactic tree were used to improve the PP prediction. The syntax tags from the common ancestor node to the leaf nodes were also considered. One can see that [9] actually combined the syntactic structure and the length information to detect PP boundaries. Some other works also employed the length information. In [10], a length optimized Chinese prosodic phrasing model was proposed. The authors used support vector machine (SVM) to generate several break candidates first, and then used a statistical model to compute the probability of an utterance given the prosodic phrase length and prosodic word breaks. The authors extended their work in [11], in which they use HMM models to predict the probability instead of the naive bayes algorithm.

Most of existing methods take the PP prediction as a classification problem. Classifiers like decision tree [12], SVM [10] and maximum entropy based methods [13, 7], etc. are used. Besides syntactic and length information,

http://dx.doi.org/10.21437/SSW.2016-26

other features used by traditional methods include Part-Of-Speech (POS), tone of a syllable, number of lexicon words in the context, length of a lexicon word, and so on.

Many works of how syntactic/semantic structure influences the location of intonational boundaries have been published in the literature of linguistics. The most related one is [14]. The authors proposed a method of predicting intonational boundaries named Left hand side/Right hand side Boundary hypothesis (LRB). Several experiments demonstrated that, at a word boundary, the size of the recently completed syntactic constituent and the size of the upcoming syntactic constituent affected the likelihood of producing a intonational boundary. The method calculate the LRB weight for each word boundary and the highest weight indicates the most likely place for a boundary to occur. Besides the sizes of the left and right syntactic constituents (e.g. phonological phrases, which is used in the paper), the method needs to consider the semantic relatedness between adjacent words because it calculates the weights word by word. In our work, we propose a top-down processing method. One only needs to consider the length of left and right phrases. The semantic relatedness is guaranteed by the syntactic structure. In [15], the authors also use a control parameter to determine the number of prosodic phrases in a sentence. The number of PPh is calculated by $m = Length\ of\ Sentence/n$, where n is set to 5 or 7. We can see n is actually the average length of a PPh. This is the only paper that gives the explicit number of PPh lengths of mandarin. In our work, we tried different lengths of a PPh from 4 to 9.

The detail of the proposed algorithm is described in Section 2. The evaluation results and the discussion are reported in Section 3. We conclude our work in Section 4.

2. PP boundary detection

We noticed the prosodic phrase is a short phrase that with relatively complete semantic meaning and with length restriction. The non-leaf nodes in the syntactic tree of a sentence often denote a noun phrase or a verb phrase. The phrases contain some semantic meaning naturally. Considering the length restriction, we count the lengths of PPs in a manually annotated data set which contains $4,821$ sentences and $14,667$ PPs. The counting results are shown in Figure 1. We can find that most of the PPs are less than 10 characters. We have checked those phases with more than 10 characters. Most of them can be further split into two or more phrases. For example, "有二十六支甲级球队参赛" (Twenty six Class A teams attended the competition) can be read as "有二十六支 | 甲级球队 | 参赛". It demonstrates the problem that the PP segmentation is a subjective matter. In this section, we will first introduce the syntactic parsing method used in this work, and the algorithm of PP prediction is described later.

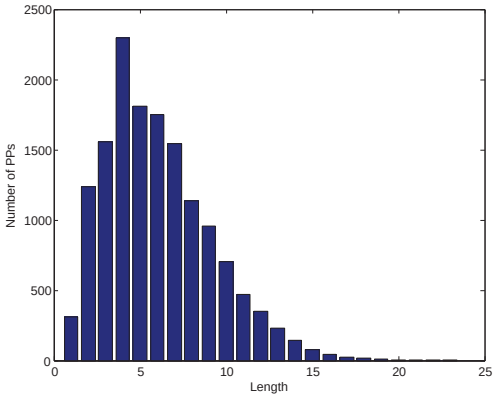

Figure 1: Statistic of the lengths of PPs.

2.1. Syntactic Parsing

The syntactic structures normally include the phrase structure and the dependency structure. The dependency structure (tree) represents the syntactic relationships between two words of a sentence, and the phrase structure (constituent) is about the relationships of the phrases of a sentence and can be converted to the dependency structure with some head finding rules. Given a sentence "剽窃他人作品所引起的纠纷由参赛者本人负法律责任" (The participant themselves will bear legal responsibility for the disputes caused by plagiarizing another's works), its phrase structure is shown in Fig. 2.

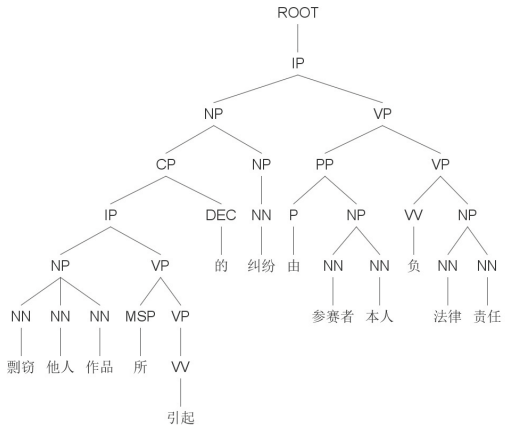

Figure 2: An example of the syntactic tree.

As using phrases to detect the boundary, we employ the phrase structure to depict the syntactic information of a sentence, and utilize Stanford parser [16, 17, 18] to generate the structure of a sentence. The parser contains models with probabilistic context-free grammars and a shift-reduce parsing model. As the shift-reduce constituency parser is much faster than the previous one with competitive accuracy, we select the shift-reduce parser and train it using the same data set-up as [19].

2.2. Syntactic Structure based PP Boundary Detection

Algorithm 1 Prosodic phrase prediction relying on the syntactic structure

Given a sentence S and its syntactic tree
Set two variants $LowThresold(LT)$ and $UpThreshold(UT)$
Set a vector v which length is the character number of S
Let all the items in v be 0 (not a PP boundary)
Find the node n at the most top level of the tree which has more than one children
Get the children list C of n
Get the phrases of the nodes in C
Let the boundaries after each phrase be 1
Set up a stack T
Push the nodes in C into T from right to left
while T is not empty **do**
    $n \leftarrow T.pop()$
    Get the phrase p of n
    Start index $SI \leftarrow$ index of p's first character in S
    End index $EI \leftarrow$ index of p's last character in S
    Last separation $LS \leftarrow \arg\max_{LS} v[LS] = 1$ where $LS <= SI$
    Next separation $NS \leftarrow \arg\min_{NS} v[NS] = 1$ where $NS > SI$
    if $length(p) \geq LT$ and $NS - EI \geq LT$ **then**
        $v[EI] = 1$
    else if $EI - LS \geq LT$ and $NS - EI \geq LT$ **then**
        $v[EI] = 1$
    else if $SI - LS \geq LT$ and $NS - SI \geq LT$ **then**
        $v[SI] = 1$
    if $length(p) \geq UT$ **then**
        Get children nodes of n
        Push the nodes into T from right to left
Return v

The proposed method is shown in Algorithm 1. The input of the algorithm is a sentence and its syntactic tree. The output is a vector v that demonstrates whether the boundary after each character is a PP boundary. If $v[i] = 1$, the boundary after the i_{th} character is a PP boundary. Otherwise, it is not. Two thresholds $LowThreshold$ and $UpThreshold$ are used to give the length restrictions. We use the sentence shown in Fig. 2 to illustrate how the algorithm works. First, we take the top level phrases of a sentence as prosodic phrases, i.e., the boundary after the character "纷" is set to be a PP boundary. Then we check and split the phrases in lower levels in the tree. A stack T is built, and we push the two nodes VP and NP in level 3 to the stack. It is worth noting that we push the nodes from right to left because we will read the NP first. Then we get the phrase of NP "剽窃他人作品所引起的纠纷". The starting index $SI = 0$ and ending index $EI = 11$. The last split index LS is 0. The next split index NS is 11. If we set $LowThreshold = 4$ and $UpThreshold = 7$. Here $length(p)$ is 12, which is bigger than $LowThreshold$. $NS - EI = 11 - 11 = 0$, and $SI - LS = 0 - 0 = 0$. Hence none of the three conditions was met. Then we push the two children of NP into the stack because $length(p) = 12 > 7$. In the next loop, the node n is CP in the $4th$ level of the tree, and the phrase p is "剽窃他人作品所引起的". Then $length(p) = 10$, $SI = 0$, $EI = 9$, $LS = 0$, $NS = 11$. Still none of the conditions was met. The children of CP are pushed into the stack because $length(p) = 10 > 7$. In the next loop, the node n is IP in the $5th$ level. The phrase p is "剽窃他人作品所引起". Then $length(p) = 9$, $SI = 0$, $EI = 8$, $LS = 0$, $NS = 11$. Still none of the conditions was met. Hence, the children of IP are pushed into the stack. In the next loop, the node n is NP in the $6th$ level. The phrase p is "剽窃他人作品". Then $length(p) = 6$, $SI = 0$, $EI = 5$, $LS = 0$, $NS = 11$. Here $length(p) > LowThreshold$ and $NS - EI > LowThreshold$, then we set $v[5] = 1$. We will not describe the following steps due to the article length limitation. The final results are "剽窃他人作品 | 所引起的纠纷 | 由参赛者本人 | 负法律责任".

One can see that the several **if** conditions are key steps of the algorithm. They can be explained in the plain language as follows.

1. If a phrase has a reasonable length, and there is a quite long distance between the end of the phrase and the next boundary, we separate the phrase and its following phrase.

2. If a phrase is very short, and the end of the phrase is far away from the last boundary. In the mean time, the end index is far away from the next boundary, we separate the phrase and its following phrase. The short phrase is combined with its previous phrase.

3. If a phrase is very short, but the start of the phrase is far away from the last boundary, and also the start index is far away from the next boundary, we separate the phrase and its previous phrase. The short phrase is combined with its following phrase.

4. If the length of a phrase is too long, we add its children into the stack to see whether it can be further separated.

Two special cases are considered after the boundaries are predicted. First, in most of the cases, the characters 的 and 了 are not labelled as the end of prosodic phrases by human. We set the boundaries after these two characters to be 0. Then we get the position NS of the nearest boundary after the character. If the distance between the index of 的 and NS is longer than the low threshold, we set the boundary of the word after 的 to be 1. The second special case is that some phrases have less than 3 characters. These short phrases are combined with their neighbours.

Method	Accuracy(%)	Pre(%)	Rec(%)	F1(%)
Proposed	90.1	70.9	67.9	69.3
CRF	90.6	78.1	59.3	67.4
CRF_N	85.0	61.6	22.4	32.8

Table 2: System performance of different methods.

We first check the lengths of its previous and next phrases. Then it is combined with the shorter one.

3. Experimental Results

We evaluate the method on a corpus with 4387 sentences. A subset of 500 sentences were randomly selected as the development set, and another 500 sentences were taken as the testing set. The proposed algorithm is compared with a Conditional Random Field (CRF) based method. The CRF based method is implemented using CRF++ [1]. The parameter c is tuned on the development set. The possible values of c are set to be $(0.1,\ 0.5,\ 1,\ 1.5,\ 2,\ 2.5,\ 3,\ 3.5,\ 4,\ 4.5,\ 5)$, and the best value is 2. The other parameters are set to be default values. The features used in the CRF based method are listed in Table 1. In which, the POS, Syntactic Structure, and Dependency Structure are obtained using Stanford Parser [20, 21]. The Prosodic Word feature is generated using the algorithm proposed in [3], whose accuracy is above 0.9.

The system performance is show in Table 2. We can find that the proposed method obtained 0.693 in terms of F1 score, which is 1.9% higher than the CRF based method. The precision of the CRF based method is higher than the proposed one, while the proposed method can find more PP boundaries. The CRF based method is trained using 3387 sentences, and tuned using 500 sentences. One has to obtain many features shown in Table 1, in which the most important one is the Prosodic Word. We show the performance of the CRF method without the Prosodic Word features, named CRF_N. One can find that the system almost crashes without the Prosodic Word feature. If the Prosodic Word prediction method is not good enough, the CRF based method will not work.

The proposed method only needs the syntactic structure of the sentences. A development set of 500 sentences is required to adjust the parameters. The possible values of low threshold are set to be $(3,\ 4,\ 5,\ 6)$, and the possible values of up thresholds are $(7,\ 8,\ 9,\ 10)$. On the development set, the best performance was obtained when low threshold and up threshold were 4 and 8 respectively. The system performance with different low and up thresholds are illustrated in Fig 3. We can see that the $UpThreshold$ does not affect the F-score much. When $LowThreshold$ is set to be 3 and 4, the F-scores are better than the others.

[1] https://taku910.github.io/crfpp/

It is coincident with the experience that a prosodic phrase is at least 3 or 4 characters long. If it is too short, it can be taken as prosodic words. If it is too long, people may need breath when speaking the phrase.

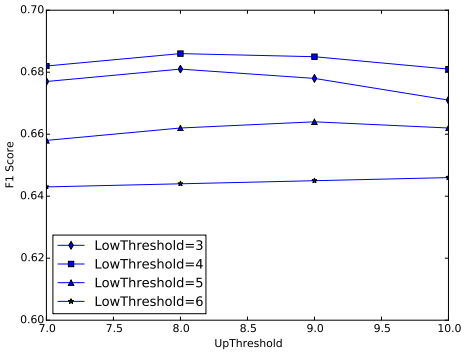

Figure 3: System performance with different threshold values obtained on the development set.

4. Conclusion

In this paper, we have proposed a rule based method of predicting prosodic phrase in mandarin speech synthesis. The method does not rely on huge manually labelled PP corpus or the results of prosodic word prediction. Experimental results demonstrated that the method was able to generate reasonable prosodic phrase boundaries. A problem is that the method relies on the syntactic parsing. When the sentence is ungrammatical, the performance will decrease. Future work should be done to make the algorithm more robust.

5. References

[1] Fu-chiang Chou, Chiu-yu Tseng, and Lin-shan Lee, "Automatic segmental and prosodic labeling of mandarin speech database.," in *ICSLP*, 1998.

[2] Zhengchen Zhang and Minghui Dong, "The power of special characters in prosodicword prediction for chinese tts," in *9th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2014*. IEEE, 2014, pp. 280–283.

[3] Zhengchen Zhang, Fuxiang Wu, Minghui Dong, and Fugen Zhou, "Mandarin prosodic word prediction using dependency relationships," in *The 19th International Conference on Asian Language Processing (IALP 2015)*, 2015.

[4] Hua-jui Peng, Chi-ching Chen, Chiu-yu Tseng, and Keh-jiann Chen, "Predicting prosodic words from lexical words-a first step towards predicting prosody

Category	Detail
Pinyin	Tone
	Syllable number of a word
	Initial and final of syllable
Part-Of-Speech (POS)	Part-Of-Speech
Prosodic Word (PW)	Is a PW boundary
	Forward and backward position of current PW
Syntactic Structure	Phrase type of father phrase (FP)
	Phrase type of grandfather phrase (GP)
	Phrase type of great-grandfather phrase (GGP)
	Level of current word in the tree
	The height of the tree
	Path Length (PL) [6] to neighbour words
	Forward and backward indices of current word in FP
	Forward index of current word in GP and GGP
	Index of FP in GP, and index of GP in GGP
	Children numbers of FP, GP, and GGP
Dependency Structure	Dependency type
	Index and POS of governor
	Number of dependent
	Level in the dependency tree
	Index and POS of grandfather

Table 1: Features used in the CRF based method.

from text," in *International Symposium on Chinese Spoken Language Processing, (ISCSLP)*. IEEE, 2004, pp. 173–176.

[5] Xiaotian Zhang, Yao Qian, Hai Zhao, and Frank K Soong, "Break index labeling of mandarin text via syntactic-to-prosodic tree mapping," in *Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on*. IEEE, 2012, pp. 256–260.

[6] Hao Che, Jianhua Tao, and Ya Li, "Improving mandarin prosodic boundary prediction with rich syntactic features," in *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.

[7] Fangzhou Liu, Huibin Jia, and Jianhua Tao, "A maximum entropy based hierarchical model for automatic prosodic boundary labeling in mandarin," in *6th International Symposium on Chinese Spoken Language Processing, 2008. ISCSLP'08.* IEEE, 2008, pp. 1–4.

[8] Yansuo Yu, Dongchen Li, and Xihong Wu, "Prosodic modeling with rich syntactic context in hmm-based mandarin speech synthesis," in *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on*. IEEE, 2013, pp. 132–136.

[9] Zhigang Chen, Guoping Hu, and Wei Jiang, "Improving prosodic phrase prediction by unsupervised adaptation and syntactic features extraction," in *Eleventh Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2010.

[10] Honghui Dong and Jianhua Tao, "Length optimized chinese prosodic phrasing model," in *ICCC*, 2005.

[11] Honghui Dong, Jianhua Tao, and Bo Xu, "Chinese prosodic phrasing with a constraint-based approach.," in *INTERSPEECH*, 2005, pp. 3241–3244.

[12] Zhao Sheng, Tao Jianhua, and Cai Lianhong, "Prosodic phrasing with inductive learning," in *ICSLP2002 Denver, USA*, 2002, pp. 2417–2420.

[13] Jian-Feng Li, Guoping Hu, and Ren-hua Wang, "Chinese prosody phrase break prediction based on maximum entropy model.," in *INTERSPEECH*, 2004.

[14] Duane Watson and Edward Gibson, "The relationship between intonational phrasing and syntactic structure in language production," *Language and cognitive processes*, vol. 19, no. 6, pp. 713–755, 2004.

[15] Keh-Jiann Chen, Chiu-yu Tseng, and Chia-hung Tai, "Predicting prosody from text," *Lecture notes in computer science*, vol. 4274, pp. 179, 2006.

[16] Dan Klein and Christopher D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the*

41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan., 2003, pp. 423–430.

[17] Dan Klein and Christopher D. Manning, "Fast exact inference with a factored model for natural language parsing," in *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada*, 2002, pp. 3–10.

[18] Yue Zhang and Stephen Clark, "Transition-based parsing of the chinese treebank using a global discriminative model," in *Proceedings of the 11th International Workshop on Parsing Technologies (IWPT-2009), 7-9 October 2009, Paris, France*, 2009, pp. 162–171.

[19] Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu, "Fast and accurate shift-reduce constituent parsing," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1*, 2013, pp. 434–443.

[20] Roger Levy and Christopher Manning, "Is it harder to parse chinese, or the chinese treebank?," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 439–446.

[21] Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D Manning, "Discriminative reordering with chinese grammatical relations features," in *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*. Association for Computational Linguistics, 2009, pp. 51–59.