Improving Prosodic Phrase Prediction by Unsupervised Adaptation and Syntactic Features Extraction

Zhigang Chen, Guoping Hu, Wei Jiang

iFly Speech Lab, University of Science and Technology of China
Hefei, Anhui, 230027
{chenzhigang, applecore}@ustc.edu

Abstract

In the state-of-the-art speech synthesis system, prosodic phrase prediction is the most serious problem which leads to about 40% of text analysis errors. Two optimization strategies are proposed in this paper to deal with two major types of prosodic phrase prediction errors. First, unsupervised adaptation method is proposed to alleviate the mismatching problem between training and testing. Second, syntactic features are extracted from parser and integrated into prediction model to ensure that the consistency between the predicted prosodic structure and the syntactic structure. We examine our methods on an in-house Mandarin speech synthesis system and experiment results show that both strategies yield positive effects and the sentence unacceptable rate significantly drops from 15.9% to 8.75%.

Index Terms: speech synthesis, prosodic phrase prediction, unsupervised adaptation, syntactic structure

1. Introduction

One of the most important measurements of a speech synthesis system is naturalness that depends on a lot of prosodic features as accent, stress, and prosodic structure. The correctness of the prosodic structure has a serious influence on naturalness of synthesized speech. Hierarchical prosodic structure is widely employed to distinguish different levels of pauses between words in speech. Normally, the prosodic structure consists of four layers: syllable, prosodic word, prosodic phrase, and breath phrase group. The detailed definitions of each layer in Chinese have been brought forward in [1]. In this paper, we discuss about prosodic phrase prediction, which is to split a sentence into several prosodic phrases. Prosodic phrase is also called intonation phrase by some researchers. For the sake of consistency, the former is used in this paper.

Methods based on rules or statistics have been proposed to solve the problem of prosodic phrase prediction for a long time. Based on the text level information, handcrafted rules were designed to predict the prosodic phrase [2]. This approach is simple and convenient, but it is such a time-consuming work that it is impossible to handle all possible linguistic phenomena. Recently, large scale corpus based machine learning methods were widely used in this field and have achieved quite good performance, including CART [1], maximum entropy (ME) [3], memory-based learning [4], and so on.

Although prosodic phrase prediction has been studied for many years, it still remains as the most critical problem in speech synthesis system, especially for Mandarin system. According to our analysis on the performance of Interphonic 6.0 system, which is one of the most popular Mandarin speech synthesis products, about 40% of text analysis errors are caused by prosodic phrase errors.

After analyzing the prosodic phrase prediction errors in detail, we find that more than half of errors have arisen due to two reasons. One reason is that our prediction model is out-of-date when it is employed to process the latest news text because our model is trained on a labeled sentence corpus selected from People’s Daily 1993 to 1996. By comparing the prediction performance on different-style corpus or even People’s Daily of different years, we find that the mismatching problem is also encountered in the prosodic phrase prediction. Another major type of errors arise because the predicted prosodic phrase structures of some sentences are too incompatible with their syntactic structures. These incompatible prediction results are always evaluated as extremely unacceptable in the subjective evaluation.

To minimize the above two major types of errors in prosodic phrase prediction, two targeted optimization strategies are proposed in this paper. Firstly, unsupervised adaptation is proposed to deal with the mismatching between training and testing, and the unsupervised method is realized by sentence selection with the help of comma in sentences. Secondly, syntactic features are extracted from syntactic parser and integrated into prediction models to handle the second major type of errors, and how to integrate syntactic structure information with other linguistic features is investigated. We examine our methods on our Interphonic 6.0 system, and the experimental results show that both of the two strategies yield positive results and the sentence unacceptable rate significantly drops from 15.9% to 8.75%.

The rest of the paper is organized as follows. In section 2, the strategy of unsupervised adaptation is introduced in detail. Section 3 gives the procedure of syntactic features extraction to enhance the feature set of prediction model construction. The experimental results are presented in section 4 and conclusions are drawn in section 5.

2. Unsupervised adaptation of prosodic phrase prediction

The task of labeling prosodic phrase of sentences is very time-consuming. And even if we can put lots of effort to develop a large scale training corpus, it is never once and for all. As we all know the evolvement of language is accompanied by the development of society. And the usage of language in various domains may be different from each other. And current state-of-the-art prosodic phrase prediction modeling method, such as ME and Conditional Random Fields (CRF), always utilize the word text information itself as prediction features, so it will suffer more from the mismatching between training and testing, because the text information evolves much more quickly and changes more than the syntactic grammars. So it is worthy to study how to automatically adapt for different periods or domains, which can be formulated as unsupervised adaptation task of prosodic phrase prediction, similar as
language model unsupervised adaptation for speech recognition.

2.1. Comma in sentence

As it is all known, there exist three kinds of punctuations in sentences that can represent different pause: comma, semicolon, and colon. Notice that a lot of sentences are acceptable no matter whether commas are used or not. The only difference is the duration of the pause in the position of comma when the sentence is read.

A: 今年 4 月，火山灰对欧洲航空业造成了影响。
B: 今年 4 月，火山灰对欧洲航空业造成了影响。

(Volcano ash impacted on the European aviation industry in this April.)

As an example, sentence A and B are the same except the comma. The position of comma in sentence A can be considered a prosodic phrase boundary (“|”) in sentence B. So if we have collected lots of sentences similar to A, we can automatically generate enormous corresponding artificial sentences as B with one indicated prosodic phrase boundary information.

2.2. Unsupervised adaptation

Given a test data set, it is very time-consuming to obtain a large number of human labeled sentences with similar style, but it is easy to gather a huge data set of unlabelled sentences in similar style. With the idea of section 2.1, enormous artificial sentences (such as the sentence B) with one indicated phrase boundary can be obtained easily to build an initial sentence pool. Next we will demonstrate how to realize the unsupervised adaptation by properly selecting sentences from the initial pool and recover the full labeling of prosodic phrase boundary for these selected sentences. Assume we have constructed a human labeled training set and a development dataset, the baseline system has been trained based on ME model [3] with sliding window smoothing [6] on the training dataset. The unsupervised adaptation includes following steps:

a) An additional CRF model (similar to [9]) is also trained on same training set, which will be used in following sentence selection step.

b) For each artificial sentence in the initial pool, do prediction with the ME system and the additional CRF system respectively.

c) One sentence is selected as candidate sentence and added into sentence set $S$ if it satisfies the following two conditions. The first is that the phrase boundary indicated by comma should be recalled in the prediction result of ME model. The second is that the ME model’s prediction result can be found in the top three best predictions of CRF model.

d) Select sentences randomly in set $S$ and add them into the training set incrementally. More specifically, every 1000 sentences are randomly selected and added in each iteration. Then ME model is retrained and F-score of the development set is recalculated. The 1000 sentences will be accepted if the loss of F-score does not exceed a predefined threshold (set to 0.1 in our experiments). Then it continues to consider another 1000 sentences from the training set just like the previous step. This procedure will be terminated if F-score fails to exceed the best F-score after five iterations.

It should be noticed that normally the development set should also be in similar type as the test set, but with the goal of implementing the entire unsupervised adaptation method, we just use a dataset split from training corpus to make sure we do not need to annotate any new sentences in the adaptation solution.

The new training dataset is closer to testing dataset than the old one. The labeled results of added sentences have high confidence to be correct ensured by the sentence selection step c) with the help of comma information and the additional CRF model. A better prediction performance on testing set is assumed and will be verified by experiments.

3. Syntactic features

3.1. Syntactic structure and prosodic structure

There has been much discussion on the relationship between prosodic structure and syntactic structure [1, 7, and 8]. Some researchers believe that a satisfactory prosodic structure can not be predicted using the results from the syntactic analysis, and the others stressed the importance of syntactic analysis.

According to [8], longer duration of pause often occurs at boundary with higher level syntactic phrases while shorter one occurs at boundary with lower level syntactic phrases. But the influence is not clear because different durations of pauses often share the same level of syntactic phrases while the same durations often have different syntactic levels.

So it is concluded that syntactic structure cannot be directly used for the prosodic structure prediction though it has much information.

3.2. Features of syntactic structure

Each boundary between prosodic words is a candidate of prosodic phrase boundary, and these boundaries are always corresponding to two adjacent leaf nodes in the syntax tree.

For example, a Chinese sentence “法国巴黎出现持续高温天气” (Hot weather continues in Paris, France.) has the lexical analysis result, “法国/NR 巴黎/NR 出现/VV 持续/JJ 高温/NN 天气/NN”. The syntax tree of the Chinese sentence is generated using the Stanford Parser [10] with the parsing model file named chineseFactored.ser.gz in the distributed package. The meaning of syntax tags labeled in the inner corresponds to the definition of the Penn Chinese Treebank 3.0.

Figure 1: Syntax tree of the example sentence.

Given the syntactic structure tree, three features which are considered to be useful for the prosodic phrase prediction are introduced below.
The first is named as path length. Concretely, it is defined as the length of path between two adjacent leaf nodes after the edge of unary parse rule is excluded. Path length of two words should be greater than 4 because the existence of lexical entries. To some extent, this feature can reflect the tightness between two adjacent nodes. According Figure 1, the path length of “法国巴黎” is 4 while it is 6 for “巴黎出现”。It indicates that the former is tighter than the later. And the path length of “持续高温” is also 4 because two unary rules are excluded.

The second is sub-path length. The latest common ancestor node of two adjacent leaf nodes splits the path between the two nodes into two sub paths. The length of each sub path is defined as sub-path length. The sub-path length of “巴黎出现” is <3,3> and that of “出现持续” is <2,4>. This feature can distinguish the cases with the same path length.

Last feature is the sub-paths themselves which is defined as the sequence of first M syntax tags from the latest common ancestor node to the two leaf nodes. M is turned up using a development dataset. Although a bigger value of M indicates more elaborate feature, data sparseness and over-fitting problems may lead to performance degradation. Data sparseness is serious if M is greater than 4. Specially, we only care about the syntax tag of the latest common ancestor node if M equals 1. To the boundary of “出现持续” when M equals 3, the left sub-path is “VP->VV” and right sub-path is “VP->NP->NP”. M is set as 2 in our experiments.

4. Experiments

We verify the two new strategies on our existing Mandarin speech synthesis system, Interphonic 6.0, which is trained only on a corpus from People’s Daily 1993 to 1996. The contribution of unsupervised adaptation is tested by comparison of the prediction performance on web news of 2009 with or without the adaptation strategy. Syntactic features are evaluated both in the matched condition and the mismatched condition.

4.1. Datasets

A set of Chinese sentences was selected randomly from People’s Daily 1993 to 1996. Prosodic word boundary was generated automatically by our lexical analyzer. The accuracy of word segmentation is 96% and 91% for POS tagging. Four annotators were employed to label prosodic phrase boundaries by reading the sentences themselves. 70394 sentences were labeled in total. Through testing, the labeling consistency among the four annotators is 75%, which can be regarded as the upper bound for automatic prosodic phrase prediction.

One tenth of the entire 70394 sentences are randomly selected as the objective testing dataset named TestSet1. Similarly, 10% are divided as the development set named DevSet. The remains are used for training, named TrainSet. To evaluate the subjective performance, one test set consisting of 1500 sentences, which is named TestSet2, is selected randomly from the web news corpus of 2009.

4.2. Evaluation metrics

We utilize the F-score as the evaluation metrics in the experiments, which is defined as follows:

\[
\text{precision} = \frac{\text{number of correctly identified breaks}}{\text{number of identified breaks}}
\]
\[
\text{recall} = \frac{\text{number of correctly identified breaks}}{\text{number of correct breaks in test set}}
\]
\[
F\text{-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

F-score is an objective metric. It is not enough to evaluate the real performance because of the uncertainty in prosody of natural speech [5]. So a subjective metric is also considered in our experiments. The automatic labeled sentences are scored by human. A 5-grade scoring system is applied, and those sentences scored under 3 are considered as unacceptable ones. After human scoring, average score and sentence unacceptable rate (SUR) are computed.

4.3. Experiment results

4.3.1. Experiment configuration

Two open source software are used in the experiments. MaxEnt [11] is used for ME model training. And CRF++ [12] is used for CRF model training. Our baseline system uses ME model which includes many common lexicalized features in a window of [-3, +3]. The cutoff value equals to 7 because our training dataset have tens of thousands of sentences. It has been discussed in [3] about the influence of the cutoff parameters. The same feature templates are used for CRF training.

4.3.2. Details of unsupervised adaptation

We use more than 600M web news corpus from 2005 to 2008 to carry out the adaptation experiments because TestSet2 is from news web corpus of 2009. This corpus is considered more homologous with testing dataset than TrainSet because it is closer in time.

TrainSet is used in the step a) proposed in section 2.2. After steps of b) and c) are employed, more than one million artificial sentences from the initial sentences pool are selected as candidate sentences. Because we want to develop pure unsupervised adaptation method, we employ DevSet as development dataset, though it is not similar to TestSet2.

As result of step d), a set of 29000 sentences are generated, which is named AdaptSet. The F-score changes from 61.82% to 63.80% on DevSet.

4.3.3. Details of syntactic features

It is very easy to integrate the syntactic features in ME model. Abbreviation of extracted syntactic features is given in Table1.

<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL</td>
<td>Total length of Path</td>
</tr>
<tr>
<td>LL</td>
<td>Length of left Sub-path</td>
</tr>
<tr>
<td>RL</td>
<td>Length of right Sub-path</td>
</tr>
<tr>
<td>LP</td>
<td>Left Sub-path</td>
</tr>
<tr>
<td>RP</td>
<td>Right Sub-path</td>
</tr>
</tbody>
</table>

Our template of syntactic features is defined as TL[k], LL[k], RL[k], LP[k], RP[k], LL[k], RL[k] and LP[k], RP[k], in which k is the offset from current boundary. And k has allowed value set of {-1, 0, +1} in our experiment which means the syntactic information of previous boundary and next boundary is also considered. The last two templates are
called jointed template while the others are atomic template. Evaluated in the DevSet, the contribution of these new features is listed in Table 2.

Table 2. Contribution of each syntactic feature template on DevSet

<table>
<thead>
<tr>
<th>ID</th>
<th>Feature Templates</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Baseline</td>
<td>55.21</td>
<td>70.25</td>
<td>61.83</td>
</tr>
<tr>
<td>1</td>
<td>0+TL[k]</td>
<td>56.34</td>
<td>69.22</td>
<td>62.12</td>
</tr>
<tr>
<td>2</td>
<td>0+LL[k]_RL[k]</td>
<td>55.00</td>
<td>71.60</td>
<td>62.21</td>
</tr>
<tr>
<td>3</td>
<td>0+LL[k]_RL[k]</td>
<td>54.19</td>
<td>71.44</td>
<td>61.63</td>
</tr>
<tr>
<td>4</td>
<td>0+LP[k]+RP[k]</td>
<td>56.35</td>
<td>70.55</td>
<td>62.65</td>
</tr>
<tr>
<td>5</td>
<td>0+LP[k]+RP[k]</td>
<td>55.01</td>
<td>70.91</td>
<td>61.96</td>
</tr>
<tr>
<td>6</td>
<td>1+LL[k]+RL[k]</td>
<td>59.42</td>
<td>68.42</td>
<td>63.60</td>
</tr>
<tr>
<td>7</td>
<td>6+LP[k]+RP[k]</td>
<td>58.46</td>
<td>70.33</td>
<td>63.85</td>
</tr>
</tbody>
</table>

Comparison of the results from ID 1 to 5 illustrated in Table 2 shows that atomic templates significantly improve performance while combined templates (ID 3 and 5) has a uncertain effect. Maximum F-score is obtained when all proposed syntactic features generated by atomic templates.

4.3.4. Objective and subjective test results

At last, we evaluate the combination of two proposed strategies on the two test sets. Evaluation results are given in Table 3 and Table 4.

Table 3. Objective evaluation results on TestSet1

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TrainSet</td>
<td>70.25</td>
<td>55.21</td>
<td>61.83</td>
</tr>
<tr>
<td>TrainSet+AdaptSet</td>
<td>67.98</td>
<td>56.68</td>
<td>61.82</td>
</tr>
<tr>
<td>TrainSet+Syntactic Features</td>
<td>68.63</td>
<td>59.17</td>
<td>63.55</td>
</tr>
<tr>
<td>TrainSet+AdaptSet+Syntactic Features</td>
<td>67.31</td>
<td>59.58</td>
<td>63.21</td>
</tr>
</tbody>
</table>

Table 4. Subjective evaluation results on TestSet2

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Average Score</th>
<th>SU(R) (%)</th>
<th>Relative reduction of SUR(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TrainSet</td>
<td>4.14</td>
<td>15.9</td>
<td>-</td>
</tr>
<tr>
<td>TrainSet + AdaptSet</td>
<td>4.19</td>
<td>13.1</td>
<td>17.6</td>
</tr>
<tr>
<td>TrainSet + Syntactic Features</td>
<td>4.24</td>
<td>10.2</td>
<td>35.8</td>
</tr>
<tr>
<td>TrainSet+AdaptSet+Syntactic Features</td>
<td>4.30</td>
<td>8.6</td>
<td>45.0</td>
</tr>
</tbody>
</table>

4.4. Discussion

The purpose of unsupervised adaptation strategy is to make use of training dataset more similar with TestSet2. The improved average score and the reduction of sentence unacceptable rate in Table 4 prove that unsupervised adaptation have achieved the same effect we anticipated. The unsupervised adaptation does not give any gain in the objective evaluation on TestSet1, which can be easily justified because only TrainSet matches TestSet1 while the AdaptSet does not matches with TestSet1.

Consistent improvements are obtained on objective and subjective measure metrics when syntactic features are used, which proves that new syntactic features have a robust effect on prosodic phrase prediction. As discussed in [3], each feature in ME model denotes one classification rule. The bigger feature weights, the more reliable the rule. We augment the discrimination of statistic model by adding “syntactic rules” with automatically trained “rule weights”.

Additionally, it is found that performance of sentence unacceptable rate will degrade about 10% if another Stanford parsing model file named xinhuaFactored.ser.gz is used. This is because chineseFactored.ser.gz is trained on more labeled training data. It means that accuracy and consistency of syntactic structure play an important role in the second strategy. Features from wrong syntactic structure may lead to performance degradation.

5. Conclusions

In this paper, we improve the performance of prosodic phrase prediction by two different strategies. First, with the help of comma in sentences, unsupervised adaptation is proposed to solve the mismatch problem between training and test data sets. Second, three features extracted from syntactic structure are integrated in our maximum entropy model training methods. Sentence unacceptable rate significantly drops from 15.9% to 8.75% on an in-house Mandarin speech synthesis product when the two methods are employed.

In the future work, we will continue to study prediction of prosodic phrase in two ways. As discussed before, the first direction is to refine the syntactic parser itself. Another one is to explore other features of different syntactic structure. It is an interesting work to make full use of rich information in the syntactic structure.

6. References