A CART-Based Hierarchical Stochastic Model for Prosodic Phrasing in Chinese

SHEN Xipeng, XU Bo

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing
{xpshen, xubo}@nlpr.ia.ac.cn

ABSTRACT
A CART-Based stochastic model for prediction of prosodic phrase breaks from input text of Chinese is provided in this work. All the features used in this model are almost obtained automatically. A novel and efficient algorithm—LLW algorithm is proposed here. Experiments demonstrate a high success rate of prosodic phrase breaks prediction from input sentences with little syntactic information (81% success rate, 6.1% false rate).

1. Introduction

One of the important problems in speech synthesis systems is the automatic generation of prosodic information from text input. A good prosody model plays an important role in both naturalness and intelligibility of synthesized speech, and also is much useful for speech understanding. Prosody elements include: prosodic phrase break, accent, intonation, duration and so on. In English there have been many works on deriving prosodic models of spoken language for TTS[1][2][3]. Although, in Chinese, such works are not so many as in English, there have been also some studies[4][5][6][7] in the last two decades. There are two chief approaches in these studies. One is rule-based approach[4]. The derivation of rules, however is laborious and time wasting. Moreover, it is difficult to obtain the complete rules to describe the prosody diversity because the interactive affection between the various linguistic features and the phonological characteristics. The other approach is through data-driven model such as neural networks. But, the network sometimes will be trapped in a local minimum of the error function, thus arriving at an unacceptable solution when a better one exists[7]. Chinese is quite different from many western languages in various structural features [8]. Lack of an appropriate prosodic model that describes the prosodic phrase structure of spoken language is one of the fundamental problems in TTS system[5]. Based on analysis of the features of this special language, we provide a CART-based hierarchical stochastic model with a novel algorithm, LLW(Linear-Logical-Weight) algorithm, for prediction of prosodic breaks in Chinese.

Prosodic breaks are relative both with input text information and acoustic cues. But here we focus on synthesis and are concerned only with the relationship between text input and the prosodic phrase breaks. We make use of only those features such as POS, position of words, cue words, punctuation and simple syntactic structure, which can be automatically obtained without complicated syntactic analysis. It is therefor very practical for speech synthesis systems. A novel algorithm, called LLW (Linear-Log-Weighted) algorithm, is provided and used in the model. This algorithm is much simpler than other stochastic algorithms for prediction such as Viterbi algorithm and is effective to such tasks as prediction of prosodic prosody breaks. The final experimental results demonstrate a high success rate for prosodic major and minor breaks (81% correct prediction with 6.1% false prediction).

In the following parts, Section 2 gives the formalism of the stochastic model. In Section 3, we introduce the implementation of the model. In Section 4, the experimental results of prediction are given. Finally, we discuss the results in Section 5.

2. The Stochastic Model

2.1 Hierarchic Prosody Phrase Structure

Prosody phrase structure can be represented by different phrase break markers. In TOBI labeling system[9], prosodic phrase breaks are classified into 7 levels. In our work, we classify prosodic breaks into only 3 levels: minor breaks, major breaks and inter-sentence breaks. The relative size of these breaks is from little to big and large. The following is an example:

某些 || 商业银行 | 还 | 采取 | 故意 | 截留 | 借方 | 余额的 | 手段 | | | 某些 | 商业银行 | 还 | 采取 | 故意 | 截留 | 借方 | 余额的 | 手段 | | | 某些 | 商业银行 | 还 | 采取 | 故意 | 截留 | 借方 | 余额的 | 手段 |

"|" denotes "minor break", " || " denotes "major break", " ||| " denotes "inter-sentence break".

2.2 the Hierarchic Stochastic Model

The hierarchic prosodic phrase structure includes major phrases and minor phrases. In our prediction of the prosodic breaks, we first predict the general breaks, which include all the three kinds of breaks. Secondly, based on the results obtained from the first step, we predict the specific kind of each break. The two steps are very similar. Both are to predict one level of units from the sub-units. The basic model [10] is showed as follows:

\[
P(M_i \mid S) = P(m_{i1}, \ldots, m_{i_n} \mid S) = P(m_{i1}, \ldots, m_{i_n} \mid S, n_i)P(n_i \mid S) = P(n_i \mid S)P(m_{i1} \mid S) \prod_{j=2}^{n_i} P(m_{ij} \mid S, m_{i1}, \ldots, m_{(j-1)})
\]

\(S\) denotes the given orthographic transcription of an input sentence, \(M_i\) is a unit composed of a sequence of sub-units \(m_{ij}(j = 1, 2, \ldots, n_i)\).

\(^1\)The research work described in this paper was supported by the National Key Fundamental Research Program (the 973 Program) of China under the grant G1998030504, the National Natural Science Foundation of China under the grant 69835030 and also the National 863 Hi-Tech Program under the grant 863-306-ZT03-02-2.
In this model, \( P(m_{ij} | S, m_{ij(j-1)}) \) is difficult to get. Generally, it is simplified to be \( P(m_{ij} | S, m_{ij(j-1)}) \). However, in our experiments, results acquired from the above model are not better than results from our model in which \( P(m_{ij} | S, m_{ij(j-1)}) \) is assumed to be \( P(m_{ij} | S) \) with some compensatory rules. In our model, a novel algorithm, LLW algorithm, other than Viterbi algorithm is used and satisfactory results are acquired. Our model is as following.

We assume:
- \( S \): an input sentence which is composed of a sequence including \( L \) words \( \{w_1, w_2, \cdots, w_L\} \).
- \( H \): one general prosodic phrase sequence \( (P_1, P_2, \cdots, P_n) \) aligned to \( S \) after being phrased in one way.
- \( H_i \): a prosodic phrase composed of a sequence of some adjacent words \( \{w_{i1}, w_{i2}, \cdots, w_{iL}\} \) in \( S \).
- \( P(b_i | S) \): the probability of word \( w_i \) to have a general break following it.
- \( \hat{P}(b_i | S) \): the probability of word \( w_i \) not to have a general break following it, which is equal to \( (1 - P(b_i | S)) \).

We sort \( P(b_i | S) \) \( (i = 1,2,\cdots,L) \) from maximum to minimum into a sequence:
\[
(P(b_{m1} | S), P(b_{m2} | S), \cdots, P(b_{mL} | S))
\]

The chief judgment criterion is as following:

\[
\hat{n}_{op} = \text{MAX} \{ r(n_1) \}
\]
\[
= \text{MAX} \{ \alpha \log P(n_i | S)
\]
\[
+ \log \prod_{j=1}^{n} P(b_{mj} | S) \times \prod_{k=n+1}^{L} \hat{P}(b_{mj} | S) \}
\]
\[
+ \beta \log (n_i) \}
\]

Where
\[
\alpha + \beta = 1, \quad \alpha = (2 + L^2)/2L^2
\]

\( \hat{n}_{op} \) is the final optional selection of \( n_i \), \( \alpha, \beta \) are two weight parameters ,whose sum is 1. The reason to let \( a \) equal to \((2 + L^2)/2L^2\) is to keep \( a \) increase if \( L \) increases and meanwhile keep \( a + \beta = 1 \) and \( a > 1/2 \). The results of our experiments show that this format is effective.

Our previous criterion does not have this part: \( \beta \log (n_i) \). In experimental results, \( \hat{n}_{op} \) is usually less than the real number of breaks \( (n_i) \) in a sentence. The reason is analyzed below.

We assume:
\[
\hat{r}(n_1) = \lambda_1 \log P(n_1 | S) + \lambda_2.
\]
\[
\log \prod_{j=1}^{n} P(b_{mj} | S) \times \prod_{k=n+1}^{L} \hat{P}(b_{mj} | S) \}
\]
\[
(\lambda_1 + \lambda_2 = 1)
\]

Suppose \( n_1 < n_2 \) and \( P(n_1 | S) < P(n_2 | S) \). Since most \( P(b | S) \) is smaller than 0.5 in fact, \( \hat{r}(n_1) \) is often bigger than \( \hat{r}(n_2) \) although \( n_2 \) may be much better than \( n_1 \). Therefore, we add \( \beta \log (n_i) \) to the criterion to compensate for such loss. In Section 4, there is a comparison of two results got from the two criteria, one of which has \( \beta \log (n_i) \) and the other doesn’t.

Once \( \hat{n}_{op} \) is obtained, the final general break locations then are determined, each of which follows one of these words:
\[
w_{m1}, w_{m2}, \cdots, w_{m\hat{n}_{op}}
\]

Above is our novel algorithm, called LLW (Linear-Logical-Weighted) algorithm. As to the effect of the previous unit \( b_{mi(j-1)} \) on unit \( b_{mj} \), we add such a rule as follows:

If the number of characters between \( b_{mi(j-1)} \) and \( b_{mj} \) equals to 1, then
\[
P(b_{mj} | S) = \kappa \cdot P(b_{mj} | S) (\kappa = 0.1)
\]

The reason to adopt this rule is that from statistic results we can see that the probability to have two breaks with only one character between them is very low. It can be seen that the complexity of our model is much less than other models using Viterbi algorithm. However, the results got from our model are satisfactory. Perhaps this is because the effect of one unit on the next unit is much more complicated than what have been considered. Thus, simple analysis of the effect is not very helpful to improve the success rate of prosodic phrase prediction.

Above is our model to predict the general breaks. The procedure to predict the specific kind of breaks is much similar to the above procedure. The models are much similar. The only difference is that in the specific break prediction, the unit is not the general phrase but is the major prosodic phrase; and each general prosodic phrase is one sub-unit. The implementation of our model for prosodic phrase prediction is showed in the next section.

3. Implementation of the Stochastic Model

To implement the stochastic model, four parameters must be determined. We developed four CARTs (Classification and Regression Tree) to predict the four parameters. In the following, we first outline the structure of our model and then introduce the text processing work.

3.1 Structure of the CART-Based Hierarchic Model

Based on the LLW algorithm, we construct our Hierarchic Stochastic Model. Through this model, we get a high rate of success for predicting major and minor phrase boundaries from text with less computation. As shown in Section 2, in this model, there are four categories of parameters to be predicted:
\[
P(n_1 | S), P(N_1 | S), P(b_i | S), P(B_i | S).
\]

- \( P(n_i | S) \): the probability of a sentence \( S \) having \( n_i \) general prosodic breaks;
- \( P(N_i | S) \): the probability of a sentence \( S \) having \( N_i \) major prosodic breaks;
- \( P(b_i | S) \): the probability of a word \( w_i \) having a general break following it;
- \( P(B_i | S) \): the probability for a general break following a word \( w_i \) to be a major break;

Four CARTs are developed to predict these four categories of
parameters from almost automatically labeled training data and use LLW algorithm to combine them to perform the prediction task. The structure of our model is showed in figure 1. C1, C2, C3, C4 denote the four CARTs.

3.2 Extracting Features from Text

In our model, we used only five categories of features which all can be obtained automatically or partly automatically from input sentences easily what is practical for speech synthesis system. The five categories of features are: POS, positions of words in a sentence, some cue words, punctuation and a simple syntactic structure.

Our corpus includes 256 sentences which are designed and recorded especially for Mandarin synthesis study. All of them are read professionally. These sentences include most of the various speech phenomena in Mandarin. There are altogether 509 sub-sentences, 3414 words and 6459 characters except punctuation. We used an automatic tool—Segword to segment characters into words. Then by the help of another automatic tool—CPT, we labeled POS with each word and corrected them by hand. There are 14 kinds of POS in our system, which are showed in Table 1. The position of each word in a sentence is classified into three categories: in the last quarter of the sentence, before the predicate of the principle clause. The reason why we use this structure information is that there is often a major break between the subject-clause and the predicate of the principle clause. These are the features that are extracted and used in our model. The development of the four CARTs and experimental results are introduced in the next section.

4. Experimental results

4.1 Development of CARTs

CART (Classification and Regression Tree) is a powerful tool for classification based on training data. Decision tree construction is equivalent to successive refinement of equivalence classes driven by answers to questions. Our questions are about the five kinds of features referred above. The structure of CART-1 is showed in figure 2 (the number in each circle is the question number selected in this dot; the numbers in each rectangle box are the numbers of those words having breaks following them and the total words in this class). Through this decision tree, the parameter \( P(b_i | S) \) is obtained. The questions asked in this decision tree are listed below.

1. Does a comma, colon or period follow this word?
2. Is this word a function followed by a substantive or a substantive followed by a function?
3. Is this word a noun followed by a verb, a preposition, a conjunction or an adverb?
4. Is this word “的” not followed by a noun?
5. Is this a substantive not followed by an auxiliary word, a modal word or word “的”?
6. Is this an auxiliary word or a modal word followed by a substantive?
7. Is a caesura sign following this word?
8. Is this word in the last quarter of the sub-sentence?
9. Is this word in the last 4 words of the sentence?

\( P(n_i | S) \) is estimated by use of CART-2. We assume that \( P(n_i | L) \) can be represented by \( q(n_i | L) \), the frequency of a sentence with \( L \) characters having \( n_i \) breaks. We need to classify \( L \) into several classes which are a series of arranges \((L_{id}, L_{iu})\). In each class, the difference of \( q(n_i | L) \) between two sentences of different length is small, while the frequencies of every two classes are much more different. The decision tree is showed in figure 3 (what is in each circle is the question about the arrange of the length of each sentence; the number in each rectangle box is the number of each class).

<table>
<thead>
<tr>
<th>1-level POS</th>
<th>2-level POS</th>
<th>Meaning</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substantive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>General Noun</td>
<td>苹果, 工人</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>Numeral Words</td>
<td>三, 多少</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Qualifier Words</td>
<td>个, 双</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Adverb</td>
<td>纷纷, 认真</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>Verb</td>
<td>打仗, 降落</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Adjective</td>
<td>好看, 新新</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Idiom</td>
<td>就是说</td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>Proper Noun</td>
<td>毛泽东</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>Pronoun</td>
<td>他, 我们</td>
<td></td>
</tr>
<tr>
<td>Function Words</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Conjunction</td>
<td>如果, 若要</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Preposition</td>
<td>在, 于</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>Auxiliary Words</td>
<td>看, 过</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>Modal Words</td>
<td>吗, 吧</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. POS of Chinese words in this work
4.2 Prediction Results and Comparisons

To evaluate our model, we take 19 sentences out of the corpus introduced in Section 3 but are not included in the training data. We got 81% correct rate, 6.1% false rate. The confusion matrix for the model using LLW algorithm is showed in table 2. And the confusion matrix use the criterion without \(\beta \log(n_i/v)\) is showed in table 3. The decrease of the break lost-rate is 9.5%, while the false rate only increases 1.9%. Therefor, adding \(\beta \log(n_i/v)\) into the algorithm is helpful to improve the success rate of prediction.

<table>
<thead>
<tr>
<th>Predicted Prosodic Phrase Breaks</th>
<th>Actual Prosodic Phrase Breaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>major</td>
<td>27</td>
</tr>
<tr>
<td>minor</td>
<td>2</td>
</tr>
<tr>
<td>no</td>
<td>2</td>
</tr>
</tbody>
</table>

5. Discussion

In summary, based upon the special features of Chinese, we constructed a hierarchical stochastic model with four decision trees to predict the parameters in the model and a novel algorithm, LLW algorithm, to combine them to perform the prediction of prosodic breaks. The experimental results demonstrate a high rate of success (81% correct rate and 6.1% false rate). Since the features used in this model are almost automatically obtained, this model can be easily used in a speech synthesis system and will be helpful to improve the naturalness and intelligibility of the synthesized speech.

The prosodic phrase break prediction model can be improved in many aspects, for example, to take more consideration to the syntactic structure information. After analysis of many experiments, we found that part of the errors is easy to be avoided with more complete syntactic structure.

This model can also improve the performance of a discourse segmentation system by serving as part of the language model. Although this model is constructed in Chinese, it is applicable for any other language.

6. Reference