International Symposium on Chinese Spoken
Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

PROSODIC PHRASE DETECTION FOR CHINESE TTS USING CART AND STATISTICAL MODEL

Minghui DONG, Kim-Teng LUA

School of Computing, National University of Singapore
S15, 1 Science Drive 2, Singapore 117543

{dongming, luakt}@comp.nus.edu.sg

ABSTRACT

Determination of prosodic phrase break from text is one of the important problems in generating good prosody for Chinese text-to-speech system. In this paper, we propose a statistical approach for detecting prosodic phrase breaks. Part-of-speech sequence information is used as the primary information. The history of the previous breaks is considered as constraint in this work. The probabilities are calculated using CART approach. During the prediction process, the breaks are dynamically determined. Viterbi algorithm is applied to find the best break sequence. We achieved a result of recall of about 89% and precision of about 85 % for our testing data.

1. INTRODUCTION

When speaking, people tend to group words into small prosodic unit groups, hence insert breaks within sentence. Grouping words into phrases helps speaker to speak easier and helps listeners to understand sentence better. Sometimes it solves ambiguities that could exist in written form.

Punctuation marks are explicit symbol in sentence to indicate breaks. However, there are examples of long sentences without a break. In such case, more breaks will be added within a long sentence when we utter a sentence. A long sentence is therefore cut into short prosodic phrases.

Prosodic phrase boundary can be identified by some pause, pitch change or duration change of boundary syllables in speech. To realize all these effects in synthetic speech, phrase boundary need to be determined in a TTS system. Phrase boundary is realized by inserting pauses, changing the pitch contour and lengthening duration of the boundary syllables. It is common that words that represent a meaningful concept are grouped into phrases.

Many ways have been proposed for the determination of the breaks or units from the text input. Typical approaches include Rule-based and corpus-based. Rule based approaches were first used for locating phrase boundary. Typical corpus based approaches include classification and regression tree (CART)[1], neural networks[2], and hidden Markov models (HMM)[3].

For Chinese, there are a few approaches for phrase determination. Chou et al [5] proposed an approach to first form a lattice to include the possible phrase grouping, then find the best path from the lattice according to the frequencies of POS grouping. Chen et al [6] proposed an approach based on inductive learning algorithm and extension matrix theory. POS sequence and syntactic structure are used in the phrase model. POS type and length of constituents in terms of the number of characters and words are used as features for prediction. A success rate of 93% achieved on 371 training sentences and 188 testing sentences.

Generally, some of the above researches for Chinese phrase break reported good prediction results. However, there are uncertainties in the above approaches. (1) Prediction of prosodic phrase on the base of syntactic structure can achieved good prediction result. However, the accuracy of automatic syntactic structure parsing was not reported. Considering the errors occurred in syntactic analysis, the accuracy of the prosodic phrase break will be lower than the reported value. (2) Most of them used small number of sentences in the experiments. The performance is unknown for bigger corpus.

In this research, we propose a statistical method for phrase break prediction. We use CART to calculate the probabilities.

2. CORPUS

The text script of the corpus in this research was initially from Chinese newspapers. To have a good coverage of different types of text, we selected different types of articles. 40 articles are selected for recording of the speech. A male speaker read each the article in a neutral manner with a normal speed in laboratory environment. Totally 2788 utterances and around 42000 syllables are collected.

The script of the corpus is processed to meet the requirement of training. The text is first normalized. The symbols and numbers are all converted into Chinese text. All non-Chinese symbols and numbers in the raw Chinese text are converted into pure Chinese text. The text is segmented into word sequence by a word segmentation program. Next the part-of-speech tagging program tags each word with a part-of-speech category.

In this research, Chinese sentence is decomposed into the following prosodic units. Syllable is the smallest unit. Syllables are combined together to form words. Words are further combined to form prosodic words. Prosodic words form longer unit, prosodic phrase. And one or more prosodic phrases form sentence. We labeled three levels of prosodic break in the corpus. The three labels are prosodic word break, minor phrase break and major phrase break.

Our corpus includes 2788 sentences, which includes 22300 words, 4140 minor phrase and 3352 major phrases.

3. THE DEPENDENCY MODEL

The problem of phrase break determination is to determine each word break is a phrase break or not. The main cue for break determination is the POS of each word in a sentence. To determine phrase break type of a word break, the POS types of the words around the word break are usually considered.

3.1 Distribution Of Phrase Length

A long phrase will be cut into shorter ones, while some short phrases will be combined to a longer one. We has a statistics of the lengths of minor phrases in our corpus. Figure 1 shows the distribution of length (number of syllables) of minor phrase in the corpus. From the figure we can see most of the minor phrases have a length from 3 to 11. That means the distance between two phrase breaks is not too long or too short. This also indicates a dependency between the breaks within a sentence.

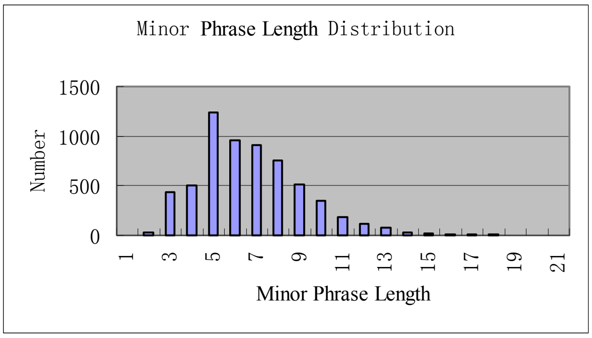

Figure 1. Distribution of length of minor phrases

In this approach, we assume the break between two words is dependent on the previous break sequence in this sentence.

3.2 The Dependency Model

The probabilistic approach to prediction of minor breaks uses a stochastic model $P(a_1^n \mid Y_1^n)$ that represents the conditional dependence of the sequence of the breaks the sequence of feature vectors $Y_1^n = \{Y_1, Y_2, \ldots, Y_n\}$. a_i is the break of the syllable i, and Y_i is a vector of features that are relevant to the break. Using the chain rule:

$$P(a_1^n \mid Y_1^n) = p(a_1 \mid Y_1^n) \prod_{i=2}^{n} p(a_i \mid a_1^{i-1}, Y_1^n)$$

Under the mth-order Markov assumption, it is assumed that a_i is only dependent on a_{i-1} and Y_i which gives:

$$P(a_1^n \mid Y_1^n) = p(a_1 \mid Y_1) \prod_{i=2}^{n} p(a_i \mid a_{i-1}^{i-m-1}, Y_i)$$

3.3 Calculation Of Probabilities

To calculate the $p(a_i \mid a_{i-1}^{i-m-1}, Y_i)$, CART[4] approach is applied. a_{i-1}^{i-m-1} and Y_i are used as input features of the tree and $p(a_i \mid a_{i-1}^{i-m-1}, Y_i)$ is the output value of the tree. a_{i-1}^{i-m-1} means the previous m break types, and Y_i means POS's of a windows of word sequence around a break.

Each terminal node assigns not only the most likely classification but also a discrete distribution, which represents the conditional probabilities for being a break or not. That is, we have the values of $p(a_i = l \mid a_{i-1}^{i-m-1}, Y_i)$ from the tree, where l=0 for no break and l=1 for being a break.

The calculation of probability can be illustrated as in the following figure.

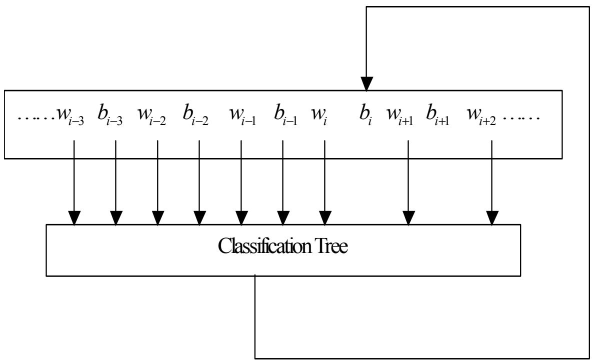

Figure 2. Calculation of probability using decision tree

3.4 Searching Algorithm

The determination of breaks needs a dynamic programming process to find the best one. The Viterbi search algorithm is chosen for its ability to search a graph in a time-synchronous manner and its excellent pruning characteristics. Viterbi search algorithm works as similar as that for prosodic word prediction.

The algorithm can be described as following.

1. *Initial state $P_0 = 1$*
2. *Search*
   for i = 1 to n, do
   a. *for k = 1 to number of paths*
      for choice j = 0 to 1 do
      $P_{i+1}(k) = P_i (k) * p(I,j)$
   b. *sort P_{i+1}*
   c. *keep the first m items*
3. *Back trace*
4. *Output*

4. EXPERIMENT RESULTS

As there are not enough data in our corpus for major phrase detection, in this research, we only work on minor phrase detection. Two methods have been tested to evaluate our dependency model for minor phrase break prediction. Method 1 only uses POS sequence, while method 2 is our dependency model, which uses POS sequence and dependency information. In both methods, we use 80% of our data as training data and the rest 20% as testing data.

Method 1. CART approach with POS Sequence: In the first approach, we only use POS sequence to predict breaks between words. A window consists of 2n surrounding words (n words before and n words after) around a word break. Features for predicting the break includes the POS of n words before and n words after the break. For the cases that there are no enough words to fill the window, a NULL value is assigned as POS. In this experiment, n is given a value from 1 to 8.

Method 2. Dependency model with POS sequence and previous break sequence: In the second approach, a window of 2n words is also selected together with n-1 breaks before the words. Therefore the features for the prediction include POS of n words before the break, n words after the break and n break types before he break. The n value varies from 1 to 8 too.

The results of prediction are as in the following table. We calculated the precision and recall values for the prediction. The recall and precision are calculated as the following:

$$\text{Recall} = N_{correct} / N_{actual}$$
$$\text{Precision} = N_{correct} / N_{predicted}$$

Where $N_{correct}$ means the number of correctly predicted breaks, N_{actual} means the number of labeled breaks in corpus, $N_{predicted}$ means the number of predicted breaks.

The results show in the first approach, we achieved a recall of around 88% and highest precision of 76 % when n = 8. The second approach has a better performance. The highest recall is 89% and the highest prediction value is 85.7% when n = 8.

n	Recall	Precision
1	87.6%	70.9%
2	87.9%	72.2%
3	88.9%	74.2%
4	88.2%	75.4%
5	88.4%	75.5%
6	88.1%	76.2%
7	88.1%	76.4%
8	88.3%	76.0%

Table 1. Result of break prediction using CART and POS sequence

N	Recall	Precision
1	87.9%	77.1%
2	88.0%	82.2%
3	88.9%	84.9%
4	89.2%	85.0%
5	89.7%	84.5%
6	89.1%	85.4%
7	89.2%	85.5%
8	89.1%	85.7%

Table 2. Result of break prediction using dependency model

We further compare the precision and recall values obtained using the two methods in Figure 3 and Figure 4.

For precision, we found the following facts: method 1 achieves the highest value when n = 5, while method 2 achieves the highest value when n = 3. However, method 2 obtained a higher precision than method 1. This shows that dependency model is better than the model without dependency information. There is a decrease in precision for both methods. This may be caused by noise. It can be interpreted as words more than 5 words away are less related with the break type, hence can be thought as noise.

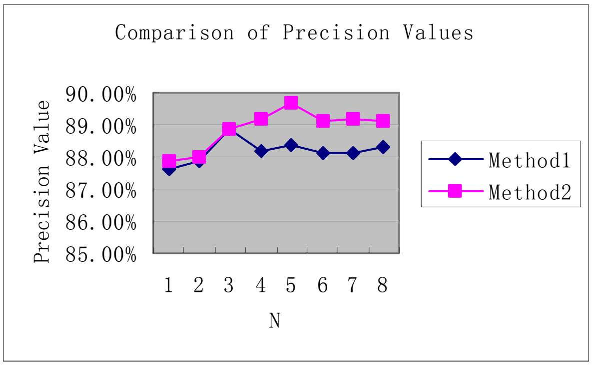

Figure 3. Comparison of precision values using the two methods

Observing the figure of recall, we found that method 2 obtained higher values than method 1. The recall values for method 1 slightly increase with n, while the recall values of method 2 increase quickly when n < 3. This shows method2 works better than method 1.

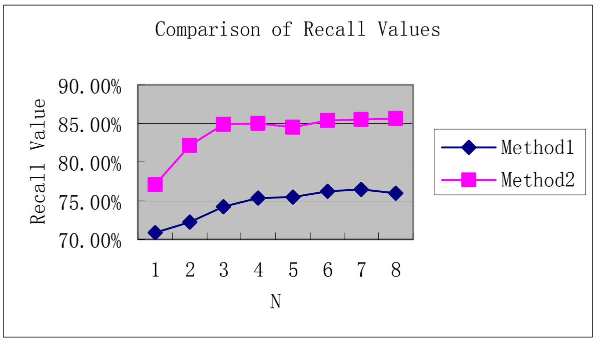

Figure 4. Comparison of precision values using the two methods

Cosidering the precision and recall value, we think our dependency model achieves best performance when n = 5.

5. CONCLUSION

We proposed a statistical approach for detecting prosodic phrase breaks. Part-of-speech sequence information is used as the primary information. The history of the previous breaks is considered as constraint. The probabilities are calculated using CART approach. We achieved a result of recall of about 89% and precision of about 85 % for our testing data. We compared our model with the model that does not use dependency information. The result shows the proposed model has a good performance in Chinese prosodic phrasing.

REFERENCES

[1] Wang, M.Q. and Hirschberg, J. Automatic classifiction of intonational phrase boundaries. Computer Speech and Language, 6: 175-196,1992.

[2] Fujio, S., Y. Sagisaki, and N. Higuchi. Prediction of Major Phrase Boundary Location and Pause Insertion Using a Stochastic Context-free Grammar, in Computing Prosody (Yoshinori Sagisaka, Nick Campbell, Norio Higuchi, editors), pp.271-284. 1996.

[3] Taylor, P. and A.W. Blank. Assigning Phrase Breaks From Part-of-speech Sequences. Computer Speech and Language, 12:99-117,1998.

[4] Breiman, L.; Friedman, J.; Olshen, R. and Stone, C. Classification and Regression Trees. Wadsworth & Brooks, Pacific Grove, CA., 1984.

[5] Chou, F. C.; Tseng C. Y. and Lee, L. S. Automatic Generation of Prosodic Structure for High Quality Mandarin Speech Synthesis, in proceeding of ICASSP, pp. 1624-1627, 1997.

[6] Chen, Weijun; Lin, Fuzong; Li, Jianmin; Zhang, Bo. A New Prosodic Phrasing Model for Chinese TTS Systems, NLPRS 2001, Taipei, 2001.