SLIDING WINDOW SMOOTHING FOR MAXIMUM ENTROPY BASED INTONATIONAL PHRASE PREDICTION IN CHINESE

Jian-Feng Li, Guo-Ping Hu, Ren-Hua Wang, Li-Rong Dai

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

ABSTRACT

In Chinese TTS (Text-To-Speech) system, Intonational phrase prediction has great influence on naturalness of synthesized speech. Different kinds of statistic models have been applied to this domain, and achieved good performance. In this paper, we first build a maximum entropy model to yield the probability of each word boundary to be an intonational phrase break, and then a sliding window smoothing algorithm is proposed, in which the length distribution curve of intonational phrase acts as the sliding window. The maximum entropy model and the distribution curve are trained from 19,000 sentences and tested on a test set of 1,000 sentences. Experiment results shows that, the sliding window smoothing algorithm makes an improvement of 5.3% in terms of F-Score, 10.0% in terms of average score, and 55.6% in terms of unacceptable rate. From the results, we draw the conclusion that the length distribution information is of great usefulness for intonational phrase break prediction, and the sliding window smoothing method is quite effective to improve the performance significantly.

1. INTRODUCTION

In Chinese TTS systems, a widely used hierarchical prosody structure system consists of syllable, prosody word, intermediate phrase, intonational phrase and breath group[1] . For convenience sake, the 5 hierarchical layers are denoted by L0, L1, L2, L3 and L4. Among them, intonational phrase plays an important role on affecting the naturalness of synthesized speech in our system. In this paper, we discuss about intonational phrase break prediction, which is to split a sentence into several intonational phrases, and also corresponds to decide whether a word boundary is an L3 break.

Recently, various kinds of statistic models were applied to this field, including CART[1] [2] (Classification And Regression Tree), Markov Model[3] , Memory Based Learning[4] , Maximum Entropy Model[5] and Artificial Neural Networks. In these models, similar information was exploited, including POS (Part-Of-Speech), syllable number and the word itself in local context. The theory of machine learning[6] tells us that significant improvements can be achieved if new valuable properties are included.

Constrained by physiology, people is inclined to make an obvious pause (L3 break) after a certain number of syllables. Hence, we assume the length distribution of intonational phrase subjects to some statistic laws. We try to investigate the length distribution in this paper and apply it to intonational phrase break prediction.

According to [5] , maximum entropy model makes an improvement of 9.4% on F-Score over decision tree, the mainstream method. In this paper, performance of maximum entropy model is regarded as the baseline, and the contribution of sliding window smoothing method is investigated.

The remainder of the paper is organized as following. In section 2, maximum entropy model based intonational phrase break prediction is simply introduced. In section 3, sliding window smoothing method is discussed in detail. Experiments are carried out in section 4, and conclusions are drawn in section 5.

2. MAXIMUM ENTROPY BASED INTONATIONAL PHRASE BREAK PREDICTION

2.1. Maximum entropy model

Maximum entropy model is a probability model, which estimates probabilities based on the principle of making as few assumptions as possible, other than the constraints imposed. A constraint can be expressed by a binary feature function $f_i(x,y)$, in which, x denotes the context, and y denotes the outcome. If some constraint is satisfied, $f_i(x,y)$ is set to 1, otherwise 0.

A maximum entropy model can be represented as[7] :

$$p(y \mid x) = \frac{1}{Z(x)} \exp(\sum_i \lambda_i f_i(x,y)) \qquad (1)$$

0-7803-8874-7/05/$20.00 ©2005 IEEE

In which, λ_i is the weight of feature $f_i(x,y)$, which can be estimated by IIS algorithms[7] . Z(x) is the normalization factor. For more information about maximum entropy model, please refer to [7] .

2.2. Feature selection

Intonational phrase break prediction could be regarded as a classification problem, where each word boundary needs to be decided by a classifier whether it is a L3 break. In this paper, maximum entropy model acts as the classifier, which will give the probability that a word boundary is a phrase break. We refer to the probability by L3 probability below.

Following [5] , 67 feature templates were constructed manually from context information, including the neighbor words themselves and their POSs and syllable numbers. The neighbor words are restricted to three at both sides. After that, we use CCFS (Count Cutoff Feature Selection) to extract features from training corpus. Please refer [5] for more details.

3. SLIDING WINDOW SMOOTHING

The maximum entropy model is used to produce L3 probability of each word boundary, and the probabilities are estimated independently. This leads to the fact that the interaction between the adjacent L3 breaks is ignored. In fact, it is unlikely that two L3 breaks occur too near to or too far from each other, and the length distribution of intonational phrase is subjects to some statistic law. If the length distribution information is integrated in the predicting system, the performance may be further improved.

It is not easy to integrate the length distribution information into the maximum entropy model discussed in section 2. So we introduce a post-processing module to implement it. This module is called smoothing module, for its effect is to make the result more smoothing (neither too short nor too long intonational phrases).

3.1. Length distribution of intonational phrase

The length distribution of intonational phrase can be easily estimated from a training corpus with human labeled L3 breaks. In our experiments, 19,000 sentences were labeled as training corpus (discussed in section 4), and the length distribution curve is estimated from it. Figure 1 shows the curve. From Figure 1 we can see most intonational phrases are of length of 4 to 9 Chinese characters. The probability of phrases with a length larger than 11 characters is rather small.

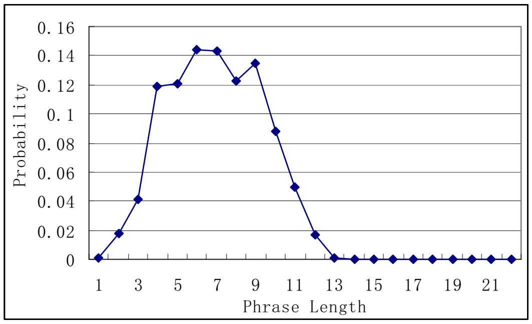

Figure 1: Length distribution of intonational phrase

3.2. Sliding window smoothing algorithm

The length distribution curve of intonational phrase is utilized as the sliding window to smooth the result of maximum entropy model. The detail algorithm is expatiated in Algorithm 1.

```
Algorithm 1:
1)  Set S to be the head of the sentence;
2)  Move the sliding window to S (let the left
    side of the sliding window start at S);
3)  For each word boundary B behind S,
    compute its L3 Confidence, and select the
    boundary M with the largest L3 Confidence
    as the next L3 break;
4)  Set S = M;
5)  If S arrives at the tail of the sentence, stop;
    otherwise, go to step 2).
```

At step 3) in Algorithm 1, the L3 Confidence is defined as:

$$Conf = P_{ME}(B) * P_{Win}(B\text{-}S) \qquad (2)$$

In which, $P_{ME}(B)$ is the L3 probability produced by maximum entropy model. *B-S* is the distance from the current word boundary B to the start S of the left side of sliding window. $P_{Win}(B\text{-}S)$ is the probability of intonational phrase with a length of *B-S* and can be obtained from Figure 1.

The sliding window smoothing process can be illustrated intuitionally by Figure 2. The histogram stands for the L3 probabilities of each word boundary, and the curves denote the sliding windows. At first, the sliding window stands at the head of the sentence. By computing the L3 confidence of each boundary after the head of sentence, the word boundary at position 6 is selected as the first L3 break. Then, the sliding window moves to position 6, and L3 confidences of boundaries after position 6 are re-computed, and position 12 is accepted as the second L3 break. This process continues, till the tail of the sentence is selected as the last L3 break.

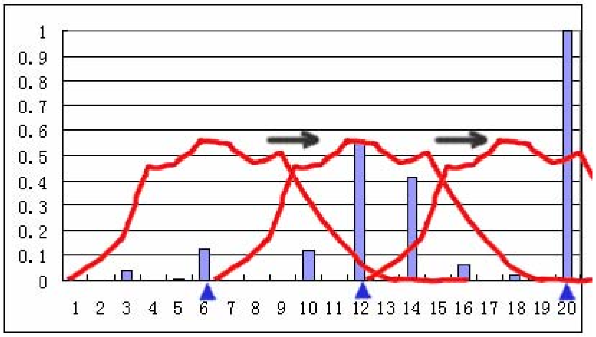

Figure 2: Sliding window smoothing process (FSW)

In Algorithm 1, the sliding window moves from left to right, so it is called FSW (Forward Sliding Window). We can also start from the end of the sentence and moves the sliding window from right to left, which is called BSW (Backward Sliding Window). Note that, the shape of BSW is different from that of FSW. BSW is obtained by turn FSW from left to right, so the start point of BSW is at the right side. Following Algorithm 1, we easily get the smoothing algorithm for BSW, in which the different part from Algorithm 1 is printed in bold.

Algorithm 2:

1) Set S to be the **tail** of the sentence;
2) Move the sliding window to S (let the **right** side of the sliding window start at S);
3) For each word boundary B **before** S, compute its L3 Confidence, and select the boundary M with the largest L3 Confidence as the next L3 break;
4) Set S = M;
5) If S arrives at the **head** of the sentence, stop; otherwise, go to step 2).

4. EXPERIMENTS

4.1. Experiment settings

4.1.1. Corpus

20,000 sentences were random selected from People's Daily, and used in the experiments. Word segmentation, POS tagging and person name recognition were carried out by a preprocessing program. The accuracy of word segmentation is 96% and the accuracy of POS tagging is 91%.

Tags need be labeled at each word boundary to indicate L3 breaks or non L3 breaks. Lack of the corresponding speech, the annotators labeled word boundaries by reading the sentences themselves. As it is known, different people might label the same sentence differently. Through testing, the labeling consistency among the four annotators was 75%, which is the upper limit for automatic prediction.

All of the sentences were divided into two parts, 1000 sentence for testing and the others for training.

4.1.2. Evaluation metrics

We utilize the F-Score as the evaluation metric in the experiments, which is defined as follows:

$$precision = \frac{number\ of\ correctly\ identified\ breaks}{number\ of\ identified\ breaks}$$

$$recall = \frac{number\ of\ correctly\ identified\ breaks}{number\ of\ correct\ breaks\ in\ test\ set} \qquad (3)$$

$$FScore = \frac{2 \times precision \times recall}{precision + recall}$$

F-Score is an objective metric. A subjective metric is also adopted in our experiments, which have the automatic labeled sentences 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 unacceptable rate are computed. Average score is the arithmetic average of all sentences, and unacceptable rate is the percentage of unacceptable sentences.

4.2. Experiment results

4.2.1. Contribution of smoothing module

Following [5] , a maximum entropy model was trained from the training corpus containing 19,000 sentences, and tested on the test set. Not using the smoothing module, those word boundaries with an L3 probability over 0.5 are accepted as L3 breaks. This model is called MEO (Maximum Entropy model Only). In this way, we got an F-Score of 66.2%.

The length distribution probabilities of intonational phrase were also estimated from the same 19,000 sentences, and the distribution curve has been illustrated in Figure 1. We use MES to denote the maximum entropy model with smoothing module. Following Algorithm 1, an F-Score of 69.7% was achieved.

Table 1 shows the F-Score of human labeling consistency, MEO and MES. The relative F-Score compared to human labeling consistency are also listed in it. From Table 1 we can see, although the absolute value of F-Score is not high, they are around 90% relative to human labeling consistency. We compute the relative improvement of MES over MEO in Table 2.

	F-Score	Relative to human labeling consistency
Human labeling consistency	75.0%	100%
MEO	66.2%	88.3%
MES	69.7%	92.9%

Table 1: F-Score of different models

We also let an annotator score those test sentences labeled by MEO and MES, and compute the average score and unacceptable rate separately. The results are listed in Table 2. Table 2 shows that, the MES model makes an improvement of 5.3% in terms of F-Score, 10.0% in terms of average score, and 55.6% in terms of unacceptable rate. It is obvious that the improvement on average score and unacceptable rate is much more significant than F-Score. So, the main contribution of the smoothing module is to make the prediction result more acceptable by human, and at the same time, make it more accurately match the standard test set. It is simultaneously proved that the length distribution information of intonational phrase is of great usefulness for intonational phrase prediction.

	MEO	MES	Relative Improvement
F-Score	66.2%	69.7%	5.3%
Average score	4.0	4.4	10.0%
Unacceptable rate	1.8%	0.8%	55.6%

Table 2: Performance comparison of MEO and MES

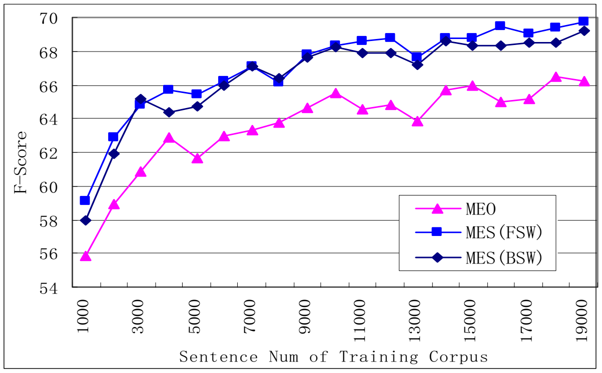

Figure 3: Performance curves of MEO and MES

4.2.2. Performance variation along with training corpus size

In order to investigate the performance variation of MEO and MES along with the training corpus size, we trained them from different size of corpus. In MES model, we applied FSW and BSW separately. The three performance curves are illustrated in Figure 3. The two curves of MES are higher than MEO obviously, and averagely, at each point, about 3 percent of improvement is achieved. Between the two curves of MES, there is no significant difference. The MES with FSW seems to be very slightly better than the MES with BSW.

5. CONCLUSIONS

In this paper, we proposed a sliding window smoothing method for intonational phrase prediction. We first build a maximum entropy model to yield the L3 probability of each word boundary, and then the length distribution curve of intonational phrase is applied as a sliding window to do smoothing. In this way, local context information and statistical length distribution information are efficiently integrated into the prediction system. Experiment results show that the length distribution information is of great usefulness for L3 break prediction, and the sliding window smoothing algorithm can improve the performance significantly.

Sliding window smoothing method is a post-processing module for L3 break prediction, which exploits the length distribution information efficiently. Generally speaking, it can be attached after any other statistic model to improve the performance, as long as the statistic model is able to yield the probability of each word boundary to be an L3 break. In the future, we will combine this smoothing module to other statistic models such as Decision Tree, to investigate its contribution.

Performance of FSW and BSW are similar in our experiments. In applications, when the results of FSW and BSW are different, could we achieve a better result by selecting one between them? This will be investigated later.

ACKNOWLEDGEMENT

This research is supported by the National Natural Science Foundation of China (60475015).

References

[1] M. Chu, Y. Qian, “Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts”. *Computational Linguistics and Chinese Language Processing*, February 2001, Vol.6, No.1: 61-82.

[2] M. Wang, J. Hirschberg. “Automatic Classification of Intonational Phrase Boundaries”. *Computer Speech and Language*, 6:175-196, 1992.

[3] NIE Xin, WANG Zuo-ying. “Automatic Phrase Break Prediction in Chinese Sentences”. *Journal of Chinese information Processing*, 2003, 17(4):39-44.

[4] G. J. Busser, W. Daelemans, Van den Bosch, A. “Predicting phrase breaks with memory-based learning”. *Proceedings 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Perthshire Scotland, August 29th - September 1st, 2001.

[5] Jian-feng Li, Guo-ping Hu, Wan-ping Zhang, Ren-hua Wang. “Chinese Prosody Phrase Break Prediction Based on Maximum Entropy Model”. *International Conference on Spoken Language Processing (ICSLP 2004)*. Oct 4-8, Korea.

[6] Richard O. Duda, Peter E. Hart, David G. Stork. “Pattern Classification”.

[7] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. “A maximum entropy approach to natural language processing”. *Computational Linguistics* 1996, 23(4): 597-618.