Syntactic Structure to Prosodic Structure Mapping with Inductive Learning Method

Jianhua Tao[†], **Shen Zhao**[‡]

† National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

‡ Dept. of Computer Science and Technology, Tsinghua University

E-mail: jhtao@nlpr.ia.ac.cn, szhao00@mails.tsinghua.edu.cn

ABSTRACT

During the last several years, there has been a rapid progress in Chinese speech synthesis. Now, the method of unit selection and concatenation, accompanying with large corpus, is used widely in the systems design. Nevertheless, the prosodic structure was still proved to be the essential links between linguistics and acoustics, and behaves as an important parameter for prosody processing and unit selection. Features related to prosodic boundaries are extracted with the corresponding boundary labels to establish a templates set. A transformational learning based method is applied to establish the prediction model for the prosodic phrasing and prosodic word parsing. The paper generates general evaluation parameters for prediction model. The importance of the features related to the prosody boundaries are also analyzed in the paper. The experiments show that the method approach can achieve the accuracy rate of 84% for prosodic word boundary and 76% for prosodic phrase boundary.

1. INTRODUCTION

During the last several years, there has been a rapid progress in Chinese speech synthesis. Now, the method of unit selection and concatenation, accompanying with large corpus, is used widely in the systems design. Nevertheless, the prosodic structure was still proved to be the essential links between linguistics and acoustics, and behaves as an important parameter for prosody processing and unit selection.

Traditionally, Chinese syntactic structure can be divided into four level, syllable, word, phrase and sentence. However, segmenting a sentence into a string of syntactic words and syntactic phrases is still far from enough for generating natural prosody. To get the relationship between syntactic structure and prosodic structure, prosodic word and prosodic phrase are induced in [8].

Some methods have been introduced to predict prosodic phrase. Such as, CART model (Wang and Hirschberg, 1992), HMM Model (Paul and Alan, 1998), which yields an accuracy of 86.6%. Ying and Shi used Recurrent Neural Network (RNN) to do the Chinese prosodic phrasing, and get an accuracy of 84.7% within 2000 Chinese sentences. Part-of-speech bigram and CART based methods are also tried in (Yao and Min, 2001), which reports a rather high recall and precision rate. But, the definition of the problem set of CART is usually time-consuming. Statistical methods like HMM and bigram need large training corpus or special techniques to avoid sparse data problems. RNN has good learning ability but the learned knowledge is represented in network weights, which are difficult to be explained and understood. Due to the difference in training corpus and evaluation methods between researchers, their results are generally less comparable.

The paper is focused on prosodic boundaries prediction with phonetic and syntactic information. An inductive learning method, transformational based learning (TBL) method, is introduced here. TBL is used for POS tagging and syntactic chunking originally. In this paper, it is proposed to predict prosodic boundaries. The paper is organized as following. In section 2, the acoustic features of prosodic boundaries are analyzed briefly. Detailed description of TBL and prediction model is described in section 3. Section 4 reports the evaluation results of prosodic boundary prediction. The context features are also compared here according to the different prediction results. Section 5 presents the conclusion and the view of future work.

2. ACOUSTIC FEATURES OF PROSODIC BOUNDARIES

Chinese is a tonal language. Prosodic rhythm (structure) plays a very important role in prosody processing. Although the perceptual reality of prosody rhythm has not been unambiguously observed in acoustic correlates, speech has always been considered as rhythmic one way or another. (Shen, 1986) mentioned in his paper, the bottom line of syllabic pitch range usually shows the trend of decline within the prosodic phrases, and top line of pitch range reflect the accents. The rhythm is the result generated

ISBN 1-876346-48-5

by the joint effect of syllabic pitch range, duration of syllable and duration of silence [3][4][5]. Normally the bottom line of the syllable above prosodic boundary shows the character of deep sunken while the corresponding syllable duration is also lengthened somewhat, though those are not the definitive rules for the various situation in spontaneous speech [13]. In some other situation, the duration of silence is also lengthened according to different layers of the prosody structure.

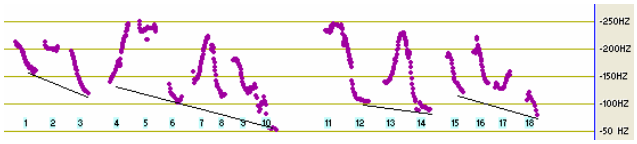

Figure 1, the decline of the bottom line in phrases

3. PROSODIC BOUNDARY PREDICTION

3.1 Context information

Context information generated by text parsing is the initial features for the prosodic boundary predicting model, which can be classified into three sets, phonetic information <P>, syntactic information <S> and other information <O>.

Phonetic information contains syllabic tone and syllabic structure (the types of initial and final) which are related to prosody. The features are essential to influence the detailed pitch contours of the intonation, such as tone sandhi, pitch continuity.

Syntactic information has close relation with prosodic structure. *POS*, which denotes part-of-speech of words, is a basic syntactic feature much easier to be obtained with automatic *POS* taggers. It has the strong influence to accent in speech. We use POS features from three POS sets simultaneously. The first one is the POS set of the tagger having 30 POS tags. The second one is much larger, in which the top 100 frequent words themselves are treated as independent POS tags in addition to those in the first set. The last one has only two tags: content words or functional words. The content words are those belonging to POS tags that are open word set. The functional words are on the contrary. To improve the results and get more context information, a POS window of 3-word width, one to the left and one to the right of the boundary, is adopted here. The features can be denoted as $<S>=(<S_1>,<S_2>,<S_3>)$, where $<S_1>$, $<S_2>$ and $<S_3>$ denote the three POS sets separately.

From the statistical figures of the corpus, both prosodic word and phrase have limitation in length. The length of syntactic word, the length of the sentence in syllable and word are length features to be considered. And the type of the sentence that is declarative, interrogative, imperative, or exclamatory may be useful, too. In HMM-based methods, the chain of boundary labels in a sentence is supposed to conform to Markov assumption. Thus the label of previous boundaries (*BTYPE*) and the distance from them to current position are also possible features. The other feature set includes some additional features: (1) the length of each word in the POS window, in Chinese characters; (2) the length of the sentence, in words and Chinese characters; (3) the position from the current boundary to the start and end of the sentence, in words and Chinese characters; (4) the distances from the current boundary to the first pervious break or non-break boundary, in words and Chinese characters.

3.2 Corpus design and labeling

The corpus contains 4000 sentences, which are randomly chosen from newspaper and read by a professional female speaker. Two experienced annotators labeled the sentences with two-level prosodic boundaries (prosodic phrase and prosodic word) by listening to the record speech. The criteria of the prosodic boundary labeling is described as below.

Bottom line of the syllabic pitch range reflects the prosodic layers very well and shows the trend of decline within the prosodic unit. Most probably, the bottom line of the syllable shows discontinuity and sunken in prosodic boundaries, the duration of the syllable above to the boundaries is also lengthened.

The labeling results of them achieve a consistency rate of 85% firstly. After the discussion and revision, it reaches 96%. The number of prosodic word boundaries is 35231 and 15620 for prosodic phrase boundaries. The average length of syntactic word, prosodic word, prosodic phrase and sentence are 1.5, 2.5, 6.2 and 14.0 in syllable

3.3 Transformational based learning

A classifier is a function that maps the input feature vector $\vec{F} = (x_1, x_2, ..., x_n)$ to a confidence that the input belongs to a class. In the case of prosodic phasing, the features are from linguistic information around the boundary and the classes are the boundary labels. As shown in Fig 2, TBL starts with a supervised training corpus that specifies the correct values for some linguistic feature of interest, a baseline heuristic for predicting initial values for that feature, and a set of rule templates that determine a space of possible transformational rules. The patterns of the learned rules match to particular combinations of features in the neighbourhood surrounding a word, and their action is to change to system's current guess as to the feature for that word.

To learn a model, one first applies the baseline heuristic to produce initial hypotheses for each site in the training corpus. At each site where this baseline prediction is not correct, the templates are then used to form instantiated candidate rules with patterns that test selected features in the neighbourhood of the word and actions that correct the

ISBN 1-876346-48-5

currently incorrect tag assignment. This process eventually identifies all the rule candidates generated by that template set that would have a positive effect on the current tag assignments anywhere in the corpus.

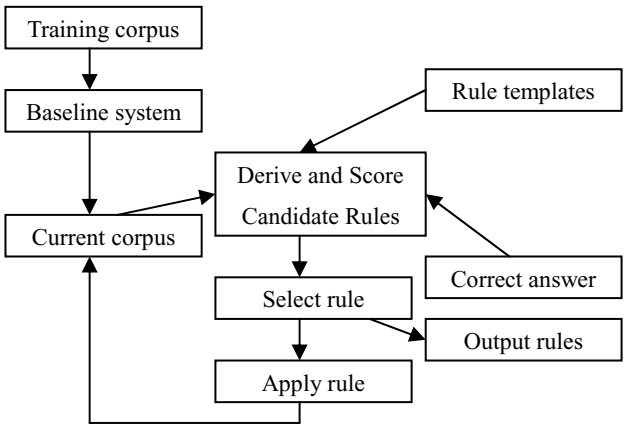

Figure 1, Transformational based learning

Those candidate rules are then tested against the rest of corpus, to identify at how many locations they would cause negative changes. One of those rules whose net score (positive changes minus negative changes) is maximal is then selected, applied to the corpus, and also written out as the first rule in the learned sequence. This entire learning process is then repeated on the transformed corpus deriving candidate rules, scoring them, and selecting one with the maximal positive effect. This process is iterated, leading to and ordered sequence of rules, with rules discovered first coming before those discovered later. The predictions of the model on new text are determined by beginning with the baseline heuristic prediction and then applying each rule in the learned rule sequence in turn.

Transformational learning begins with some initial "baseline" prediction, which here means a baseline assignment of prosodic boundary tags to syntactic words. In our work, baseline heuristics after a text has been tagged for context information, such as tone, types of initial and final, POS and other features. We tested both approaches, and the baseline heuristic using POS tags turned out to do better, so it was the one used in our experiments. The POS tags used by this baseline heuristic, and then later also matched against by transformational rule patterns, were derived by running the raw texts in a prepass through Brill's transformational POS tagger(Brill,1993)

4. RESULTS AND EVALUATION

To evaluate the ability of generalization of the learned rules, 2-fold cross validation tests are executed on the corpus for TBL. As a classification task, prosodic boundary prediction should be evaluated with consideration on all the boundary labels. The rules induced from examples are applied on a test corpus to predict the label of each boundary. The predicted labels are compared with labels given by human, which are thought to be true, to get a confusion matrix as follows:

True labels	Predicted labels	
	B_0	B_1
B_0	C_{00}	C_{01}
B_1	C_{10}	C_{11}

Table 1, Confusion matrix

C_{ij}s are the counts of boundaries whose true label are B_i but predicted as B_j. B_0 means the boundary of prosodic word and B_1 is the boundary of prosodic phrase. From these counts, we can deduce the evaluation parameters for prosodic boundaries prediction.

$$\mathrm{Re}call_i = \frac{C_{ii}}{C_{i0} + C_{i1} + C_{i2}} (i = 0,1)$$

$$\mathrm{Pr}ecison_i = \frac{C_{ii}}{C_{oi} + C_{1i} + C_{2i}} (i = 0,1)$$

$\mathrm{Re}call_i$ defines the recall rate of boundary label B_i. $\mathrm{Pr}ecison_i$ defines the precision rate of B_i.

To know which features in context information are more important for prediction, several tests were tried here.

Features	B_0		B_1	
	Recall	Precision	Recall	Precision
<P>,(<S_1>,<S_2>,<S_3>),<O>	0.83	0.84	0.72	0.76
(<S_1>,<S_2>,<S_3>),<O>	0.82	0.84	0.73	0.70
(<S_1>,<S_2>,<S_3>)	0.81	0.83	0.71	0.75
<S_1>,<S_2>	0.75	0.82	0.70	0.71
<S_1>	0.78	0.79	0.71	0.68
<P>,<O>	0.62	0.67	0.65	0.65
<P>	0.54	0.48	0.43	0.49
<O>	0.64	0.69	0.60	0.53

Table 2, prediction results of prosodic boundaries with different input features

From the results of the experiments (shown in table 2), we can draw some conclusions on the effect of the features involved for prosodic boundaries prediction. (1) Part-of-speech is a basic and useful feature. (2) Large *POS* set performs better than the small one. That's because small *POS* sets leads to small feature space, which is not big enough to distinguish the training examples. (3) Length information is beneficial to prosodic boundary prediction. (4) Phonetic information is less useful than syntactic information. Further test also shows that predicting history is helpful to make decision on current label. Although the error prediction of former labels may lead to another error

ISBN 1-876346-48-5

on current label, the result shows the whole performance is improved.

5. CONCLUSIONS

In this paper, we describe an effective and inductive learning method to generate rules for Chinese prosodic structure prediction. The main idea is to extract appropriate features from the linguistic information and apply rule-learning algorithms to automatically induce rules for predicting prosodic boundary labels. TBL learning algorithms are experimented in our research. The learned rules can achieve a good accuracy rate around 84% for prosodic word and 76% for prosodic phrase on test data. However, it's difficult to compare the results with those reported in [11] [12] for the two reasons. First, the models are tested in different testing data set with different content and different size. Second, there may be some different criteria in the definition of the prosody boundaries in spontaneous speech.

After the training from the corpus, TBL method generates lots of rules which are very important for prediction. Then, it is possible to classify the rules into different types and revise them according to experience or other methods. Further work will be focused on testing the rules generated by TBL method, compared them with statistic results, and try to make them work better.

REFERENCES

[1] Abney Steven. (1995) Chunks and dependencies: bringing processing evidence to bear on syntax. Computational Linguistics and Foundations of Linguistic Theory, CSLI

[2] Brill,Eric.(1995). Transformation-Based Error –Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging.Cmputational Linguistics 21(4):543- 565

[3] Hu weixiang, Xu bo, Huang Taiyi, "The study of detection and recognition of Chinese speech prosody boundaries", 6th National conference on human computer communication, Shengzhen, 2001,11

[4] Li Aijun, "A national database design for speech synthesis and prosodic labeling of standard Chinese", Oriental COCOSDA'99 Taipei, 1999

[5] Lin Maocan, "The acoustic manifestation of prosodic phrase boundaries in standard Chinese", Proc. of Conference on Phonetics of the languages in China, City University of Hong Kong, 1998

[6] Wang Pei, Yang Yufang, "Prosodic Structure and Syntactic Structure", The proceedings of the third international Conference on Cognitive Science, P491-496, 2001

[7] Grace Ngai and Radu Florian.(2001) Transformation-Based Learning in the Fast Lane, Proceedings of North Americal ACL 2001, 40-47

[8] Julia Hirschberg, Owen Rambow.(2001) Learning Prosodic Features using a Tree Representation, Eruospeech 2001

[9] Li Aijun, Lin Maocan. (2000) Speech corpus of Chinese discourse and the phonetic research. ICSLP2000

[10] Michelle Wang and Julia Hirschberg. (1992) Automatic classification of intonational phrase boundaries. Computer Speech and Language 6:175–196.

[11] Yao Qian, Min Chu, Hu Peng. (2001) Segmenting unrestricted Chinese text into prosodic words instead of lexical words, ICASSP2001

[12] Zhiwei Ying and Xiaohua Shi. (2001) An RNN-based algorithm to detect prosodic phrase for Chinese TTS, ICASSP2001

[13] Wang Pei, Yang Yufang, Lv Shinan, "Acoustic correlation of Chinese prosodic structure", 5th National Conference on Modern Phonetic, P161-165, 2001, Beijing

[14] Goldman-Eisler, F. 1972. "Pauses, clauses, sentences." Language & Speech, 15, 103 – 113.

[15] O'Malley, M. H., Kloker, O. R., & Dara-Abrams, D. 1973. "Recovering parentheses from spoken algebraic expressions." I.E.E.E. Transactions on Audio and Electroacoustics, AU–21 (3), 217 – 220.

[16] Streeter, L. A. 1978. "Acoustic determinants of phrase boundary perception". JASA, 64 (6), 1582 – 1592.

[17] Cooper, W. E., Paccia, J. M., & Lapointe, S. G. 1978. "Hierarchical coding in speech timing". Cog. Psychology, 10, 154 – 177.

[18] Lehiste, I. 1972. The timing of utterances and linguistic boundaries. JASA, 51 (6), 2018 – 2024.

[19] Huggins, A. W. F. 1974. An effect of syntax on syllable timing. Quarterly Progress Report, MIT, 114, 179 – 185.

[20] Zhao Sheng, Tao Jianhua etc, "Prosodic phrasing with inductive learning", ICSLP2002, Denver