CHINESE PROSODIC PHRASING WITH EXTENDED FEATURES

†Zhao Sheng ‡Tao Jianhua §Jiang DanLing

Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China
{†szhao00, §jdl00}@mails.tsinghua.edu.cn ‡jhtao@mail.tsinghua.edu.cn

ABSTRACT

Prosodic phrasing is an important component in modern TTS systems, which inserts natural and reasonable breaks into long utterance. This paper reports the study of prosodic phrasing in unrestricted Chinese text. A text corpus of 500 sentences is collected from our speech database and manually labeled with syntactic structure and prosodic structure. Features and target prosody labels are extracted from the corpus and used as training examples for a rule-learning program. The acquired rules are evaluated on unseen sentences. The experiments show that the tree-level syntactic features are the most effective ones for Chinese prosodic phrasing. And chunk-level features can also help to improve the prediction accuracy.

1. INTRODUCTION

Prosodic phrasing or prosodic phrase prediction plays an important role in improving the naturalness and intelligence of TTS systems. Linguistic research shows that the utterance produced by human is structured in a hierarchy of prosodic units, including phonological phrase, intonation phrase and utterance [1]. Prosodic structure makes the utterance sound natural and sometimes can help resolve syntactic ambiguity. But the output of syntactic analysis in TTS framework is often a structure of syntactic units, such as words or phrases, which are usually not equivalent to the prosodic ones. Therefore the object of prosodic phrasing is to map the syntactic structure into its prosodic counterpart.

A lot of methods have been introduced to predict prosodic phrase in English text. These methods are mainly data-driven based procedure such as Classification and Regression Tree (CART) [2], Hidden Markov Model (HMM) [3], Neural Network Autoassociators[4]. For Chinese prosodic phrasing, the traditional method is based on handcrafted rules. And Recurrent Neural Network (RNN) [5] as well as part-of-speech (POS) bi-gram and CART based methods [6] is experimented recently. An HMM based statistical method for prosodic structure prediction is also reported in [7]. However, due to the difference in training corpus and evaluation methods between researchers, the results are generally less comparable.

This paper explores prosodic phrasing with three sets of syntactic level features: base part of speech features, chunk-level features and tree-level features. All the features together with the boundary labels are collected at each word boundary of a speech corpus to establish training and testing datasets, which are used by a rule-learning program. Section 2 of the paper briefly describes the framework and evaluation methods for prosodic phasing. Section 3 proposes three sets of syntactic features. Experimental results are presented in section 4. In section 5 we discuss the results, followed by our conclusions.

2. PROSODIC PHRASING

2.1. Prosodic Phrasing Framework

It has been shown that Chinese utterance is also structured in a prosodic hierarchy, in which there are mainly three levels of prosodic units: prosodic word, prosodic phrase and intonation phrase [8]. Since intonation phrase is usually indicated by punctuation marks, what we need to consider is the prediction of prosodic word and phrase. Figure 1 shows the prosodic structure of a Chinese sentence. In the tree structure, the non-leaf nodes are prosodic units and the leaves are syntactic words. A prosodic phrase is composed of several prosodic words, each of which in turn consists of several syntactic words.

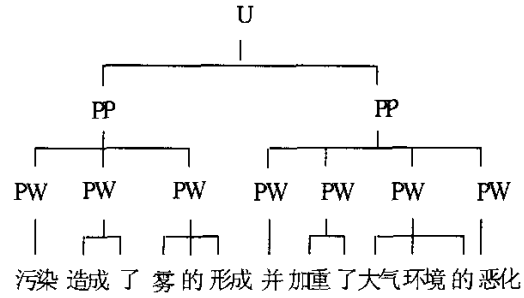

Figure 1: Two-level prosodic structure tree
(U for intonation phrase, PP for prosodic phrase, PW for prosodic word)

Suppose we have a string of syntactic words i.e. $w_1, w_2, \ldots w_n$, the boundary between two neighbouring words w_i, w_{i+1} is the object to be studied. There are total three types of boundaries, which can be labelled as B_0 (w_i, w_{i+1} are in the same prosodic word), B_1 (the words are in the same prosodic phrase, but not the same prosodic word), or B_2 (the words are in different prosodic phrases). Assume the label of a boundary is determined by its contextual linguistic information represented by a feature vector $\vec{F}$, prosodic phrasing can be viewed as a classification problem that in essence can be handled with any trained classifiers, taking the feature vector $\vec{F}$ as input and giving the most probable boundary label as output.

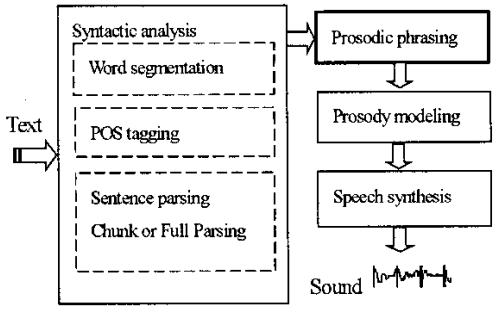

Figure 2: Prosodic phasing in TTS framework

2.2. Evaluation Parameters

Prosodic phrasing can be evaluated with subjective or objective measure. The subjective measure is generally performed by perceptive tests, which are undoubtedly convincing but time-consuming to conduct on large corpus. In this paper, only the objective measure is adopted. As a classification task, prosodic phrase prediction should be evaluated with consideration on all the boundary labels. The trained classifiers are applied on a test corpus to predict the label of each boundary. Then the predicted labels are compared with labels given by human, which are thought to be true, to get a confusion matrix shown in table 1.

True labels	Predicted labels		
	B_0	B_1	B_2
B_0	C_{00}	C_{01}	C_{02}
B_1	C_{10}	C_{11}	C_{12}
B_2	C_{20}	C_{21}	C_{22}

Table 1: Confusion matrix

C_{ij}s are the counts of boundaries whose true label are B_i but predicted as B_j. From these counts, we can deduce the evaluation parameters for prosodic phrasing.

$$\mathrm{Re}c_i = C_{ii} / \sum_{j=0}^{2} C_{ij} \qquad (i = 0,1,2) \tag{1}$$

$$\mathrm{Pr}e_i = C_{ii} / \sum_{j=0}^{2} C_{ji} \qquad (i = 0,1,2) \tag{2}$$

$$F_i = 2 * \mathrm{Re}c_i * \mathrm{Pr}e_i / (\mathrm{Re}c_i + \mathrm{Pr}e_i)(i = 0,1,2) \tag{3}$$

$$Acc = \sum_{i=0}^{2} C_{ii} / \sum_{j=0}^{2} \sum_{i=0}^{2} C_{ij} \tag{4}$$

$\mathrm{Re}c_i$ defines the recall rate of boundary label B_i, while $\mathrm{Pr}e_i$ defines the precision rate of B_i. Since the counts of different boundary labels are usually unbalanced in the corpus, F_i is used as a combination of the recall and precision rate [10]. Acc is the overall accuracy of all the labels. If the number of labels is reduced to two, the evaluation parameters can be deduced similarly.

3. EXTENDED FEATURES

Linguistic information around word boundary is the main source of features. The features may come from different levels including syllable, word, phrase, and sentence level. And the type of features can be phonetic, lexical, syntactic, and semantic. Which features have most close relation with prosodic phrasing and how to represent them are still open research problems. A good feature set can improve the prediction accuracy but the design of it is usually work intensive and needs much linguistic experience [9].

3.1. Base POS features

Part-of-speech (POS) sequences are the most popular features used in the previous research. And it's much easier to automatically get POS tags from unrestricted Chinese text than other deep syntactic structures such as syntactic phrase or components. 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 most frequent 100 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 sets. The functional words are on the contrary. The adoption of multiple POS sets results in POS features of different granularity. For a word boundary, a context window of 5 words is applied with three words to the left and two words to the right.

3.2. Chunk-level features

Text chunking consists of dividing a text in syntactically correlated parts of words [11]. For example, the sentence in Figure 1 "污染造成了雾的形成并加重了大气环境的恶化." can be divided as follows:

[NP 污染] [VP 加重 了] [NP 雾 的 形成] [O 并] [VP 加重 了] [NP 大气 环境 的 恶化] .

In the above, NP, VP and O are syntactic chunks, which are non-overlapping regions of a text and non-recursive.

There are five types of chunks recognized after segmentation and tagging: NP (noun chunk), VP (verb chunk), PP (prepositional chunk), ADJP (adjective chunk) and ADVP (adverb chunk). Furthermore, chunk tags can be defined for each word in the sentence. The tags now we use are B-NP, I-NP, B-VP, I-VP, B-PP, I-PP, B-ADJP, I-ADJP, B-ADVP, I-ADVP and O. B-NP is for the first word of a noun chunk and I-NP is for words in a noun chunk that are not B-NP. O is for words that are not in any chunk. Other chunk tags are similar to the case of noun chunks. The sentence with chunk tags would look like:

污染/B-NP 加重/B-VP 了/I-VP 雾/B-NP 的/I-NP 形成/I-NP 并/O 加重/B-VP 了/I-VP 大气/B-NP 环境/I-NP 的/I-NP 恶化/I-NP

Just like POS features, the chunk tags of words around a word boundary can be extracted as features to predict prosodic chunking.

3.3. Tree-level features

Full syntactic parsing builds a phrase structure or dependency structure of a sentence. We adopt phrase structure as the tree representation of grammar. The phrase tree of the example sentences is:

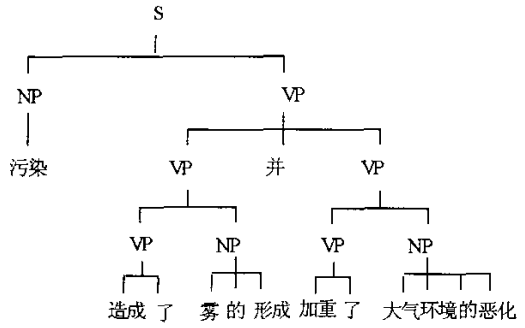

Figure 3: The syntactical structure tree

For each word juncture w_i, w_{i+1}, we extract the following tree-level features:

- SBP: the smallest syntactic phrase dominating both words of the juncture.
- LSP: the largest syntactic phrase dominating the left word but not the right.
- LRP: the largest syntactic phrase dominating the right word but not the left.
- The length of the phrases SBP, LSP and LRP, in both words and characters.

3.4. Three feature sets

We also compute some length features such as the length of the words in the context window, the length of the sentence, the position of the boundary in the sentence etc. The base POS features together with length features forms the first feature set *FEATA*; Chunk-level features plus *FEATA* constitutes *FEATB; FEATC* is composed of tree-level features and *FEATA*. The three feature sets, *FEATA*, *FEATB* and *FEATC*, incorporate different level syntactic information respectively. Their effects on prosodic phrasing are inspected in the following experiments.

4. EXPERIMENTS

4.1. The corpus

In our experiments, a speech corpus for our TTS system is used for training and testing. The corpus has 500 long sentences, which are randomly chosen from newspaper and read by a radiobroadcaster. Two experienced annotators label the sentences with two-level prosodic boundaries by listening to the record speech. The labeling results of them achieve a high consistency rate. There are 9754 Chinese characters in the corpus, which constitute 6378 words. The number of prosodic word boundaries (B_1) is 2749 and that of prosodic phrase ones (B_2) is 1208.

The sentences of the corpus are also processed with a text analyzer, where Chinese word segmentation and part-of-speech tagging are accomplished in one step using a statistical language model. The segmentation and tagging results are corrected manually.

Finite state machine technique is adopted to recognize chunks from POS sequences. The words are grouped into chunks and then chunk tags are assigned to each word. The chunking results are also checked manually.

Full syntactic parsing is totally done by hand. We labeled the syntax structure of the sentences and got a small tree-bank style corpus.

4.2. Phrasing experiments

There are three boundary classes (B_0, B_1, B_2) in the corpus, the prediction of which is a multi-class classification task. To simplify the problem, we merge the three classes into two since in most systems only one prosodic level is used to generate target pitch contours. It's possible to merge B_0, B_1 into a class B_{01}, or merge B_1, B_2 into a class B_{12}, which gives rise to two different classification problems. The former one is prosodic word prediction, while the latter is prosodic phrase prediction. For learning algorithms, the main difference is that the training data of prosodic phrase prediction is more heavily unbalanced than that of prosodic word prediction..

To estimate the generalization ability of a learning algorithm, we apply five-fold cross validation test on the corpus to obtain the generalized results. The corpus data is divided equally into five portions. At each step we train the algorithms on four portions and test them on the rest one.

4.3. Results

We conducted several experiments on the same corpus

Features	*Classes*	Rec_0	Pre_0	F_0	Rec_1	Pre_1	F_1	*Acc*
FEATA	B_{01}, B_2	*0.947*	*0.841*	*0.890*	*0.470*	*0.750*	*0.578*	*0.826*
	B_0, B_{12}	*0.830*	*0.862*	*0.846*	*0.912*	*0.890*	*0.900*	*0.879*
FEATB	B_{01}, B_2	*0.980*	*0.933*	*0.956*	*0.496*	*0.777*	*0.605*	*0.920*
	B_0, B_{12}	*0.888*	*0.888*	*0.888*	*0.925*	*0.926*	*0.925*	*0.900*
FEATC	B_{01}, B_2	*0.975*	*0.936*	*0.955*	*0.774*	*0.900*	*0.832*	*0.929*

Table 2. Experimental results using three feature sets

using the three feature sets to predict prosodic word and phrase boundaries. The machine-learning program we used is C4.5 decision tree package [12]. The results of them are showed in *Table 2*. From the table, first we can see that prosodic word prediction can be accomplished with a high recall and precision rate. The F_0 and F_1 measures of it are both above 84% for *FEATA* and *FEATB*. The use of *FEATB* improves about 2% on the F measure compared with *FEATA*. Second prosodic phrasing with *FEATA* or *FEATB* produces low F_1 measures, although chunk-level features in *FEATB* result in about 10% improvement on the total accuracy Acc. The use of tree-level features in *FEATC* helps to acquire the best F_1 measure of 83.2% and Acc of 92.9%

5. DISCUSSION

From the experiments, it is shown that the carefully designed syntactic features are effective for prosodic phrasing. Since the corpus used in our experiments is somewhat small, we need to label more sentences and evaluate the features on them. For real applications, our method is to be combined with syntactic parsers, which extract syntactic features. The overall performance will be decided by both prosodic and syntactic phrasing.

However, it's difficult to compare our results with those reported in [5] [6] because the corpus used and the evaluation methods are different between different researchers. Thus a well-accepted prosodic corpus should be built up to advance the research of Chinese prosodic phrasing in the future. Although the application of machine learning algorithms on prosodic phrasing has been a popular strategy, there are still some problems remaining untouched:

- From experience, a sentence may have several prosodic structure without changing its meaning. The current framework cannot handle this case.
- How well can prosodic phrasing improve the quality of synthetic speech? What's the relationship between the speech quality and the recall/precision rate?

6. CONCLUSIONS

In this paper, we explore the application of three syntactic feature sets for the Chinese prosodic phrasing problem. Features and target prosodic boundaries are extracted from the same corpus to form training and testing data sets, on which machine-learning classifiers are trained and evaluated. The results demonstrate that the incorporation of deeper syntactic information can improve the accuracy of prosodic phrase prediction. Especially the tree-level syntactic features give the best results, and chunk-level features give the second.

7. ACKNOWLEDGEMENTS

We would like to thank the reviewers of ICASSP2003 for their serious comments. This work is supported by 863 Program of China (No:001AA114072).

8. REFERENCES

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

[2] Michelle Wang and Julia Hirschberg, Automatic classication of intonational phrase boundaries, Computer Speech and Language 6:175–196, 1992.

[3] Paul Taylor and Alan.W.Black, Assigning phrase breaks from part-of-speech sequences, Computer Speech and Language v12, 1998.

[4] Achim F. Muller, Hans Georg Zimmermann and Ralph Neuneier, Robust generation of symbolic prosody by a neural classifier based on autoassociators, ICASSP2001.

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

[6] Yao Qian, Min Chu and Hu Peng, Segmenting unrestricted chinese text into prosodic words instead of lexical words, ICASSP2001.

[7] Qin Shi, XiJun Ma, WeiBin Zhu, Wei Zhang and LiQin Shen, Statistic Prosody Structure Prediction, Proceedings of IEEE 2002 Workshop On Speech Synthesis

[8] Li Aijun and Lin Maocan, Speech corpus of Chinese discourse and the phonetic research, ICSLP2000.

[9] Julia Hirschberg and Owen Rambow, Learning Prosodic Features using a Tree Representation, EuroSpeech2001.

[10] C.J. van Rijsbergen, Information Retrieval, Butterworths, London, 1979.

[11] http://cnts.uia.ac.be/conll2000/chunking/

[12] Quinlan,J.R., Induction of decision trees, Machine Learning, 1(1):81-106, 1986.