Chinese Prosody Structure Prediction Based on Conditional Random Fields

Jingwei Sun, Jing Yang, Jianping Zhang, Yonghong Yan
Institute of Acoustics, Chinese Academy of Science
Thinkit Speech Laboratory
Beijing, China

Abstract

In this paper, a novel statistical method based on Conditional Random Fields (CRF) is proposed for hierarchical prosody structure prediction, which is a key module in speech synthesis systems. We will discuss how to build the prosody models for mandarin Chinese using Conditional Random Fields in detail, including corpus preparation, feature selection, feature template design, model training and evaluation. Comparison is conducted between the new method and the classical decision tree based one. The experimental results show that CRF-based method can significantly improve the overall performance with the same feature set.

1. Introduction

Prosody structure modeling/prediction plays an important role in speech synthesis systems. Linguistic research shows that the utterance produced by human is structured in a hierarchy of prosodic units. In Chinese Text to Speech (TTS) systems, a typical hierarchical prosody structure consists of lexical word, prosodic word, prosodic phrase and breathe group, which are denoted by B0, B1, B2 and B3 for convenience sake. Figure 1 gives an example of prosody boundary labeling and shows the prosody structure of the sentence. Each lexical word is marked by a prosodic boundary and the whole sentence is segmented by these boundaries. In this way, a hierarchical prosody structure is constructed. In the tree structure, the non-leaf nodes are prosodic units and the leaves are lexical words that can be derived from a lexical-based word segmentation module. Whether the prosodic boundaries are properly predicted will affect the naturalness and intelligibility of TTS directly.

A variety of researches have been done in this field and some effective methods are proposed. For Chinese prosody structure prediction, the traditional method is based on handcrafted rules [1]. This approach is simple and convenient, but it is quite time-consuming to get lots of trivial rules. In recent years, many researchers exploited statistical-based method and achieved good performance, including Classification and Regression Tree (CART) [2], Markov Model [3], Maximum Entropy Model [4], Memory Based Learning [5] and Artificial Neural Networks.

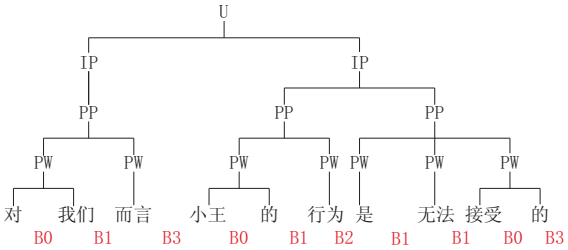

Figure 1. Three-level prosody structure tree (U for utterance, BG for breathe group, PP for prosodic phrase, PW for prosodic word, the bottom layer is lexical word).

Recently, a new framework for building probabilistic models, called Conditional Random Fields [6], were proposed to segment and label sequence data and have been successfully applied to a variety of NLP (Natural Language Processing) tasks, such as POS tagging and topic segmentation. CRFs have several advantages over HMMs (Hidden Markov Models) and stochastic grammars for such tasks, such as the ability to relax strong independence assumptions made in those models. CRFs also avoid a fundamental limitation of MEMMs (Maximum Entropy Markov Models) and other discriminative Markov models based on directed graphical models, which can be biased towards states with few successor states.

Prosody structure prediction can also be treated as a sequence labeling task. The boundary decision of the current unit is strongly correlated to that of the context units. Therefore, we use CRFs for prosody structure prediction, which is also called CRF-based prosodic boundary labeling in this paper. We will discuss how to build the prosody models for mandarin Chinese using CRF, including corpus preparation, feature selection, feature template design, model training

978-0-7695-3736-8/09 $25.00 © 2009 IEEE
DOI 10.1109/ICNC.2009.44

and evaluation. We will also conduct comparative experiments to confirm the effectiveness of the new method.

This paper is organized as follows: Section 2 gives a brief introduction of CRFs principle. In section 3, feature selection and evaluation metric of CRF-based prosodic boundary labeling are described. Detailed experiments design and implementation are discussed in section 4 and conclusion remarks are presented in the final part.

2. A Brief introduction of Conditional Random fields

CRFs can be considered as a generalization of logistic regression to label sequences. They define a conditional probability distribution of a label sequence $\vec{y}$ given an observation sequence $\vec{x}$. In this paper, $\vec{x} = (x^1, x^2, ..., x^n)$ denotes a sentence of length n and $\vec{y} = (y^1, y^2, ..., y^n)$ denotes the label sequence corresponding to $\vec{x}$. In prosodic boundary labeling, x^t denotes a boundary and y^t is a label noting the type of the boundary.

CRFs specify a linear discriminative function F parameterized by Λ over a feature representation of the observation and label sequence $\Psi(\vec{x}, \vec{y})$. The model is assumed to be stationary, thus the feature representation can be partitioned with respect to positions t in the sequence and linearly combined with respect to the importance of each feature Ψ_k, denoted by λ_k. Then the discriminative function can be stated as in Equation 1,

$$F(\vec{x}, \vec{y}) = \Sigma_t(\Lambda, \Psi(\vec{x}, \vec{y})), \tag{1}$$

Then, the conditional probability is given by

$$p(\vec{y}|\vec{x}; \Lambda) = \frac{1}{Z(\vec{x}, \Lambda)} F(\vec{x}, \vec{y}; \Lambda), \tag{2}$$

where $Z(\vec{x}, \Lambda) = \sum_{\tilde{y}} F(\vec{x}, \tilde{y}; \Lambda)$ is a normalization constant which is computed by summing over all possible label sequences $\tilde{y}$ of the observation sequence $\vec{x}$. Since CRFs condition on the observation sequence, they can efficiently employ feature representations that incorporate overlapping features, i.e. multiple interacting features or long-range dependencies of the observations, as opposed to HMMs which generate observation sequences. In this work, the overlapping features are acquired by designing lots of feature templates based on the atomic features.

In CRFs, the objective function is the log-loss of the model with parameters with Λ respect to a training set D. This function is defined as the negative sum of the conditional probabilities of each training label sequence y_i, given the observation sequence x_i, where $D \equiv \{(x_i, y_i) : i = 1, ..., m\}$. CRFs are known to overfit, especially with noisy data if not regularized. To overcome this problem, we penalize the objective function by adding a Gaussian prior (a term proportional to the squared norm $\| \Lambda \|^2$) as suggested in [7]. Then the loss function is given as:

$$\begin{aligned} L(\Lambda; D) &= -\sum_i^m \log p(y_i|x_i; \Lambda) + \frac{1}{2} c \| \Lambda \|^2 \\ &= -\sum_i^m F(x_i, y_i; \Lambda) + \log Z(x_i, \Lambda) + \frac{1}{2} c \| \Lambda \|^2, \end{aligned} \tag{3}$$

where c is a constant. Given an observation sequence $\vec{x}$, the best label sequence is given by:

$$\hat{y} = \arg\max_y F(x, y; \hat{\Lambda}), \tag{4}$$

where $\hat{\Lambda}$ is the parameter vector that minimizes $L(\Lambda; D)$. The best label sequence can be identified by performing the Viterbi algorithm.

3. Feature Selection and Evaluation Metric for CRF-based prosodic boundary labeling

3.1. Feature Selection

No matter which model used, feature selection is crucial to the classification of prosodic boundary labels. Linguistic information around the word boundary is the main source of features. The features can be derived from different levels, including syllable, word, phrase, and sentence level. And the type of features may be phonetic, syntactic, semantic or pragmatic. Many researches have been done to find the most closely related features with prosodic phrasing. In this work, atomic features are selected based on Zhao Sheng's research [8], including the part-of-speech (POS), the length in syllables and the word itself of the words surrounding the boundary. The window size of 1+1 is adopted. Because CRFs can make labeling decisions based on the whole sentence, we do not need to extract position information as features, such as distance from the sentence beginning or sentence end to the current boundary , although they are proved to be useful. Because prosodic boundary labeling is a complicated labeling problem, atomic features of words and POS tags are not sufficient to describe actual language phenomenon. Based on the atomic features, combined features are created. By this means, context information can also be taken into account. For example, prediction result of the previous boundary can be used as an input feature for the current boundary for labeling decision. For CRFs, feature selection and combination are realized by designing a lot of corresponding feature templates, which will be described in detail in Section 4.

	B0	B1	B2	B3
B0	C_{00}	C_{01}	C_{02}	C_{03}
B1	C_{10}	C_{11}	C_{12}	C_{13}
B2	C_{20}	C_{21}	C_{22}	C_{23}
B3	C_{30}	C_{31}	C_{32}	C_{33}

Table 1. Confusion matrix for Prosody Labeling.

3.2. Evaluation Metric

As a classification task, prosodic boundary labeling should be evaluated with consideration of all kinds of classes. The predicted labels are compared with labels given by human, which are thought to be true, to get a confusion matrix as shown in Table 1. C_{ij}s are the counts of boundaries whose true label are Bi, but predicted as Bj. From these counts, we can deduce the evaluation parameters for prosody prediction.

$$Rec_i = C_{ii} / \sum_{j=0}^{3} C_{ij}, i = 0, 1, 2, 3 \quad (5)$$

$$Pre_i = C_{ii} / \sum_{j=0}^{3} C_{ji}, i = 0, 1, 2, 3 \quad (6)$$

$$F_i = 2 \bullet Rec_i \bullet Pre_i / (Rec_i + Pre_i), i = 0, 1, 2, 3 \quad (7)$$

$$Rec_a = \frac{1}{4} \sum_{j=0}^{3} Rec_j \quad (8)$$

$$Pre_a = \frac{1}{4} \sum_{j=0}^{3} Pre_j \quad (9)$$

$$F_a = \frac{1}{4} \sum_{j=0}^{3} F_j \quad (10)$$

Rec_i defines the recall rate of boundary label Bi. Pre_i defines the precision rate of Bi. Fi is a combination of recall and precision rate, called F-score, suggested by [9]. Rec_a, Pre_a and F_a define the recall rate, precision rate and F-score of prosodic boundary labeling with the overall consideration of different boundary types.

4. Experiments

4.1. The corpus

Totally 20000 sentences randomly selected from People's Daily were used in our experiments. Word segmentation and POS tagging were carried out by a front-end preprocessing program. The accuracy of word segmentation is 96% and the accuracy of POS tagging is 90%.

Tags need to be labeled at each boundary to indicate B0, B1, B2, B3 breaks in advance. Lack of the corresponding speech, annotators labeled prosodic boundaries by reading the sentences themselves. As we know, different people might have different judgment on the same boundary. Through testing, the labeling consistency among the annotators was 73%.

Sentences in the corpus were divided into two groups, 1% for testing and the other for training.

4.2. Performance Comparison between CRFs and CART

As described in Section 3.1, the atomic features used for model training include W-1, W+1, P-1, P+1, L-1, L+1, with their explanation given in Table 2. As a result, the definition of tokens for a boundary can be denoted as "W-1 W+1 P-1 P+1 L-1 L+1 Bi", where Bi is the answer tag of this boundary. Table 2 shows the training tokens for the example sentence. Tokens for each sentence in the training corpus can be acquired automatically from the annotated data through a feature extraction module.

Feature Tag	Feature Explanation
W-1	The lexical word before the boundary
W+1	The lexical word after the boundary
P-1	POS of the lexical word before the boundary
P+1	POS of the lexical word after the boundary
L-1	Length of the lexical word before the boundary
L+1	Length of the lexical word after the boundary

Table 2. Explanation of atomic features used in CRFs training.

	W-1	W+1	P-1	P+1	L-1	L+1	Bi
1	对	我们	p	r	1	2	B0
2	我们	而言	r	u	2	2	B1
3	而言	，	u	w	2	0	B3
4	小王	的	nr	u	2	1	B0
5	的	行为	u	n	1	2	B1
6	行为	是	n	v	2	1	B2
7	是	无法	v	d	1	2	B1
8	无法	接受	d	v	2	2	B1
9	接受	的	v	u	2	1	B0
10	的	。	u	w	1	0	B3

Figure 2. Training tokens of a sentence for CRFs.

Apart from these atomic features, we designed many feature templates to combine different features together. By this means, more correlated context information were considered. This is the key point why CRFs is better than

other statistic modeling methods for sequence data labeling. Totally 99 templates were designed and some of them are given in the appendix. In each template, special macro %x[*row, col*] will be used to specify a token in the input data. *row* specifies the relative position from the current focusing token and *col* specifies the absolute position of the column. The template B is a bigram template, which denotes the prediction result of the previous boundary. We take boundary 6 in Figure 2 as example, give some samples of templates and their expanded features and presented them in Figure 3.

Templates	Expanded Features
%x[0,0]	行为
%x[-2,2]/%x[-2,3]	nr_u
%x[-1,0]/%x[1,2]/%x[-1,4]	d_v_1
B	B1

Figure 3. Examples of templates and their expanded features.

F-score, as described in Section 3.2, was used for performance evaluation. Besides measurement of the overall labeling accuracy of the four boundary types, we also evaluated the labeling performance of B2 separately, for it plays a very important role in Chinese prosodic structure and is more difficult to detect.

The CART-based modeling, which is the classical prosody structure prediction method, is applied as a comparison with the same corpus and features used.

The experiment result is presented in Table 3. From the table, we can see that CRFs achieve better performance than CART. CRFs modeling makes an improvement of 3.3% in terms of overall F-score and 6.1% in terms of F-score of prosodic phrase boundary.

Model	Pre_a	Rec_a	F_a	Pre_2	Rec_2	F_2
CART(%)	87.4	86.8	88.1	63.9	62.9	63.4
CRFs(%)	90.9	91.0	91.0	66.2	68.3	67.3

Table 3. Labeling perfermance comparision between CRFs and CART.

In order to investigate the performance variation of the two models along with the training corpus size, we trained them from different size of corpus. Figure 4 shows the performance curves of the two modeling method.

Furthermore, we applied the two prosody models to our HMM-based TTS system. To evaluate the prosody modeling performance, we just changed the prosody prediction module of the system. 100 open-set sentences were synthesized by the two systems and 46 pairs of them had different

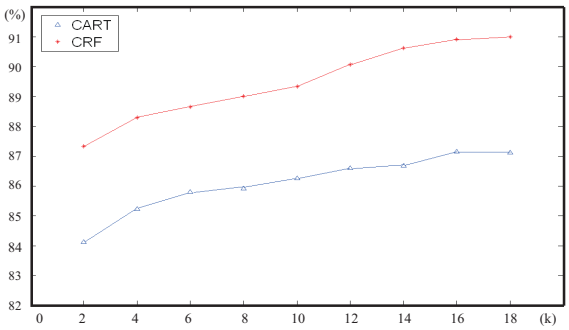

Figure 4. Performance curves of CART and CRFs.

prosody prediction results. A preference listening test was conducted on these 46 pairs of speech waves. 6 trained listeners were asked to give preference score of naturalness for each pair. Figure 5 shows the result of this test. From the figure, we can find that CRF-based prosody modeling can achieve better naturalness of synthesized speech with a preference score of 76%.

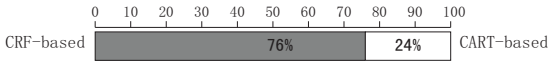

Figure 5. Listening preference score of CRF-based modeling to CART-based modeling for TTS.

5. Conclusion

In this paper, a CRF-based method was proposed for Chinese prosody structure prediction. We treated prosody structure prediction as a sequence labeling task firstly and then introduced Conditional Random Fields to solve this problem, which is very effective for sequence labeling. The experiments showed that the CRFs can get better prediction accuracy than CART. Preference listening test of the synthesized speech also confirmed the effectiveness of the new method when applied it to a TTS system.

More complex syntactic and phonetic information will be explored for prosody structure prediction under the CRFs framework in the future. We will also focus on data balance and model optimization. Our ultimate goal is to build a new TTS system that can achieve better synthetic naturalness by improving the prosody prediction module.

References

[1] Zhengyu Niu, Peiqi Chai, "Segmentation of Prosodic Phrases for Improving the Naturalness of Synthe-

sized Mandarin Chinese Speech", *Proceeding of ICSLP Conference*, 2000, pp.350-353.

[2] 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.

[3] NIE Xin, WANG Zuo-ying, "Automatic Phrase Break Prediction in Chinese Sentences", *Journal of Chinese information Processing*, 2003, 17(4):39-47.

[4] Jian-feng Li, Guo-ping Hu, Wan-ping Zhang, and Ren-hua Wang, "Chinese Prosody Phrase Break Prediction Based on Maximum Entropy Model", *International Conference on Spoken Language Processing*, 2004, Oct 4-6, Korea.

[5] 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.

[6] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", *Proc. of 18th International Conference Machine Learning*, 2001.

[7] M. Johnson, S. Geman, S. Canon, Z. Chi, and S. Riezler, "Estimators for stochastic unification-based grammars", *Proc. of ACL'99 Association for Computational Linguistics*.

[8] Sheng Zhao, Jianhua Tao and Danling Jiang, "Chinese Prosody Phrasing with Extended Feature", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

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

6. Appendix

Mark	Template
U000	%x[-2,0]
U001	%x[-1,0]
U002	%x[0,0]
U003	%x[1,0]
U004	%x[2,0]
U010	%x[-2,1]
U011	%x[-1,1]
U012	%x[0,1]
U013	%x[1,1]
U014	%x[2,1]
U020	%x[-2,2]
U021	%x[-1,2]
U022	%x[0,2]
U023	%x[1,2]
U024	%x[2,2]
U030	%x[-2,3]
U031	%x[-1,3]
U032	%x[0,3]
U033	%x[1,3]
U034	%x[2,3]
U041	%x[-1,4]
U042	%x[0,4]
U043	%x[1,4]
U051	%x[-1,5]
U052	%x[0,5]
U053	%x[1,5]
U060	%x[-2,0]/%x[-2,1]
U061	%x[-1,0]/%x[-1,1]
U062	%x[0,0]/%x[0,1]
U063	%x[1,0]/%x[1,1]
U064	%x[2,0]/%x[2,1]
U070	%x[-2,2]/%x[-2,3]
U071	%x[-1,2]/%x[-1,3]
U072	%x[0,2]/%x[0,3]
U073	%x[1,2]/%x[1,3]
U074	%x[2,2]/%x[2,3]
U080	%x[-2,4]/%x[-2,5]
U081	%x[-2,4]/%x[-2,5]
U082	%x[-2,4]/%x[-2,5]
U083	%x[-2,4]/%x[-2,5]
U084	%x[-2,4]/%x[-2,5]
U090	%x[-2,1]/%x[-2,3]/%x[-2,5]
U091	%x[-2,0]/%x[-2,2]/%x[-2,4]
U092	%x[-1,1]/%x[-1,3]/%x[-1,5]
U093	%x[-1,0]/%x[-1,2]/%x[-1,4]
U094	%x[0,1]/%x[0,3]/%x[0,5]
U095	%x[0,0]/%x[0,2]/%x[0,4]
U096	%x[1,1]/%x[1,3]/%x[1,5]
U097	%x[1,0]/%x[1,2]/%x[1,4]
U098	%x[2,1]/%x[2,3]/%x[2,5]
U099	%x[2,0]/%x[2,2]/%x[2,4]
U102	%x[0,1]/%x[0,3]/%x[0,5]/%x[0,2]/%x[0,4]/%x[0,0]
U103	%x[1,1]/%x[1,3]/%x[1,5]/%x[1,2]/%x[1,4]/%x[1,0]
U104	%x[2,1]/%x[2,3]/%x[2,5]/%x[2,2]/%x[2,4]/%x[2,0]
U110	B
...	...

Table 4. Feature templates designed for training