HIERARCHICAL PROSODIC BOUNDARY PREDICTION FOR UYGHUR TTS

Askar Hamdulla[1] Guljamal Mamateli[2] AskarRozi[3] Sayyare Imam[4]
[1]School of Software of Xinjiang University, Urumqi, China
E-mail:askar@xju.edu.cn
[2]Institute of Mathematic Science of Xinjiang Normal University, Urumqi, China
guljamal123@gmail.com
[3]Institute of Mathematics and System Science of Xinjiang University, Urumqi, China
E-mail:askhar@xju.edu.cn
[4] College of Politics and Public Administration of Xinjiang University, Urumqi, China
E-mail:sayyarim@163.com

***Abstract*—Correct prosodic boundary prediction is crucial for the quality of synthesized speech. This paper presents the prosodic hierarchy of Uyghur-language which belongs to agglutinative language. A two-layer bottom-up hierarchical approach based on conditional random fields (CRF) is used for predicting prosodic word (PW) and prosodic phrase (PP) boundaries. In order to disambiguate the confusion between different prosodic boundaries at punctuation sites, CRF based prosodic boundary determination model is used and integrated with bottom-up hierarchical approach. Word suffix feature is considered useful for prosodic boundary prediction and added into the feature sets. The experimental results show that the proposed method successfully resolves the confusion between different prosodic boundaries. Consequently, further enhance the accuracy of prosodic boundary prediction.**

I. INTRODUCTION

The prediction of prosodic boundary is an important step in TTS system. Prosodic boundaries induce prosodic structure in sentences, thus making them more naturally sounding and intelligible.

One of the main obstacles to automatic generation of prosody is the difficulty of identifying the hierarchical prosodic constituents from texts automatically. The early research on prosodic boundary prediction was based on syntactic analysis and rule-based methods. Various data-driven approaches, such as hidden Markov model (HMM) [1], CART-based hierarchical model [2-3], maximum entropy model [4-5] have been investigated to predict prosodic boundary in most previous studies. No unique prosodic units and features have been claimed as a basic units or features for prosodic boundary prediction in previous work, since it is a language-dependent task, where prosodic units, features and a prediction algorithm need to be optimally selected for a particular language.

Almost all Punctuations are considered to imply intonation phrase (INP) boundary in many prosody prediction methods [2, 3, 5], but the influence of punctuations on different prosodic boundaries has not been considered. In practice, punctuations (especially comma) not only relate to INP boundary, but also relate to all the other prosodic boundaries.

There are 74.1% punctuations related to INP boundary, 25.8%related to PP boundary and a few punctuations related to PW boundary in the corpus of our approach. Although, most of the punctuations are related to INP boundary, but the influence of punctuation on different prosodic boundaries are cannot be neglected. We found that 70.1% of wrong predicted boundaries occurred at punctuation sites in the sentence from the PW and PP boundaries prediction results. It indicates that a technical prosodic boundary determination model of punctuation sites is very important for prosody prediction method.

At present, there has no research work investigated on Uyghur prosodic boundary prediction. In this paper, we present the prosodic hierarchy of Uyghur language according to the linguistic and syntactic analysis. Bottom-up hierarchical approach based on CRF model is used to predict the PW and PP boundaries, prosodic boundary determination model of punctuation sites is used and integrated with the bottom-up hierarchical approach to handle the confusion between different prosodic boundaries at the punctuation sites. Word suffix features are also added into the feature sets according to the agglutinative nature of Uyghur-language. Finally, the missed and wrong predicted punctuation sites are corrected in the PW and PP boundaries prediction results and more accurate prosodic boundary prediction performance is achieved by disambiguating the confusion between different prosodic hierarchies at the punctuation sites.

The rest of the paper is organized as follows: Section 2 describes the lexical and prosodic structure of Uyghur-language; Section 3 outlines our prosodic boundary prediction method; Section 4 summarizes the results; and Section 5 presents our conclusions.

II. LEXICAL AND PROSODIC STRUCTURE OF UYGHUR-LANGUAGE

In this part, we mainly present lexical and prosodic structure of Uyghur-language which relate to our work.

Uyghur is an agglutinative language and belongs to the Turkish language family of the Altaic language system.

A. Lexical Structure

Uyghur text is written from right to left in Arabic scripts with some modifications. Words are separated by space or other punctuation and formed by affixes attaching to the stem (or root).

Morpheme structure of Uyghur word is "Prefix+Stem+Suffix1+Suffix2+…". A root (or stem) is followed by zero to many (at longest 10 or more) suffixes. In this work, 108 suffix types are defined according to their syntactic and semantic functions, which have 305 surface forms. A few words have a (only one) prefix preceding a stem; seven kinds of prefixes are considered. There are two types of suffix in Uyghur-language which are derivational suffixes and inflectional suffixes. Suffixes that make semantic changes to a root are derivational suffixes and make syntactic changes to a root are inflectional suffixes. Words which have same syntactic function in the sentence are attached same inflectional suffix and it is regarded as very useful features for predicting prosodic boundary in Uyghur-language.

B. Prosodic Structure

Prosodic hierarchy of Uyghur-language consists of PW, PP and INP according to the linguistic and syntactic analysis [6]. Reliability of prosodic boundary annotation rules is verified by perceptual experiment and acoustic analysis of different prosodic boundaries. Experiments and analysis are made on 1497 sentence speeches which are recorded in a natural way by 28 year-old female broadcaster. Figure 1 shows the prosodic hierarchy of Uyghur sentence "His lecture has high level and strong persuasion".

Figure.1 Prosodic hierarchy of Uyghur sentence

According to prosodic boundary annotation rules, PW and PP boundaries are annotated according to the analysis of sentence constituent, Phrase structure, auxiliary and independent component. An INP is a simple sentence or sub sentence in a composite sentence and almost can be separated by punctuation. The proportion of comma (68.1%) is big among all punctuation marks in large text corpus of Uyghur TTS.

One of the major reasons causing the confusion between different prosodic boundaries at the punctuation site is that apposition constituent in the sentence is regarded as PP boundary in prosodic boundary annotation rules [6] and separated by comma in the sentence, since there is no caesura sign in Uyghur as in English. PP and INP boundary conflict at comma site in this case when predicting prosodic boundary.

III. METHOD

Two-layer bottom-up hierarchical approach based on CRF model is used to predict the PP and PW boundaries in our approach. There is no prediction model for INP boundary, since INP boundaries are separated by punctuation in prosodic boundary annotation rules. Prosodic boundary determination model of punctuation sites is used and integrated with two-layer bottom-up hierarchical model to disambiguate the confusion between different prosodic boundaries at punctuation sites and to correct the missed or wrong predicted punctuation sites from the results of PW and PP prediction model. Word suffix is also considered as useful feature for prosody prediction and added into the feature sets.

A. CRF

We consider the prosodic boundary prediction as a sequence labeling problem. Given feature vector sequence X of the potential prosodic boundary, we aim to find the optimal label sequence Y to maximize $P(Y|X)$, where Y indicates the predicted prosodic boundary. Considering that Conditional Random Field (CRF) [7] provides a probabilistic framework for calculating the probability of Y globally conditioned on X, we build CRF classifiers for solving this problem.

In the training process of the CRF classifier, we need to estimate the parameter $\Lambda = \{\lambda_k, \mu_l\}$ in Equation (1):

$$P(Y|X) = \frac{1}{Z_x} exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, X) + \sum_{i,l} \mu_l g_l(y_i, X)\right) \quad (1)$$

Where Z_x is the normalization constant that makes the probability of all label sequences sum to one; $f_k(y_{i-1}, y_i, X)$ is an arbitrary feature function over the entire feature vector sequence and the labels at position i and $i-1$; $g_l(y_i, X)$ is a feature function of label at position i and the feature vector sequence. λ_k and μ_l are weights learned for the feature functions, reflecting the confidence of feature functions. The feature functions describe any aspect of a transition from y_{i-1} to $y_i$as well as y_i and the global characteristics of X.

Given the parameter Λ, the most probable labeling sequence can be produced as $Y^* = \text{argmax}_Y P_\Lambda(Y|X)$. The marginal probability of labels at each position in the sequence can be computed by a dynamic programming inference procedure similar to the forward-backward procedure for HMM. We then calculate the marginal probability of each potential boundary being a prosodic boundary given the whole segment sequence by $P(y_k = \{PW, PP, INP\}|X)$. Finally we get the optimal boundary sequences according to the marginal probability values.

B. Bottom-up hierarchical approach

A bottom-up approach [2,3,5]considers the hierarchical relationship between different prosodic hierarchies and yields better prediction performance, in which prosodic units are predicted gradually from smaller to larger ones with

respective classifier model and adopts a bottom-up way of predicting different prosodic boundaries step by step.

In our approach, two-layer bottom-up hierarchical approach is used to predict PW and PP boundaries. To train the PW-CRF, all the potential boundaries (PB) of lexical words (LW) are treated as non-boundary samples and all the others (PW and PP) are boundary samples. To train PP-CRF, only PW boundaries are used as non-boundary samples, and PP were boundary samples.

C. *Prosodic boundary determination model*

We can get the PW and PP boundaries from the results of bottom-up hierarchical approach, INP boundary can be separated by punctuation, but not all the punctuation sites are related to INP boundary, some of them are also related to PW and PP boundary, and wrong predicted or missed at some times in the results of two layer bottom-up hierarchical approach. In order to determine a prosodic boundary which is more related to punctuation sites and to correct the missed and wrong predicted punctuation sites from the results of two-layer bottom-up hierarchical model, we use a prosodic boundary determination approach of punctuation sites (PS) and integrated with bottom-up approach, as shown in the Figure2.To train the punctuation-CRF, all the punctuation sites in the sentence are treated as non-boundary samples which may be LW, PW, PP or INP. The outputs of the model are whether the punctuation site is PW, PP or INP boundary. After identify the punctuation sites which can be regarded as PW or PP boundary according to the predicted results from the punctuation-CRF, rest of the punctuations are considered to be INP boundary.

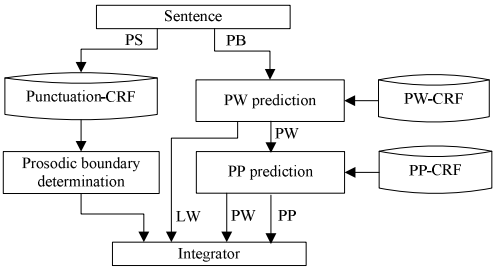

Figure.2Frame work of proposed method

D. *Feature set*

Feature selection is very important step for a classifier. Selection and combination of feature templates significantly affect the prosody prediction performance. Three sets of feature are used for predicting the prosodic boundary by mainly considering the linguistic knowledge. The first is Part-of-Speech (POS) set, which is a commonly used feature in the related research [1-5]. The second is position set which includes the features relating to position or text length. The third is word suffix which is the unique feature of agglutinative language. Considering that the apposition constituent, which causes one of the major confusion of prosodic boundary at punctuation sites, has same suffix in the sentence. For example, the apposition constituents in the sentence below "She has been writing, studying, working and caring for baby." have same suffix "دى", like a suffix "ing" in the English sentence.

ئۇ يازدى، ئوقۇدى، ئىشلىدى ۋە بالىسىغا قارىدى.

All the feature sets used in our approach, including their offset range and value range, are listed in table 1.

TABLE.1 FEATURE SETS

Feature Type	Offset range	Value range
Pos	±2, ±1, 0	41 types
Syllable in Word	±2, ±1, 0	continuous
Length of Word	±2, ±1, 0	continuous
Position of word in the sentence(from beginning& from ending)	-	continuous
Distance from current boundary to previous & next punctuation boundary	-	continuous
Word Suffix	±2, ±1, 0	108 types

IV. EVALUATION AND DISCUSSION

A. *Experimental data*

In this paper, The 40630 sentences which were annotated with summarized prosodic boundary annotation rules [6]are used for experimental test, the length of sentences is 10~25 Uyghur syllables. LWs were annotated manually by space and other punctuation between the words. PWs and PPs were annotated manually by reading the text transcriptions according to the prosody annotation rules, since lacking of corresponding speeches. The punctuation between the sub sentence and end of each sentence was labeled INP boundary. POS tagging were carried out by a preprocessing program in which a POS tag set consists of 41 POS tags and accuracy of 90.74% for POS tagging were achieved.

The corpus includes 281,114PW, 114,238PP and 70,953INP. Average syllable numbers of each prosodic boundary are 5/13/21Uyghur syllables respectively. There are altogether 40,911 items of punctuation in our corpus, 30,324 (74.1%) of which relate to INP, 25.8% relate to PP and only a few punctuations relate to PW. The corpus has been divided in two parts, one for training, with 80% of the sentences (total of 32,504), and the remaining for testing (total of 8126).

B. *Results and discussion*

Two experiments are conducted in this paper. First is baseline method with different feature sets, second is different prosody prediction method with same feature sets. Baseline approach is a one-step way of predicting PW, PP and INP boundaries with a single CRF model.

In order to investigate the influence of word suffix features on the prosody prediction, we experiment the baseline method with different feature sets. Table 2 shows the Precision, Recall and F-score of baseline method. The precision (P) is the accuracy percentage of prosodic boundary assigned by this approach. And there call (R) is the percentage of boundaries in the corpus that are found correctly. The F-score is defined as:

$$\mathrm{F-Score} = 2 \times \mathrm{P} \times \mathrm{R}/(\mathrm{P}+\mathrm{R}) \qquad (2)$$

It can be seen from table 2 that the prosody prediction performance with suffix set has about the same performance as those with POS set and position set. Meanwhile, by adding suffix set to the POS and position set, prediction results achieves about 4.22% and 7.34% higher performance in terms of F-score for PW and PP boundaries respectively. This indicates that suffix is an effective parameter for prosody prediction and more useful unique feature for agglutinative language.

TABLE.2EXPERIMENTAL RESULTS OF BASELINE METHOD WITH DIFFERENT FEATURE SETS

	PW			PP		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
suffix	53.36	61.49	57.14	79.01	50.37	61.52
pos	50.36	56.91	53.44	82.07	54.71	65.66
position	54.30	53.34	53.81	73.97	68.41	71.08
POS & Position	58.25	65.01	61.45	84.95	57.86	68.84
POS & suffix & Position	65.84	65.50	65.67	78.56	73.95	76.18

Three different prosody prediction methods with same feature sets are listed in the table 3. First is bottom-up hierarchical approach without considering prosodic boundary determination model of punctuation sites. Second is prosodic boundary determination approach described in section 3.3 PW boundary prediction failed in this experiment, since only a few PW boundaries relate to punctuation in our corpus. Third is our proposed approach which integrates the two-layer bottom-up hierarchical model and prosodic boundary determination model of punctuation sites. We only give the increasing percentage of PP boundary, since PW boundary failed in the prosodic boundary determination model and all the punctuation sites are considered to be INP boundary after identifying the punctuation which relates to PP boundary.

TABLE.3 EXPERIMENTAL RESULTS OF DIFFERENT PROSODY PREDICTION METHODS WITH SAME FEATURE SETS

		P(%)	R(%)	F(%)
Bottom-up	PW	78.4	74.9	76.6
	PP	82.4	66.8	73.8
punctuation	PP	79.1	73.9	76.4
	INP	88.1	90.9	89.5
Bottom-up & Punctuation	PP	84.1	75.6	79.6

It can be seen from the table 3 that the proposed approach achieves about 1.7% higher performances in terms of precision rate, 8.8% in terms of recall rate and 5.8%in terms of F-score.

From the results of bottom-up hierarchical model we found, there were all together 6435 boundary which were predicted wrong, 4581(70.1%) of which were related to punctuation, it indicates that technical prosodic boundary determination model of punctuation site is very helpful for prosody prediction.

Since punctuation sites are always accompanied by pause in the synthesized speech, the relationship between the pause and punctuation inferable from its text represents and people tend to expect a pause at the punctuation site intuitively from the TTS synthesis result. It seems the synthesized speech is abnormal, if the prosodic boundary at the punctuation sites are missed or predicted wrong. Therefore, we consider recall rate has a higher priority in our approach. Evaluation results also show our approach increases the performance of prosodic boundary prediction method on this aspect.

V. CONCLUSIONS

Prosodic hierarchy of Uyghur-language is presented in the paper. Bottom-up hierarchical approach based on the CRF model is used to predict PW and PP boundaries. To disambiguate the confusion between different prosodic hierarchies at the punctuation sites, the prosodic boundary determination model based on CRF is also used and integrated with bottom-up hierarchical model. The word suffix which is the unique feature of agglutinative language is also added into the feature set and shows high performance in prosody prediction. The experimental results show that the proposed method increases the performance of prosody prediction by 5.8% in terms of F-score, leading to a number of perceivable improvements in TTS quality, e.g., in regard to prosodic parameter estimation and pause insertion. This bodes well for applying an analogous strategy to other languages.

ACKNOWLEDGMENT

This work is supported by Program for New Century Excellent Talents in University (NCET-10-0969), Natural Science Foundation of China (No.61065005), and open project of Xinjiang Key Laboratory of Multilingual Information Processing (XJDX0201-2012-03).

REFERENCES

[1] Taylor, P. and Black, A. W., “Assigning Phrase Breaks from Part-of-Speech Sequences”, Computer Speech and Language, Vol. 12, 1998, pp. 99-117.

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

[3] DaweiXu, Haifeng Wang, Guohua Li, Takehiko Kagoshima, “PARSING HIERARCHICAL PROSODIC STRUCTURE FOR MANDARIN SPEECH SYNTHESIS”, ICASSP2006.

[4] Jianfeng Li, Guoping Hu and Renhua Wang, “Chinese prosody phrase prediction based on maximum entropy model”, Interspeech 2004, Jeju Island, Korea, 2004, pp. 729-732.

[5] Xiaonan Zhang, Jun Xu and LianhongCai “Prosodic Boundary Prediction based on Maximum Entropy Model with Error-Driven Modification”, Proceeding of ISCSLP 2006 Conference, Singapore, 2006, pp.149-160.

[6] GULMIRE Imam,ASKAR Hamdulla. “Research on the Rules and Regulation for Manual Labeling of Prosody Levels in Uyghur Sentence”, The third conference of natural language information processing of national minority, collection of papers of the 2nd national joint conference of construction of language knowledge base, 2010.

[7] J. Lafferty, A. McCallum, and F. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, in Proc. 18th Int.Conf. Machine Learning (ICML 2001), Williamstown, MA, June2001, pp. 282–289.

[8] Agustin Gravano, Martin Jansche, Michiel Bacchiani, “RESTORING PUNCTUATION AND CAPITALIZATION IN TRANSCRIBED SPEECH” in ICASSP 2009, pp.4741-4744.