PUNCTUATION GENERATION INSPIRED LINGUISTIC FEATURES FOR MANDARIN PROSODIC BOUNDARY PREDICTION

Chen-Yu Chiang, Yih-Ru Wang and Sin-Horng Chen

Institute of Communication Engineering, National Chiao Tung University

ABSTRACT

A novel statistical linguistic feature, called punctuation confidence, is proposed in this paper for assisting in prosodic break prediction in Mandarin text-to-speech. The punctuation confidence calculated from the input text is a measure of the likelihood of inserting a major PM at a word boundary. Since a punctuation in text tends to be pronounced as a break, the punctuation confidence associated with a punctuation estimate should provide useful information for break prediction from text. The idea is realized in this study by first employing a conditional random field (CRF)-based model to generate a predicted punctuation and its associated punctuation confidence for each word boundary. Then, the predicted punctuation and its punctuation confidence are combined with contextual linguistic features to predict the break type of the word boundary by an MLP (multi-layer perceptrons). Experiment on the Treebank speech corpus confirmed the effectiveness of the proposed approach.

Index terms **—** punctuation confidence, text-to-speech, prosodic break, punctuation generation, conditional random field

1. INTRODUCTION

Prosodic phrase boundary (or prosodic break) prediction from text plays a very important role in an unlimited text-to-speech (TTS) system. Proper prosodic break prediction would make the synthesized speech sound more natural and more intelligent in terms of intonation and rhythm without losing or destroying the original meaning of the text. Previous break prediction studies mainly focused on the following two issues: (1) Design or utilization of prediction model, and (2) Utilization of features. In the first issue, many prediction models have been proposed, including hierarchical stochastic model [1], N-gram model [2], classification and regression tree (CART) [3,4], bottom-up/sifting hierarchical CART [3], Markov model [5], artificial neural networks [6], maximum entropy model [7], etc. In the second issue, some shallow linguistic features, such as part of speech (POS), word length, sentence length, position in a sentence, etc., are very basic and popularly used. To further improve break prediction accuracy, many studies adopted higher level syntactic features, such as word chunk [6] and syntactic tree [6]. On the other hand, a statistical feature - connective degree (CD) [8] was proposed to neglect complex syntactic parsing that causes impracticality in constructing an unlimited TTS system.

This paper focuses on the second issue to propose a statistical linguistic feature called punctuation confidence which is motivated by automatic Chinese punctuation generation [9] and linguistic characteristic of Chinese punctuation system [10]. In [9], a maximum entropy (ME)-based automatic Chinese punctuation generation method was proposed to insert 16 types of punctuation mark (PM) to an un-punctuated text by using features of word and lexical-functional grammar (LFG) features. The results in [9] showed that the punctuation generation model can generate alternative/ acceptable insertion, deletion or substitution PMs. This phenomenon was also observed in a human punctuation experiment reported by Tseng [10] in which alternative punctuation strategies were found among different native Mandarin Chinese speakers. These observations reflect a fact that Chinese PMs serve as a loose reference to both syntactic structure and semantic domain, and therefore native Chinese writers would freely utilize PMs to delimit written Chinese into various linguistic elements, such as phrases and clauses, so as to clearly express the meaning of a text. Furthermore, punctuation generation of a speaker when reading written Chinese would reflect his/her prosodic phrasing strategy because pause break is highly correlated with punctuation. Therefore, an automatic punctuation generation model trained from a large text corpus may provide useful cues for prosodic break prediction.

In this study, a conditional random field (CRF)-based automatic punctuation generation model is constructed to predict punctuation and generate the associated confidence measure, referred to as punctuation confidence, from PM-removed word/POS sequences. The punctuation confidence can be regarded as a statistical linguistic feature to measure the likelihood of inserting a major PM into the text. It is reasonable to hypothesize that word junctures which are more likely to be inserted with major PMs in text, are more likely to be inserted with pause breaks in utterance. We can therefore use the punctuation confidence to help prosodic break prediction. Several advantages of the approach can be found. First, the punctuation confidence can be easily obtained from features of word/POS sequence which can be robustly obtained by current word segmentation and POS tagging technologies without using complicated statistical syntactic parsing. This makes the proposed approach more suitable for practical on-line unlimited TTS. Second, as trained using a large text corpus, the CRF-based punctuation generation model can learn alternative punctuation strategies from numerous paragraphs by various writers so as to generate more reliable punctuation confidence. Third, compared with the size of an available text corpus parsed with syntactic tree for constructing a statistical syntactic parser, the size of corpus used to train the CRF-based punctuation generator can be considerably larger. Therefore, we can expect that the punctuation confidence would be more robust than syntactic features derived from an automatic syntactic parser.

This paper is organized as follows. Section 2 introduces the experiment database and its prosody labeling. The relationship between prosodic breaks and PMs are illustrated. The proposed method is presented in Section 3. Section 4 discusses the experimental results. Some conclusions are given in the last section.

2. PROSODY LABELING OF SPEECH CORPUS

A read Mandarin speech database uttered by a female professional announcer was used to evaluate the proposed approach to prosodic break prediction. Its associated texts were all short paragraphs composed of several sentences selected from the Sinica Treebank Version 3.0 [11]. The database is further divided into two parts: a training set consisting of 376 utterances with 51,868 syllables and a test set consisting of 46 utterances with 4801 syllables.

2.1. Prosody labeling of the speech corpus

The corpus was labeled with seven break types by the PLM algorithm [12] proposed previous. As shown in Fig. 1, the seven break types, i.e. {*B*0, *B*1, *B*2-1, *B*2-2, *B*2-3, *B*3, *B*4}, delimit an utterance into a four types of prosodic units, namely syllable (SYL), prosodic word (PW), prosodic phrase (PPh), and breathe group/prosodic phrase group (BG/PG).

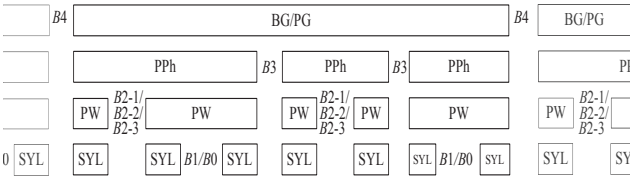

Fig. 1. The prosody-hierarchy model of Mandarin speech used in this study [12]

In the labeling system, each defined break type is characterized by its specific juncture prosodic-acoustic features: *B*4 is defined as a major break accompanying long pause and apparent F0 reset across adjacent syllables; *B*3 is a major break with medium pause and medium F0 reset; *B*0 and *B*1 represent respectively non-breaks of tightly-coupling syllable juncture and normal syllable boundary, within a PW, which have no identifiable pauses between SYLs; and *B*2 is a minor break with three variants: F0 reset (*B*2-1), short pause (*B*2-2), or pre-boundary syllable duration lengthening (*B*2-3).

2.2. Analysis on prosody labeling

2.2.1. Prosodic-acoustic features of the labeled break types

Among various types of prosodic-acoustic features, pause duration is the most salient cue to specify boundaries of prosodic units. Therefore, this study aims to improve the break prediction accuracy of those pause-related break types, i.e. *B*4, *B*3, and *B*2-2. Fig. 2 displays the distributions of pause duration for the seven break types. As can be seen from the figure, the break types of higher level were generally associated with longer pause duration. Notice that *B*4, *B*3 and *B*2-2 have apparent pause duration (>30ms), while *B*0, *B*1, *B*2-1 and *B*2-3 all have very short pause duration (<30ms). By the above analysis on the pause durations of the seven break types, this study defines four break prediction targets, including (1) *B*4, (2) *B*3, (3) *B*2-2, and (4) non-pause break type (*NPB*) which is a grouping of *B*0, *B*1, *B*2-1 and *B*2-3.

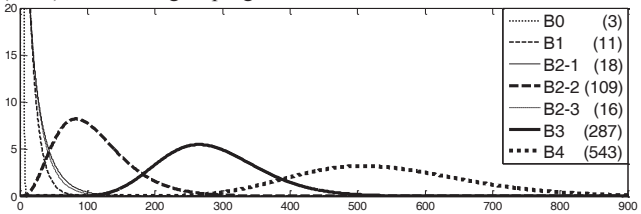

Fig. 2: The pdfs of pause duration. Numbers in () denote the mean values in ms.

2.2.2. Relationship between the labeled break types and PM types

It is generally agreed that pause breaks co-occur with PMs. Therefore, most TTSs cautiously insert pause only on major PMs, such as comma and period. This cautious strategy of pause insertion can make the synthesized speech very stable, but may be unnatural as the input sentence is very long. Table 1 shows the co-occurrence matrix of four target break types and three syllable juncture types calculated from the training set. It can be seen from the table that most PM locations co-occur with pause-related break type (*B*2-2, *B*3 and *B*4), while most intra-word locations map to *NPB*. In-between PM and intra-word, non-PM inter-word locations co-occur with *NPB*, *B*2-2 and *B*3. Actually, about 40% of prosodic phrase boundaries (*B*3s) and over 94% of *B*2-2 come from non-PM inter-word junctures. By more detail analysis, we find that 60% of non-PM *B*3s coincide with depth-1 node boundary of full parsed syntactic tree. The above discussions imply that it would be unsatisfactory to insert prosodic pause breaks only at PM locations. The study hence tries to overcome this shortage.

Table 1: Co-occurrence matrix of four target break types and three syllable juncture types.

	NPB	*B*2-2	*B*3	*B*4
Intra-word	21,970	14	2	0
Non-PM inter-word	20,288	3,148	1,391	30
PM	30	169	2,130	2,320

Table 2 shows the co-occurrence matrix of four target break types and 8 PM types. It can be found from the table that the major PM set {period ‘ ◦ ’, exclamation mark ‘ ! ’, question mark ‘ ? ’, semicolon ‘ ; ’, colon ‘ : ’, comma ‘ ， ’} is highly correlated with major breaks, i.e., *B*3 and *B*4. This implies that a word juncture which tends to insert major PM in text is more likely to be a major break in utterance. This motivates us in this study to propose a CRF-based automatic punctuation generator to predict the insertion of major PM (i.e., punctuation) and its likelihood (i.e., punctuation confidence) for each word juncture from an unpunctuated text, and use them to help the break prediction.

Table 2: Correlation matrix of 4 break types and 8 PM types

	◦	!	?	;	:	，	、	•
NPB	1	0	0	0	0	4	25	1
*B*2-2	2	1	1	0	1	88	75	1
*B*3	42	1	7	9	2	1,901	168	0
*B*4	606	39	58	63	0	1,523	30	1

3. THE PROPOSED METHOD

Fig.3 shows a diagram of the proposed break prediction scheme. As shown in the figure, the scheme contains four components, including (1) Word and POS Tagger, (2) CRF-based Punctuation Generator, (3) Context Analyzer, and (4) MLP (multi-layer perceptrons) Break Predictor. It is noted that the predicted punctuation and punctuation confidence are extra information in the feature vector (in addition to the contextual linguistic features) to help the break prediction. The system is operated as follows. First, the input raw text is segmented into word sequence with POS labeling by the Word and POS Tagger. Then, the CRF-based Punctuation Generator determines the presence/absence of punctuation and generates the associated punctuation confidence from the input PM-removed word and POS sequence. Last, the MLP Break Predictor determines the break type of each word

juncture using the feature vector which consists of the predicted punctuation, the punctuation confidence, and the contextual linguistic features formed by the Context Analyzer using the sequences of PM, word and POS.

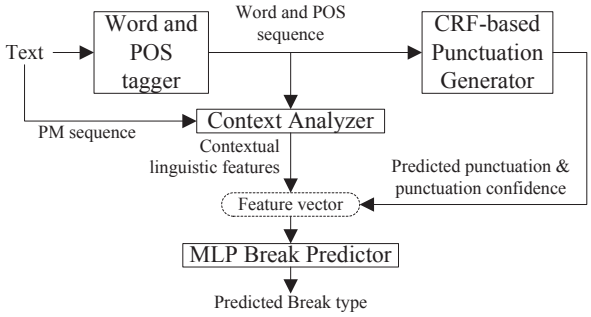

Fig.3: A diagram of the proposed break prediction scheme.

3.1. CRF-based Punctuation Generator

The task of the CRF-based Punctuation Generator can be viewed as a label-tagging problem that labels each lexical word juncture with the presence or absence of a major PM, **Y**, by using features of lexical word, **W**, and POS, **S**. It is hence formulated as

$$P(\mathbf{Y}|\mathbf{W},\mathbf{S})=\frac{1}{N(\mathbf{W},\mathbf{S})}\exp\left(\sum_{t=1}^{T}\sum_{i=1}^{I}\lambda_i f_i(Y_t=y,Y_{t-1},\mathbf{W},\mathbf{S})\right) \quad (1)$$

where $N(\mathbf{W},\mathbf{S})$ is a normalization factor to ensure that $\sum_{\mathbf{Y}}P(\mathbf{Y}|\mathbf{W},\mathbf{S})=1$; t stands for lexical word index; Y_t represents the presence or absence of a major PM between t-th and (t+1)-th lexical words; I represents the number of feature functions; and $f_i(Y_t=y,Y_{t-1},\mathbf{W},\mathbf{S})$ is a feature function defined by

$$f_i(Y_t=y,Y_{t-1},\mathbf{W},\mathbf{S})=\begin{cases}1, \text{ if } (\mathbf{W},\mathbf{S})=h_j \text{ is satisfied and } y=y_k\\ 0, \text{ otherwise}\end{cases} \quad (2)$$

where h_j represents the j-th possible feature context; and y_k is the k-th possible tag to be labeled. Generally, feature contexts are organized into several groups, referred to as 'feature templates'. Table 3 displays the feature templates used in this study.

Table 3: Targets, features and feature templates of the CRF-based punctuation generation

Target	description
Y_t	y_1 : presence of the major PM ('。', '！', '？', '；', '：', '，'), y_0 : absence of the major PM
Feature	**description**
W_t	t-th lexical word
S_t	POS of t-th lexical word
Feature templates (separated by comma)	
Lexical word context: W_t, W_{t+1}, W_t^{t+1}, W_t^{t+2}, W_{t-1}^{t+1}, W_{t-1}^{t+2},	
POS context: S_t, S_{t+1}, S_t^{t+1}, S_{t-1}^{t}, S_t^{t+2}, S_{t-1}^{t+1}, S_{t-1}^{t+2}, S_{t-2}^{t}, S_{t+1}^{t+3}, S_{t-2}^{t+1}, S_t^{t+3}, S_{t-2}^{t+2}, S_{t-1}^{t+3}, S_{t-2}^{t+3}	
Lexical word and POS context: (S_t, W_{t+1}), (W_t, S_{t+1}), (S_{t-1}, W_t^{t+1}), (W_t^{t+1}, S_{t+2}), (S_{t-1}, W_t^{t+2}), (W_{t-1}^{t+1}, S_{t+2}), $(S_{t-1}, W_t^{t+1}, S_{t+2})$	

The predicted punctuation can be obtained by

$$Y_1^*, Y_2^*, \cdots, Y_T^* = \arg\max_{Y_1, Y_2, \cdots, Y_T} P(\mathbf{Y}|\mathbf{W},\mathbf{S}) \quad (3)$$

And the punctuation confidence is given by

$$\varphi_t(\mathbf{W},\mathbf{S})=P(Y_t=y_1|\mathbf{W},\mathbf{S}) \quad (4)$$

which is the marginal probability of the presence of major PM.

3.2. Context Analyzer and MLP Break Predictor

The Context Analyzer is used to form a basic contextual linguistic feature vector for the break prediction task. The features used for each word juncture are:

- 6 Boolean flags for 6 PM types
- 5 real numbers for the length (in syllable) of the current/following/previous sentences (delimited by the major PMs), and the distances (in syllable) to the previous/following major PMs.
- (m+n)×82 POS Boolean flags which consist of
  - Level-3 POS of m previous/n following words: 47 categories proposed by CKIP [13].
  - Level-2 POS of m previous/n following words: 23 categories merged from Level-3 POS.
  - Level-1 POS of m previous/n following words: 12 categories also merged from Level-3 POS.
  - Broad class of m previous/n following words: substantial word or function word.
- (m+n) ×5 word length flags which consist of
  - Word lengths of m previous/n following words: word length of 1, 2, 3, 4 and ≥5 syllable(s).

The input feature vector of the MLP Break Predictor includes (1) the predicted punctuation, (2) the punctuation confidence $\varphi_t(\mathbf{W},\mathbf{S})$, and (3) basic contextual linguistic feature of (m+n)× 87+11 dimensions. The MLP Break Predictor is of 3-layer structure with one hidden layer. The output layer consists of five nodes corresponding to *NPB*, *B*2-2, *B*3, *B*4 and B_e (end of the utterance). Both the hidden and output layers use standard sigmoid functions. The training algorithm for the MLP is the error backpropagation algorithm.

4. EXPERIMENTAL RESULTS

The corpus used for training the CRF-based Punctuation Generator was the Academia Sinica Balanced Corpus of Modern Chinese V.4.0 [14] (ASBC4.0) which consists of 9,454,734 words (or 31,126 paragraphs). The number of target punctuation classes is 2: one for major PMs {。, ！, ？, ；, ：, ，} and another for all others. The CRF model was trained by the CRF++ [15] with a cut-off setting of the features that occured no less than 3 times in the given training data.

Table 4 displays the experimental results on the training set ASBC4.0 and the Treebank corpus used in the break prediction experiment. A high F-score of 85.8 was reached. This shows that the punctuation prediction is quite well. But, we still need to check the effects of insertion and deletion errors of the punctuation prediction on break prediction. Table 5 displays the co-occurrences of hit, insertion and deletion of major PM with the four break types for the training set of Treebank. Interestingly, the generated punctuations hit with major PMs were more likely to correlate with *B*4s while those of deletions were more likely to correlate with *B*3s. This implies that the generated punctuation has some ability to disambiguate *B*3 and *B*4. On the other hand, the insertions of

generated punctuations were likely to co-occur with *B*3 so as to provide useful information in the prediction of non-PM prosodic phrase boundary.

Fig. 4 displays the histogram of punctuation confidence corresponding to various types of word juncture. It can be clearly observed that the values of punctuation confidence were higher as target breaks are longer in pause duration. This shows that the punctuation confidence can be used to help to disambiguate various break targets on both PM and non-PM word junctures.

Table 4: Experimental results of punctuation generation

	precision	recall	F-score
ASBC4.0	89.7%	84.6%	87.1
Treebank - training set	86.5%	81.1%	83.7
Treebank - test set	86.9%	84.7%	85.8

Table 5: Co-occurrences of the predicted punctuation error types and the four break types.

	NPB	*B*2-2	*B*3	*B*4
Hit	0	0	1,418	2,049
Insertion	173	56	363	7
Deletion	2	0	649	238

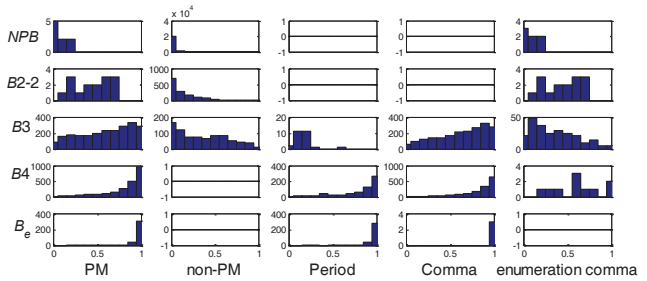

Fig. 4: Histograms of the punctuation confidence corresponding to various types of word juncture. X-axis and Y-axis represent the punctuation confidence and sample count, respectively.

In the break prediction experiment, three schemes were compared, including (1) break prediction using only PMs as input features (referred to as *PM_only*), (2) break prediction using the basic contextual linguistic features described in Subsection 3.2 (referred to as *Basic*), and (3) break prediction using the basic contextual linguistic features, predicted punctuation and punctuation confidence (referred to as *Basic+proposed*). The 2-previous and 2-following word lengths and word POSs, i.e. m=2 and n=2, were empirically set. The numbers of hidden nodes of MLP for the three schemes were all set to be 90. Table 6 displays the precision, recall and F-score for the three target break types obtained by the three schemes. As shown in the table, the *Basic+proposed* scheme outperformed the other two schemes. The results confirmed the usefulness of the proposed linguistic features of predicted punctuation and punctuation confidence.

Table 6: Precision (%)/recall (%)/F-score of the break prediction using various linguistic feature sets.

	*B*2-2	*B*3	*B*4
PM_only	0.0/0.0/0.0	39.1/40.8/39.9	90.2/19.9/32.6
Basic	48.2/13.9/21.6	46.2/36.1/40.5	73.6/62.9/67.8
Basic+proposed	52.0/16.2/24.7	49.5/39.9/44.2	76.3/65.6/70.5

5. CONCLUSIONS

Two statistical linguistic features, predicted punctuation and punctuation confidence, are introduced in this paper to help to improve the performance of prosodic break prediction. The effectiveness of the proposed approach was confirmed by the experiment on the Treebank speech corpus. In future works, it is worthwhile to incorporate the punctuation confidence with the durational information of prosodic units (i.e., PW, PPh, BG/PG) for further improvement on prosodic break prediction. Conducting a formal listening test on TTS to further evaluate the proposed method is also worth doing.

ACKNOWLEDGEMENT

This work was supported by NSC of Taiwan under contracts NSC98-2221-E-009-075-MY3, NSC99-2221-E-009-009-MY3 and NSC99-2622-E-009-005-CC2. The authors want to thank Academia Sinica, Taiwan for providing the Treebank Corpus and the Academia Sinica Balanced Corpus of Modern Chinese V.4.0.

REFERENCES

[1] M. Ostendorf and N. Veilleux, "A hierarchical stochastic model for automatic prediction of prosodic boundary location," Comput. Linguist. 20, pp.27–52 (1994).

[2] H.-J. Peng, C.-C. Chen, C.-Y. Tseng, and K.-J. Chen, "Predicting prosodic words from lexical words—A first step towards predicting prosody from text," ISCSLP 2004, pp.173–176.

[3] M. Chu and Y. Qian, "Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts," Computat. Linguist. and Chinese Language Processing, 6, pp.61-82 (2001).

[4] D.-W. Xu, H.-F. Wang, G.-H. Li, and T. Kagoshima, "Parsing hierarchical prosodic structure for Mandarin speech synthesis," ICASSP 2006, vol.1, pp.14–19.

[5] A. W. Black and P. Taylor, "Assigning phrase breaks from part-of-speech sequences," Eurospeech 1997, pp. 995–998.

[6] Z. Sheng, J.-H. Tao, and D.-L. Jiang, "Chinese prosodic phrasing with extended features," ICASSP 2003, vol.1, pp.492–495.

[7] J.-F. Li, G.-P. Hu, and R.-H. Wang, "Chinese prosody phrase break prediction based on maximum entropy model," Interspeech 2004, pp. 729–732.

[8] D.-W. Xu, H.-F. Wang, G.-H. Li and T. Kagoshima, "Parsing Hierarchical Prosodic Structure for Mandarin Speech Synthesis," ICASSP2006, pp.I-I, May 2006

[9] Y.-Q. Guo, H.-F. Wang and Josef Van Genabith, "A Linguistically Inspired Statistical Model for Chinese Punctuation Generation," ACM Trans. on Asian Language Processing, 9(2), 2010.

[10] C.-Y. Tseng, "Mandarin speech prosody: issues, pitfalls and directions," EUROSPEECH-2003, pp.2341-2344.

[11] C.-R. Huang, K.-J. Chen, F.-Y. Chen, Z.-M. Gao, and K.-Y. Chen, "Sinica Treebank: Design criteria, annotation guidelines, and pn-line interface," Proceedings of the Second Chinese Language Processing Workshop 2000, pp.29–37.

[12] C.-Y. Chiang, S.-H. Chen, H.-M. Yu, and Y.-R. Wang, "Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech," J. Acoust. Soc. Am. 125, No. 2, pp. 1164-1183 (2009).

[13] K.-J. Chen and C.-R. Huang, "Part of speech (POS) analysis on Chinese language," CKIP Technical Report No.93-05, Institute of Information Science, Academia Sinica, Taiwan, R.O.C., 1993.

[14] Available on http://www.aclclp.org.tw/use_asbc.php

[15] Available on http://crfpp.sourceforge.net/