PARSING HIERARCHICAL PROSODIC STRUCTURE FOR MANDARIN SPEECH SYNTHESIS

Dawei Xu, Haifeng Wang, Guohua Li*, Takehiko Kagoshima*

Multimedia Laboratory, Corporate Research and Development Center, Toshiba Corporation,
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki-shi 212-8582, Japan
*Toshiba (China) Research and Development Center,
5/F., Tower W2, Oriental Plaza, No.1, East Chang An Ave., Beijing, 100738, China

ABSTRACT

In Mandarin prosody synthesis by means of hierarchical prosodic structure, the naturalness of the output is reliant largely on the parsing of the prosodic structure. We propose a machine learning approach to improve prosodic structure parsing in cases where full syntax parsing is neglected due to considerations concerning practicality. The novel aspect of our approach is the new attribute in the input vector, which is named connective degree and calculated from the occurrence rate of the punctuation marks between Chinese characters by referring to a large text corpus. The results of experiments show that connective degree yield makes a remarkable contribution to parsing of hierarchical Mandarin prosodic structure.

1. INTRODUCTION

1.1. Prosodic unit and lexical unit in Mandarin Chinese

In written language, a sentence contains smaller units, such as phrases and words. Similarly, spoken language is also composed of hierarchical prosodic units. From the viewpoint of speech synthesis, research into prosody control in the scheme of hierarchical prosodic structure is undertaken in order to improve the naturalness of the synthesized speech. Since the hierarchical prosodic structure can be expressed by a series of nested prosodic unit boundaries, parsing the hierarchical prosodic structure is synonymous with predicting the boundaries of the hierarchical prosodic units.

1.2. Definition of PW/PP/IP

As proposed by Cao [1], prosodic word (PW), prosodic phrase (PP) and intonational phrase (IP) are the three prosodic units utilized in the prosodic scheme for our Mandarin speech synthesis system. These three prosodic units are in a hierarchical relation. An utterance can contain several IPs, an IP can contain several PPs, and a PP can contain several PWs, respectively.

An IP is a group of words with an obvious pause at its end. Fundamental frequency at the head of the succeeding IP is reset much higher than at the end of the preceding IP. At the same time, duration of the last syllable in an IP is often lengthened.

At the PP boundary, the reset of fundamental frequency and the lengthening of the last syllable can also be observed, but no obvious pause can be perceived.

PW is the shortest of these three prosodic units. It can contain one or several lexical words (LWs), but it is spoken as a whole unit without perceptible difference at the internal LW boundaries.

Feng [5], Cao [1] and Chu [2] indicate that most Mandarin PWs have disyllabic or trisyllabic forms. Based on this observation, only the merging of monosyllabic LW with its surrounding LWs is considered in the prosodic unit scheme. On the other hand, dividing long LWs with more than three syllables into small disyllabic or trisyllabic prosodic words may be a difficult task even for a human being. Taking these two factors into account, a polysyllabic LW is ignored in this study, unless it is merged with a monosyllabic LW. A polysyllabic LW will neither be merged with other polysyllabic LWs nor be divided into smaller prosodic words.

In Mandarin, punctuation marks such as commas, colons or semicolons almost always indicate pauses in utterances [2]. However, since a pause is not necessarily indicated with a punctuation mark, IP boundary prediction is essentially the same as pause prediction at boundaries without punctuation marks.

2. FACTORS RELATED TO THE GENERATION OF PROSODIC UNIT

In spoken language, prosodic unit generation is related to two principal factors. In this paper, the first one is called the dependency relation within words. The greater the dependence of words on one another, the less the possibility of the words being separated by a prosodic unit boundary. The second factor is called the prosody likelihood. The number of syllables in a prosodic unit usually distributes in some specific probability distribution.

2.1. Dependency between words

Dependency between words reflects both the syntactic and semantic information in the sentence. The correct information of dependency between words requires a high-level understanding of the sentence. However, at present the computer linguistics approach is incapable of attaining a full understanding of a sentence, and its inadequacy is particularly evident in the case of application systems with limited computing resources. In this study, a shallow parsing approach is used to acquire the information related to the dependency between words.

2.1.1 Definition of connective degree (CD)

In this study, a parameter named connective degree (CD) is calculated for every word boundary to show how often the boundary is not separated by a punctuation mark. This degree is a

posterior probability and it is counted by referring to a large text corpus. The bigger the CD is, the lower the probability that the word boundary is indicated by a punctuation mark, and therefore, the neighboring words are more dependent on one another. Therefore, the bigger the CD for a word boundary is, the lower the probably that the word boundary is a prosodic unit boundary.

The CD between the k_{th} Chinese word and the preceding one is calculated according to their neighboring Chinese characters by the following formula:

$$D_k = \frac{1}{3}\sum_{i=0}^{2} P_i(s^{(j\_k-i)}) \quad (1)$$

in which,

$s^{(j\_k-i)}$ is the sequence of 2 Chinese characters with indices *j_k-i* and *j_k-i+1*, where *j_k* is the index of the first Chinese character of the k_{th} word, and *i* stands for the index of the Chinese character boundary for the head, middle or tail with the value 0, 1 or 2;

$P_i(s^{(j\_k-i)})$ is the posterior probability for a punctuation mark to appear at the boundary *i* of Chinese character sequence $s^{(j\_k-i)}$, and it is calculated from a text corpus by the formula (2).

$$P_i(s) = -\log \frac{C_{punc}(s,i)}{C(s)} \quad (2)$$

in which,

s is a sequence of 2 Chinese characters;

$C(s)$ is the count of Chinese character sequence *s* in the text corpus;

$C_{punc}(s, i)$ is the count of Chinese character sequence *s* with a punctuation mark at the boundary *i*; If $C_{punc}(s, i)$ is 0, $P_i(s)$ is set to a large constant number.

As a concept, it should be easier to understand if we define D_k according to $P_i(s)$ that is calculated from a sequence of words rather than that of Chinese characters. However, the number of Chinese words is much larger than the number of Chinese characters. The former can easily exceed 100,000 in an unrestricted application, but the latter is much smaller. For example, the number of Chinese characters is 6,763 in Chinese character set GB2312. Considering again that $P_i(s)$ is calculated from a sequence, the size of the table for $P_i(s)$ is tremendously smaller when we calculate $P_i(s)$ from a sequence of Chinese characters rather than that of words. Meanwhile, the relation between character and word in Chinese is different from other languages such as English. A Chinese character links highly to the meaning of a Chinese word, but an English letter contributes little to the meaning of an English word. Take into account the quantitative and linguistic features of Chinese characters, the formula for connective degree in this paper is based on Chinese character rather than word.

2.2. Prosody likelihood

Prosody likelihood is defined as how often a character boundary can be a prosodic unit boundary according to the position information of the boundary. This is also the factor considered in the studies of [2] and [3], and it is named prosody constraint in [3]. Since it is shown in terms of probability, we prefer to call it likelihood instead of constraint.

C4.5 program [7], which is a well-known tool for decision tree, is used in this paper to train decision rules for parsing Mandarin prosodic structure. In this C4.5 approach, prosody likelihood is reflected in the way that position information is a part of the conditions of the C4.5 decision rules for prosodic unit boundaries.

3. PREDICTION OF PROSODIC UNIT BOUNDARY BY C4.5

3.1 Available attribute set for prosodic unit prediction

Three sets of attributes are used for predicting the prosodic unit boundary. The first set is Part-of-Speech (POS) set, which a commonly used attribute in the related research. For example, P. Taylor *et al* [8] proposed the POS sequence method to predict pause position for English speech synthesis. POS related attributes are also used in studies on prosodic unit boundary prediction for Mandarin [2] [3] [5]. The second set is named position set, which includes the attributes relating to position or text length. With this attribute set, we can construct rules related to prosody likelihood. The third set is the information related to CD. The bigger CD is, the lower the probability that the word boundary is indicated by a punctuation mark, and therefore, the lower the probability that the word boundary is a boundary of a prosodic unit.

The attributes for prosodic unit prediction are shown in Table 1 in detail. Among these attributes, POS set includes numbers 1 to 3, position set includes numbers 4 to 8, and CD set includes numbers 9 to 12.

Table 1. The attributes being used to predict prosodic unit by C4.5

ID	Attribute
1	Part-of-Speech (POS)
2	POS of the preceding word
3	POS of the succeeding word
4	Number of syllables from the previous pause
5	Number of syllables to the next pause
6	Number of syllables in the lexical phrase, i.e., the phrase marked by punctuation marks
7	Number of syllables in the lexical word before the boundary
8	Number of syllables in the lexical word after the boundary
9	Connective degree (CD)
10	The difference between the current CD and that at a syllable before
11	The number of syllables before which CD is less than the current CD
12	The number of syllables after which CD is less than the current CD

3.2 Three approaches of machine learning prediction

Instead of a baseline approach in which PW/PP/IP boundaries are predicted simultaneously by a whole decision tree, Chu *et al* [2] argue that a bottom-up approach yields better prediction performance, in which prosodic units are predicted gradually from smaller to larger ones with respective decision trees. In her approach, however, the larger the prosodic unit is, the smaller is the amount of training data available for the construction of its decision tree. Only the data at the boundaries for the target prosodic units and those for its 1-level-smaller prosodic units are used to construct a decision tree. Thus, the data sparseness problem is easily induced in machine learning for the larger prosodic units, and particularly so in the case of a smaller speech corpus. In this paper, we investigate the approach in which the data at all LW boundaries are used to construct the respective decision trees. Meanwhile, the baseline approach is also investigated for the purpose of comparison. So, the prediction performances of 3 approaches are compared in this paper. These 3 approaches are described in greater detail below.

3.2.1 Baseline approach

The baseline approach for predicting prosodic unit boundary is a one-step way of predicting PW, PP and IP boundaries with a single decision tree. In the case of C4.5, this approach uses an all-in-one C4.5 tree to predict these prosodic unit boundaries as a whole. This is an easily achievable approach for such a task.

3.2.2 Bottom-up hierarchical approach

This is the approach proposed by Chu *et al.* [2]. This approach considers the hierarchical relationship between PW, PP and IP, and adopts a bottom-up way of predicting PW, PP and IP boundaries step by step. Concretely, there are three prediction steps in this approach. The first step is to predict boundaries of PW, PP and IP from all LW boundaries. The boundaries passed for PW, PP or IP are used in the second step as input. The second step is to predict boundaries of PP and IP from the boundaries achieved in the first step. The PP and IP boundaries passed in the second step are used as input for the third step, in which only IP boundaries are predicted. In the final output, PW boundaries are those that passed the first step but failed to pass the second step, and PP boundaries are those that passed the second step but failed to pass the third step. The results of the experiment of Chu et al. indicate that the predictive performance of a bottom-up approach is superior.

3.2.3 Sifting hierarchical approach

The author considered a third approach in which PW, PP and IP are also predicted hierarchically. But the difference from the bottom-up approach is that in every prediction step from a lower layer to a higher layer, the C4.5 tree is trained from the data at all the LW boundaries.

In the bottom-up approach, the training data for a C4.5 tree contains only the data at 1-level-smaller prosodic units. The training data for predicting PP boundary are at least at a PW boundary, and those for predicting IP boundary are at least at a PP boundary. However, because the PW boundaries and PP boundaries are much less than the LW boundaries in the same speech corpus, the C4.5 tree to predict PP or IP is trained by less data than that to predict PW. This gives rise to the problem of data sparseness in training the decision trees for PP or IP.

In training using the sifting approach, the input data for all C4.5 trees consist of all LW boundaries. Therefore, a higher precision rate can be expected from the sifting approach than from the bottom-up approach.

4. EXPERIMENT

4.1 Materials

CDs are calculated for every 2 GB2312 Chinese characters from a large text corpus, which contains about 13 million sentences, 44 million punctuation marks, and 689,790 / 2,334,800 / 656,550 different Chinese character pairs with punctuation marks before / within / after them, respectively. The CDs for unseen Chinese character pairs are handled by add-one smoothing [6].

The speech corpus contains altogether 1,800 sentences, and the prosodic unit boundaries for the sentences are manually labeled. Altogether 3,117 PW boundaries, 874 PP boundaries and 2,732 IP boundaries are labeled. The guidelines to label the prosodic unit boundary are the same as those mentioned in section 1.2.

Since a LW boundary with a comma, colon or semicolon is almost accompanied by a pause, hence force such a LW boundary nearly equals to an IP and PP boundary, only those LW boundaries not accompanied by such punctuation marks are used in the experiments.

4.2 Experiment results

Two experiments were conducted in this study. The first one was to investigate the validation of CD for predicting IP boundary by evaluating various decision trees from different attribute sets. The second one utilized the full attribute set to compare the performances among various prediction approaches, namely, the baseline approach, the bottom-up hierarchical approach and the sifting hierarchical approach.

4.2.1 Experiment 1: Availability of CD in IP boundary prediction

In this experiment, CD attribute set is compared with POS attribute set and position attribute set in terms of the contribution to the prediction of IP boundary. Table 2 shows the results for precision rate (p), recall rate (r) and F score, where F=2*p*r/(p+r). F score reflects an overall evaluation of precision and recall rates[9]. The results listed in Table 2 are the average of 10 experiments of cross validation.

From Table 2, we can see that C4.5 tree with CD attribute set (the 2nd row) had about the same performance as those with POS attribute set and position set. Even when only CD itself is used (the 1st row), C4.5 tree achieved comparable performance. This indicates that CD is an effective parameter for pause prediction. Meanwhile, by adding CD set to the POS set and position set, the C4.5 tree achieved only about 3.0% higher performance in terms of precision rate and about 1.3% in terms of recall rate, but it is considered to be a remarkable contribution in light of the experiment result in [2], in which the manually checked syntactic information yields an approximately 4.4% improvement in precision rate. Also, in her study, considering that a computational syntactic parser cannot always provide reliable output information, the investigation on the effect of syntactic information was only conducted as a pilot experiment.

Table 2. The precision rate, recall rate and F score for IP prediction with different attributes set

Attribute set	Precision [%]	Recall [%]	F [%]
Only CD	61.07	23.25	33.90
All CD related attributes	63.29	29.08	39.80
POS set	66.16	26.16	37.43
Position set	60.48	7.99	14.07
POS set plus position set	67.82	44.40	53.58
Full set (CD, POS & position)	70.80	45.67	55.48

Table 3. The precision rate, recall rate and F score for PW/PP/IP in different prediction approaches

		Precision [%]	Recall [%]	F [%]
Baseline (all-in-one)	PW	75.59	82.50	78.89
	PP	25.00	4.17	7.14
	IP	53.80	41.82	47.06
Bottom-up hierarchical	PW	74.64	86.36	80.07
	PP	29.63	11.11	16.16
	IP	62.79	49.09	55.10
Sifting hierarchical	PW	73.18	88.16	79.98
	PP	8.86	9.72	9.27
	IP	75.90	28.64	41.58

4.2.2 Experiment 2: Performances of the three prediction approaches

The performances of the three approaches for prosodic unit prediction are listed in Table 3.

The three approaches achieved about the same performance for PW prediction. For PP prediction, all three approaches failed to attain a satisfactory result. For IP prediction, the sifting hierarchical approach achieved the highest precision rate, and the bottom-up hierarchical approach achieved the highest recall rate.

Since IP boundary is always accompanied by a pause in synthesized speech, the precision rate for IP is considered to have a higher priority than the recall rate. From this viewpoint, the bottom-up hierarchical approach has a problem in that it outputs too many IP boundaries, which may be attributable to the data sparseness problem in the machine learning steps, especially in that from PP to IP. In this set of machine learning, the data at the input side are the PP boundaries and IP boundaries. In the current data set, since there are more IP boundary data than PP boundary data, the trained decision tree tends to output IP boundaries more easily, and consequently, the precision rate for IP boundaries is lower.

We proposed the sifting approach to improve the precision rate for IP boundary. In this approach, all of the C4.5 trees to predict IP, PP or PW are constructed with all LW boundary data, thus precluding deterioration of reliability due to sparse data. Consequently, an output IP boundary is more highly qualified in this approach because it has to pass three reliable decision trees for PW, PP and IP, as if it were sifted repeatedly. On the other hand, the decision trees for PW, PP and IP in the bottom-up approach are found to become progressively less reliable because of the significantly reduced amount of training data. As the result, the decision tree for IP boundaries in the bottom-up approach performs poor in the precision rate.

5. DISCUSSION

The values listed in the 2 experiments are not as high as those reported in some related studies, such as [3] and [5], but those two studies didn't indicate whether the boundaries accompanied by punctuation marks were used in training and evaluation. But as reported by Chu [2], 99.6% of boundaries accompanied by commas, colons or semicolons are IP boundaries. In her study, the evaluations are conducted distinctly for the boundaries with or without punctuation marks, where the precision rate for IP boundaries in the best approach is 71.12% in the experiment ignoring punctuation marks [2].

In Experiment 1, it is shown that CD makes a remarkable contribution to the prediction of IP boundaries. At the same time, since the data needed to calculate CD can be acquired by scanning a large text corpus, CD can be determinately calculated not only in the phase of training data preparation, but also in the phase of prosody generation. This insures that the observed improvement in the experiment can definitely be reproduced in the prosody generation. This is a great advantage over the syntactic information with which it is difficult to achieve a reliable output for syntactic information in the phase of prosody generation. Since it is calculated in regard to the neighboring Chinese characters, CD is also valuable in that it reveals not only the syntactic information accompanying the Chinese characters, but also the semantic information to some extent. Meanwhile, POS or other linguistic units may be also reasonable when the approach is utilized in the speech synthesis for other languages or for a large footprint application.

The performance on PP boundary prediction in the Experiment 2 is quite lower than that for PW or IP boundaries. One of the reasons may be that a PP boundary is quite difficult to be distinguished from a boundary caused by emphasis. Although pitch reset is a main character of PP, pitch reset can be also caused by an emphasis. In this way, the boundary for PP may be just a boundary before an emphasis. Because an emphasis is considered as difficult to be predicted without semantic analysis, the performance on PP prediction is low in this machine learning approach.

The sifting hierarchical approach yields a much higher precision rate for IP boundaries than does the bottom-up hierarchical approach. However, a side effect of the sifting hierarchical approach is that it performs worse for the recall rate. The same relation is even observed in the F score. However, the F score is not yet a perfect overall evaluation score for the IP boundary prediction, considering that the precision rate of IP boundary is practically more important than recall rate. An inappropriate inserted pause, therefore an unnecessary IP boundary, may empirically cause more deterioration in the synthesized prosody than a missing pause. A possible solution is to accord a higher weight to the precision rate than to the recall rate in the calculation of an overall evaluation score, but this is left as a subject for future work as it requires further practical investigation to determine such a weight.

6. CONCLUSION

A novel text feature named connective degree is introduced in the machine learning approach to predict Mandarin prosodic unit boundaries. Experiments show that connective degree provides effective information for prosodic unit prediction. Furthermore, it is argued that the sifting hierarchical approach achieves the highest performance in terms of precision rate for IP boundary prediction. The approach proposed in this paper is easy to realize in Text-to-Speech system, and it is also considered effective for other languages.

References

[1] J. Cao, "Rhythm of Spoken Chinese – Linguistic and Paralinguistic Evidences," *Proc. of ICSLP 2000*, Beijing, pp. 357-360, 2000.

[2] M. Chu, and Y. Qian, "Locating Boundaries for Prosodic Constituents in Unrestricted Mandarin Texts," *Computational Linguistics and Chinese Language Processing,* vol. 6, pp. 1-22, 2001.

[3] H. Dong, J. Tao, and B. Xu, "Chinese Prosodic Phrasing with a Constraint-Based Approach," *Proc. of Interspeech 2005*, Lisbon, pp. 3241-3244, 2005.

[4] S. Feng, *Interactions between Morphology Syntax and Prosody in Chinese*, Beijing University Press, Beijing, 1997.

[5] J. Li, G. Hu, W. Zhang, and R. Wang, "Chinese Prosody Phrase Break Prediction Based on Maximum Entropy Model", *Proc. of Interspeech 2004*, Jeju Island, 2004.

[6] C. Manning, and H. Schiitze, *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, MA, 1999.

[7] J. R. Quinlan, *C4.5 Programs for Machine Learning*, Morgan Kaufmann publishers, San Manteo, CA, 1993.

[8] P. Taylor, and A. Black, "Assigning Phrase Breaks from Part-of-Speech Sequences", *Computer Speech and Language,* vol. 12, pp. 99-117, 1998.

[9] M. W. Wang, J. Hirschberg, "Automatic classification of intonational phrase boundaries", *Computer Speech and Language*, Vol. 6, pp.175-196, 1992.