RC22498 (C0206-002) June 24, 2002
Other

IBM Research Report

Syntactic and lexical constraint in prosodic segmentation and grouping

***Jianfen Cao **Weibin Zhu**

**IBM Research Division
China Research Lab
2/F No. 7, 5th Street
Shangdi, Haidian District
Beijing 100085, P.R. China

*Institute of Linguistics
Chinese Academy of Social Sciences

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Syntactic and lexical constraint in prosodic segmentation and grouping

*Jianfen Cao **Weibin Zhu

**IBM China Research Lab
*Institute of Linguistics
Chinese Academy of Social Sciences
Email:jianfencao@hotmail.com

ABSTRACT

This paper tries to discuss the interrelation between prosody and syntax by clarifying some syntactic constraints in Chinese prosodic segmentation and grouping. The main attention will be paid to search for (1) possible correlation between prosodic breaks and syntactic construction; (2) possible correlation between prosodic breaks and POS; and (3) the role of syntactic and lexical information in prosodic word chunking. Accordingly, an algorithm for the prediction of prosodic structure based on these information could be formed later on.

I. INTRODUCTION

Prosodic segmentation and grouping is referred to prosodic structure of speech, and has been a hot topic in the field of speech processing. TTS, for example, still suffers from somewhat unnaturalness, though great progress has been made in this field. A lot of specific problems seem to be caused by the dislocation in prosodic hierarchy. At the same time, the accuracy in speech recognition is difficult to get enhanced, which seems also to be caused by lacking uses of the knowledge in prosodic structure. Consequently, it is true to say " that both for speech synthesis and speech recognition, prosody is the area in which the most progress needs to be made before such technology can be used as an acceptable replacement for human speech" [4].

Because of the development of TTS, prosody generation has become the most important part to enhance the naturalness of synthetic speech. It is especially necessary to predict a prosodic hierarchy according to grammatical information from the text, since it will be of great benefit to the generation of prosody automatically at the back-end in a system. However, it is well-known that prosodic and grammatical structure do not always directly map each other. Consequently, it is a critical issue to clarify the interrelation between prosody and syntax both for theoretical approach in phonetic science and applied research in speech technology.

The main task is to segment syllable sequence into proper units and then organize them into correct prosodic layers based on text analysis. This issue is not easy to deal with. It is because that in human speech production, different prosodic units are well organized in a hierarchy, but such information is almost completely lost in text. Therefore, people have to seek a proper way to recover prosodic structural information and annotate them onto the text. Actually, here the underlying challenge is related to the interrelation between syntax and prosody.

Studying on the correlation between prosody and syntax is a classical and universal subject conducted in many languages. To achieve this goal, a lot of efforts have been made in related fields[6, 3, 8]. Recently, in China, there were also several methods [5,10] that make use of features such as part of speech(POS), position of words, cue words, punctuation and so forth to predict the prosodic breaks, but it is far from realizing the whole picture, and further study is needed.

Making prosodic segmentation and grouping at the front-end of Chinese TTS is a particularly difficult task due to following factors.

First of all, many researches have found that prosodic hierarchy is not always consistent with syntactic hierarchy in a language; therefore, it is impossible to make prosodic segmentation and grouping by directly mapping from syntactic structure. This is a common issue revealed in different languages.

The second, but more important factor, is that Chinese has many morpho-syntactic features quite different from those of western languages on one hand, such as monosyllabic structure in morpheme and its flexibility and poly-synthetism in word-formation; while on the other hand, the speech units in natural Chinese is also well-organized as a hierarchy, instead of a discrete linear alignment. According to Cao [1, 2], Tseng [11] and Qian et al [9], the prosodic hierarchy of Mandarin Chinese consists of at least three layers, i.e., prosodic word, prosodic phrase and intonation phrase. It means that, the monosyllabic written form in Chinese is completely separated from the spoken form. Consequently, how to gather the monosyllables into prosodic chunks has become a very thorny problem .

In addition, we neither can expect to determine prosodic words directly according to lexical words, even though the monosyllabic morphemes have been combined into lexical form in the dictionary. It is because of that the lexical word may be smaller or larger than a prosodic word, and in fact, there is only 70.71% of lexical words which are identified with prosodic words in real speech[7], not to mention the fact that the definition and segmentation of lexical word itself is originally a controversial issue in Chinese linguistics. Consequently, how to make prosodic segmentation and grouping at the front-end, so that to build a prosodic hierarchy automatically at the back-end of TTS system, has become an urgent subject raised in both fields of phonetics and speech technology.

Bearing this idea in mind, the present study tries to discuss this issue by clarifying the syntactic constraints occurred in prosodic segmentation and grouping in Mandarin Chinese. The main attention will be paid to following aspects: (1) exploring possible correlation between perceived pause (i.e., prosodic break) and syntactic construction; (2) searching for

possible correlation between perceived pause and the part of speech; (3) observing the role of syntactic and lexical information in prosodic word chunking.

II. EXPERIMENTAL MATERIALS AND METHODS

This study is based on the observation and analysis of two sets of sentences: the first set was selected randomly from a discourse corpus that is uttered by multi-speakers. Another set was extracted randomly from a large scale speech corpus of sentences read aloud by a male speaker. Among these materials, the length of sentence is 9-20 syllables on average, and the longest one consists of 41 syllables, so all of them are involved in the prosodic segmentation and grouping taking place within a sentence.

Experimental investigation and analysis includes four steps: (1) Text analysis, in which grammatical word segmentation, part of speech annotation and syntactic parsing of each sentence were conducted. (2) Perception test, by which breaks and their strength were labeled, so that to provide prosodic segmentation information and form as a prosodic hierarchy. (3) Acoustic analysis, through which the suprasegmental features related to the prosodic hierarchy, like boundary marked pitch movement and temporal variation, were obtained. These information are very helpful in determining the prosodic hierarchy objectively. (4) At last, to conduct a comparative analysis of the data obtained from (1) to (3), i.e., to look into possible correlation between prosody and syntax, through the comparing of the perceived pause and the results of syntactic parsing.

III. RESULTS AND DISCUSSION

Perceptually, prosodic hierarchy is roughly identified with perceived pause (i.e., break), so the hierarchy can be represented through pausing distribution, and we can explore the prosody-syntax interrelation by investigating the relationship between pausing distribution and syntactic construction.

Generally, there are three degrees of pause that can be perceived by normal listening, but there is a fourth degree of pause, a mini-pause, could be perceived by a careful perception test. Usually, such mini-pause is being ignored. However, it has been revealed that such mini-pause is even more important for the improving of TTS naturalness, since whether the mini-pause setting is proper or not will obviously affect not only the naturalness, but also the sound quality and intelligibility to some degree. Therefore, we try to observe four degrees of pause in the present study, that is p1-p4 from major to minor. A part of experimental results are listed in the Table on the next page of this paper.

3.1Distribution of pausing and syntactic construction

According to the results obtained from this study, we find that there does exist some correlations between pausing distribution and syntactic construction. The main points can be summarized below.

3.1.1Distribution of the first degree of pause (hereafter P1)

The first degree of pause is major break perceived from perception test, and the majority of this degree of pause occur between the subject and predicate of a sentence, especially in a relatively balanced sentence. Whereas, the P1 in an unbalanced sentence is usually occurred (1) between a prepositive adverbial modifier and the subject; (2) between the clauses in a complex sentence and (3) between different modifiers in the complex attribute or adverbial modifier.

3.1.2 Distribution of the second degree of pause (P2)

The data obtained here show that the majority of P2 is occurred (1) between different modifiers in the complex attribute or adverbial modifier ; (2) between comment and object, including verb and object, pronoun and object, or the copulative verb and the predicative; (3) between the subject and predicate of an unbalanced sentence and (4) between the modifier and the part being modified.

3.1.3 Distribution of the third degree of pause (P3)

Most of the P3 is distributed (1) between the modifier and the part being modified; (2) between the comment and object, including verb and object, pronoun and object, or the copulative verb and the predicative.

3.1.4 Distribution of the fourth degree of pause (P4)

The distribution range of P4 is mainly located (1) between the words in a word compound or mini-phrase without auxiliary word “de”; (2) after the auxiliary word “de” in a noun phrase or word compound.

In summary, the situations described above indicate that the major prosodic breaks P1 in natural speech do correspond to most but not necessarily all major syntactic boundaries. While the minor prosodic breaks P4 usually occur within a small size syntactic phrase or word compounds. Generally, the P2 and P3 function as the most powerful boundary marker of intermediate prosodic chunks, mainly occur in the syntactic levels lower than that of P1 but higher than that of P4, but may be some overlap between P2 and P3 as well as P1, especially in an unbalanced sentence.

3.2 Distribution of pausing and lexical information

According to the results obtained here, some correlation also found between pausing distribution and lexical information. That is, a pause, longer or shorter, is likely occurred in the following cases:

(1) After a location name or person name, especially after a foreign name;
(2) After a time or quantitative word;
(3) After a polysyllabic verb and followed by a polysyllabic structure;
(4) After a position word and directional verb served as a suffix and followed by a polysyllabic structure;
(5) Before and /or after a coordinative conjunction, connecting two complex constituents, and the pause before the conjunction word is usually stronger than that after the conjunction word;
(6) Before and / or after a turning conjunction connecting two complex constituents;
(7) After an auxiliary word followed by a polysyllabic structure;
(8) After a disyllabic adverb followed by a polysyllabic structure.

3.3 The role of syntactic and lexical information in prosodic word chunking

Grammatically, there are a lot of monosyllabic words in Chinese, but according to the information obtained from this

Table. Pause distribution among various syntactic components in the first set of materials (total 57 sentences)

Syntactic components	P1	P2	P3	P4
Between(B/w) subject and predicate	23	11	3	9
B/w clauses	3			
B/w prepositive adverbial modifier and subject	5	3		
B/w coordinative objects or predicatives	2			
B/w verb/ preposition and object, or b/w linking verb and predicative	2	16	29	8
B/w modifier and the part being modified	2	10	48	3
B/w different modifiers	3	17		
B/w coordinate predicates	2			
B/w prepositive object and verb	1			2
B/w object and complement	1	2	2	
B/w subject and predicate of a clause		2	3	
B/w coordinate component 1 and conjunction		3	2	
B/w conjunction and coordinate component 2		1		2
B/w verb and complement		1	4	2
B/w shorter noun phrases			4	
B/w turning conjunction and the component being connected			2	
After auxiliary word “de” of a word compound with “de”				13
B/w the words of a word compound without “de”				47
B/w modifier and verb in a verb phrase				10
At the center of quadri-syllabic idioms				8
B/w former two numerals and later two numerals in year’s name, B/w numeral and quantity, or at the center of a compound numeral				8
B/w preposition and object in a shorter phrase				4
B/w the names of year, month or season				2
B/w “di” and numeral in a polysyllabic ordinal number				1

study, we found that such kind of words is likely to form a prosodic word by self-lengthening or attaching to another mono- or disyllabic constituent so as to gather into a foot in order to fulfill the requirement of speech prosody. However, such prosodic grouping must be under certain rules, which can be summarized as follows.

(1) Independently forms as a foot by self-lengthening when it is stressed in speech or when there is no any mono- or disyllabic constituent could be attached;
(2) To form as an independent foot by attaching with another one or two monosyllabic words when neighboring (before and / after) other monosyllabic word;
(3) To form as a super-foot by pre- or post- attaching to a standard foot, i.e., a disyllabic constituent, when it is unstressed in speech.

In addition, the prosodic grouping process described above must be under certain syntactic and lexical constraint.

Specifically, for example, if the monosyllable word is a noun, then the chunking process is usually achieved by pre-attaching a mono- or disyllabic adjective/pronoun/numeral/conjunctive/verb/another noun, and / or by post-attaching a suffix /auxiliary word “de”/position and direction word /another noun. Whereas, if it is a monosyllabic verb, then the chunking process will be achieved by pre-attaching a mono- or disyllabic adverb/auxiliary verb/interrogative pronoun/prefix /another verb, and / or by post-attaching a monosyllabic noun/adjective/ noun suffix/auxiliary word “de”/direction and position word /verb suffix / preposition.

Similarly, other parts of speech, like adjective, adverb, preposition and so forth, all have special lexical constraints in their chunking process. These situations tell us that each part of speech must be under certain lexical condition when it is combined with other word to chunk into a prosodic word, otherwise, it will serve as a prosodic word independently.

IV. SUMMARY

According to the data obtained from this investigation, we find that the main inconsistency between prosody and syntax is in their hierarchical strength. For example, the major syntactic boundary in a sentence should be located between the subject and predicate, while the major prosodic boundary in natural speech is often occurred at some lower syntactic level. However, the data also indicates that prosodic hierarchy is not completely inconsistent with syntactic structure. There do exist certain mapping relations between prosodic structure

and syntactic structure. It means that people can extract a lot of prosodic structural information from available syntactic information, and then, apply them for the prediction of prosody in speech processing. In the fact, the findings obtained from this investigation have been partially applied to certain TTS system, the test is in progress, but the preliminary result sounds satisfactory. Of course, this is just an initial effort, and much more work, especially for some statistic approach is needed. A possible algorism for the prediction of prosody based on text information in TTS is ongoing, it may be reported in later papers.

REFERENCES

[1]Cao, J., Acoustic-phonetic characteristics on the rhythm of Standard Chinese, The Proc. Of 4th National Conference on Modern Phonetics, Beijing, August, 1999.

[2]Cao, J., Rhythm grouping and speech timing, Proc. Of ICSLP'2000, Beijing, Oct. 17-20.

[3]Gee, J. P., Performance structures: A psycholinguistic and linguistic appraisal, Cognitive Psychology, 15, 1983.

[4]Hirst, D., Bel, B. and N. Campbell, Speech Prosody 2002 committee –Message, 2001.

[5]Lee, L., Structural features of Chinese language – Why Chinese spoken language processing is special and where we are", Proc. ISCSLP'98, Singapore, 1998.

[6]Lehiste, I., The timing of utterances and linguistic boundaries, JASA, 51, 1972.

[7]Lu, S., He, L. et al., A comparison between synthetic speech and natural speech of Chinese, ISCSLP'00, Oct. 13-15, 2000, Beijing.

[8]Mertens, P., Intonational grouping boundaries and syntactic structure in French, In ESCA Workshop on Prosody, Lund Working Papers, 41, 1993.

[9]Qian Y., Chu, M. and Peng, H., Segmentating unrestricted Chinese text into prosodic words instead of lexical words, ICASSP'2001.

[10]Shen, X. & Xu, B., A CART _based hierarchical stochastic model for prosodic phrasing in Chinese, Proc. of ISCSLP'00, Oct. 13-15, 2000, Beijing.

[11]Tseng, C., The interplay and interaction between prosody and syntax: evidence from Mandarin Chinese, Proc. Of ICSLP'2000, Beijing, Oct. 17-20.