Acoustic Correlates of Hierarchical Prosodic Boundary in Mandarin

Yang Yufang & Wang Bei

The Institute of Psychology, Chinese Academy of Sciences

{Yangyf; Wangb}@psych.ac.cn

Abstract

The aim of this paper is to present a systematical analysis of the acoustic correlates of hierarchical prosodic boundaries. This analysis is based on a large labeled corpus of Mandarin Chinese. The acoustic correlates include the lowest value of pitch and the duration of the silence. The prosodic structure is defined thanks to a perception experiment. The main results are : a) the declination of intonation in Chinese is achieved through the bottom line of the intonational contour ; b) the acoustic correlates of prosodic word boundaries are pre-boundary lengthening and a slight pitch reset of the bottom line of intonation ; c) the acoustic correlates of prosodic phrase boundaries and intonational phrase boundaries are a significant pitch reset of the bottom line of intonation and the insertion of a silence. Moreover, the higher the prosodic boundary is, the higher the extent of the pitch reset is and the longer the silence is. There is no significant difference on pre-boundary lengthening between the syllables on these two boundaries.

In conclusion, pre-boundary lengthening is the acoustic correlate of weak boundary. Pitch reset is that of medium boundary, while silence is that of strong boundary. The acoustic correlate of lower boundaries can also occur on larger boundaries, but the acoustic correlate of larger boundaries usually does not occur on lower boundaries.

1. Introduction

The aim of Text-to-Speech is to transfer written text to understandable and natural speech. The generation of good prosody is of the greatest importance to synthesized speech. The first step to prosody generation might be the assignment of a prosodic boundary index.

Although there is a general consensus in the literature on the hierarchical prosodic structure, it is not the case as far as the number and the types of the prosodic constituents are concerned. Nespor and Vogel (1978) proposed a 7-layer prosodic structure, that is, the syllable, the foot, the phonological word, the clitic group, the phonological phrase, the intonational phrase and the phonological utterance[1]. Wightman points out that the pitch accent phrase lies between the prosodic word and the intermediate phrase [4]. We can say that there is a general agrreement about the following prosodic constituents : the syllable, the foot, the prosodic word, the prosodic phrase, the intonational phrase and the sentence.

The main difficulty when we study the acoustic correlates of prosodic boundaries is to determine the location and the degree of prosodic boundaries. The paradigm of rating on scale in psychological experiment is a good solution [1][3]. De Pijper proved that untrained subjects can reliably determinate the degree of prosodic boundaries [1]. They also found that there is high agreement between the degree of the prosodic boundaries determined by a perception experiment and the theoretically predicted prosodic structure. A perception experiment offers several advantages : it is possible to construct straight away the hierarchical prosodic structure, the listeners can be naïve listeners and not linguists, thus avoiding a circular process [3]. It is generally agreed that the acoustic correlates of hierarchical prosodic boundaries are pitch reset, pre-boundary lengthening and silence. Streeter argues that pitch contour and duration are more important cues than amplitude in parsing ambiguous algebraic expressions [2]. Furthermore, duration has a larger range of effectiveness than pitch [2]. Swerts proposes that break, pitch range and the tone of the boundary are prosodic cues to speech structure [3].

The intonation of the sentence has been shown to chunk an utterance into consecutive phrases [3]. There are two ways of changing the pitch on the prosodic boundary : one is the discontinuity of the intonation line, the other is the pitch reset of the declined intonation line [1]. Though there is intonation declination in Chinese just as in other languages, since Chinese is a tonal language, the intonation of the sentence is combined with the pitch contours of the tones. To describe Chinese intonation, we should apply to the top and bottom line the intonational model which is described by the highest value of of pitch and the lowest value of pitch respectively. The low value of pitch reflects the integration of the rhythm constituent while the high value of pitch reflects the variation of pitch accents. The low value of pitch is lower on the big rhythm constituent, at the same time, the hierarchical structure of the prosodic constituents is also realized by the lowest value of pitch [5].

To indicate the degrees of the prosodic boundaries, duration is relevant as far as syllable lengthening, silence and foot lengthening are concerned. The results of Wightman's (1992) experiments are that the pre-boundary lengthening differs according to the type of prosodic boundary on the basis of standard duration and normalized speech rate.

All the experiments mentioned above were mainly based on a small set of experimental sentences. In addition, there is some disagreement about how the cues in the acoustic signal actually mark the boundaries [4]. Moreover, it is not clear precisely what boundaries are actually signaled. This kind of research is quite rare in Chinese, which maks it all the more difficult. We should also consider the characteristics of Chinese. In this paper, we used a large corpus labeled on the sentence level to analyze systematically the acoustic correlates of hierarchical prosodic boundaries, including pitch reset, pre-boundary lengthening and the duration of silence.

1. Method

1.1. Corpus

The corpus is composed of 500 sentences with 7940 syllables. A 24 year-old female professional radiobroadcaster read all sentences in a natural way. The speech rate is 3 to 4 syllables per second. The average length of sentence is 15 syllables with 4 to 5 syntactic layers.

1.2. Multi-level labeling

The prosodic labeling of the corpus is achieved thanks to a perception experiment using degree scaling. There are 24 participants who were born in Beijing. They are 20 to 23 years old without obvious hearing obstacles. After listening to the sentence twice, they are asked to give a degree for each prosodic boundary on a 4-degree scale, that is, strong, medium, weak and no boundary.

The acoustic labeling of the corpus includes the highest value of pitch, the lowest value of pitch, mean pitch, duration and silence using the Multispeech software.

1.3. Labeled prosodic structure

The agreement among participants on degree scaling is higher than 0.8. The prosodic boundary degree is 0-3, while the degree of 69% of the boundaries is lower than 0.05. The average degree of prosodic boundaries is 1.08 with a standard deviation of 0.81. Three kinds of highly agreed prosodic constituents are studied, that is, the prosodic word, the prosodic phrase and the intonational phrase. According to the perceived degree, the break index of the prosodic word boundary is 0.05-1, that of the prosodic phrase boundary is 1-2, and that of the intonational phrase boundary is higher than 2. We call the boundaries “no break”, “weak break”, “medium break” and “strong break” respectively.

1. Results and Discussion

1.1. Pitch reset on prosodic boundary

In Chinese, the minimum value of pitch is related to rhythm while the maximum value of pitch is related to pitch accent [5]. If we compare the two extreme values of pitch, the variation of the highest value occurs more freely on syllables with different degrees of accent [6]. Since the lowest value of pitch never occurs on syllables of tone one and neutral tone, they are excluded of calculation. Syllables at the end of sentences are also excluded. The low points of syllables of tone 2, 3 and 4 at different prosodic boundaries are shown in figure 1.

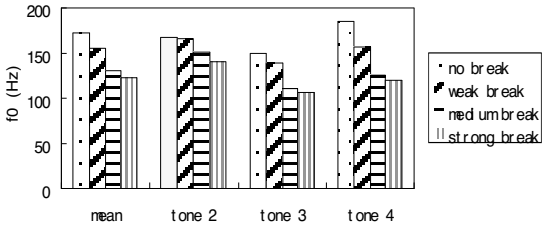

Figure 1. The minimum value of pitch on syllables of tone 2, 3 and 4 at different prosodic boundaries

It is shown in figure 1 that the lowest value of pitch decreases as the prosodic boundary degree increases. There is a significant difference among prosodic boundaries on the lowest value of pitch for all the tones ($F(3)=250$, $P<0.001$). It is also true for tone 2, 3 and 4. Furthermore the declination effect is more significant for tone 3 and tone 4 because the lowest value of pitch for these syllables is almost the minimum value of the pitch range box, so that the link between the lowest value of pitch and the prosodic constituent is based on the pitch range box. Table 2 shows the average minimum value of pitch on syllables before and after the different boundaries.

Table 2. Average minimum value of pitch on syllables before and after boundaries

	Before boundary	After boundary	Deviation
No boundary	171	156	-15
Prosodic word boundary	155	167	12
Prosodic phrase boundary	130	179	49
Intonational phrase boundary	122	187	65

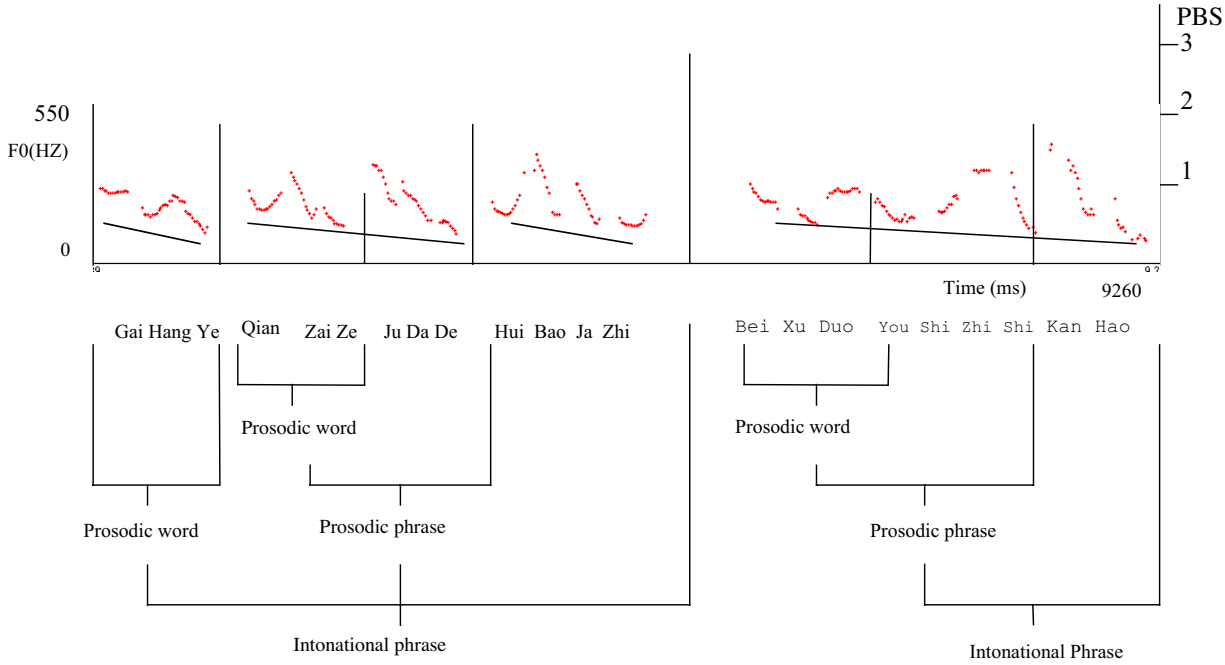

Figure 2. An example of hierarchical prosodic structure realized by bottom line of intonation, where the vertical lines stand for the perceived degrees of boundary.

It is easy to see in table 2 that the minimum value of pitch is lower for the syllables which are after the boundary when there is no break. At the prosodic word boundary, the minimum value of pitch on the syllables after the boundary is slightly higher than that on the syllables preceding the boundary. At the prosodic phrase boundary, the deviation is significant because a pitch reset takes place. At the intonational phrase boundary, the deviation is more significant than that of medium boundaries.

To sum up figure 1 and table 2, we can say that the bottom line of intonation reflects the hierarchical prosodic structure which is shown in figure 2. When there is no boundary, the bottom line declines. When there is a prosodic word boundary, there is a slight pitch reset of the bottom line of intonation. When there is a prosodic phrase and an intonational phrase boundary, there is a significant pitch reset of the bottom line of intonation. The degree of pitch reset is higher for intonational phrase boundaries than for prosodic phrase boundaries.

1.2. Pre-boundary lengthening on prosodic boundary

The pre-boundary lengthening on prosodic boundaries is shown in table 3.

Table 3. Duration and increased percentage of duration (IPD) of syllables preceding prosodic boundaries (ms)

	Duration (ms)	IPD(%)
No boundary	273	-7
Prosodic word boundary	325	11
Prosodic phrase boundary	392	29
Intonational phrase boundary	388	21

The results in figure 3 show that the duration of the syllables preceding boundaries increases as the degree of prosodic boundaries increases. The effect of pre-boundary lengthening is more significant on prosodic word boundaries and prosodic phrase boundaries. There is no significant difference of pre-boundary lengthening between prosodic phrase boundaries and intonational phrase boundaries (F(1)=0.713, P=0.399).

In conclusion, the duration of the syllables preceding boundaries increases as the prosodic boundary degree increases. It is interesting that there should be no significant difference between syllables at the boundary of prosodic phrases or intonational phrases.

1.3. Silence and prosodic boundaries

The correlation between the prosodic degree and the duration of the silence is shown in figure 4.

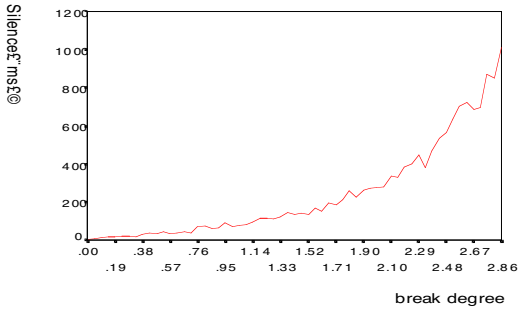

Figure 4. Correlation between the prosodic boundary level and the duration of the silence

It is shown in figure 4 that the higher the degree of the prosodic boundary is, the more rapidly the duration of the silence increases. The longest silence is 1151ms. Yang's (1997) experiment shows that there is a linear correlation between the duration of the silence and the degree of the prosodic boundary. Figure 4 shows that the duration of the silence increases dramatically as the degree of the prosodic boundary increases.

There is no silence on prosodic word boundaries. The duration of the silence on prosodic phrase boundaries ranges from 0 to 617ms with a mean duration of 146 ms. For intonational phrase boundaries, the silences last from 0 to 1151 ms with a mean duration of 408 ms.

2. Conclusion

Researchers pay more and more attention to corpus analysis because large corpora contain many linguistic phenomena, especially more variations in the prosody. In addition, the multi-level labeling of corpora allows statistic analyses useful to study the links between the different layers of speech such as syntax, semantics and prosody.

The main results presented here are:

a) the declination of intonation in Chinese is achived through the bottom line of the intonational contour ;
b) the acoustic correlates of prosodic word boundaries are pre-boundary lengthening and a slight pitch reset of the bottom line of intonation ;
c) the acoustics correlates of prosodic phrase boundaries and intonational phrase boundaries are a significant pitch reset of the bottom line of intonation and the insertion of a silence. Moreover, the higher the prosodic boundary is, the higher the extent of the pitch reset is and the longer the silence is. There is no significant difference for pre-boundary lengthening between syllables on these two boundaries.

In conclusion, pre-boundary lengthening is the acoustic correlate of weak boundary. Pitch reset is that of medium boundary, while silence is that of strong boundary. The acoustic correlates of lower boundaries can also occur on larger boundaries, but the acoustic correlates of larger boundaries usually do not occur on lower boundaries.

The corpus used in this research is read relatively slowly so that the function of silence to indicate prosodic boundary is significant. The variations of tempo should be considered further. This research will be extended from sentence to discourse, from read aloud speech to spontaneous speech.

Acknowledgement

We own great thanks to Professor Lu Shinan for his constructive comments on the earlier version of the manuscript. We also gratefully acknowledge Dr. Zheng Bo for his effort on corpus labeling. Thanks are due to Motorola Inc. China Center for the foundation.

3. References

[1] Pijper J. R.; Sandemrman A.A., 1994. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America*, 96 (4), 2037-2047,

[2] Streeter L.A, 1978. Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America* 64(6), 1582-1592,

[3] Swerts M., 1996. Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, 101(1), 514-521

[4] Wightman C.W., 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 91 (3), 1707-1717,

[5] J. Sheng, 1985. Pitch range and intonation of Chinese tone. *Experimental Phonetics of Chinese* edited by D. Ling & L.J. Wang., Pecking University Press, 75-107 (in Chinese).

[6] B. Wang, S. N. Lu; Y.F. Yang,, 2001. The pitch movement of stressed syllable in Chinese. *Journal of Acoustics* (in press) (in Chinese).