Boundary and Lengthening—On Relative Phonetic Information

Tseng Chiu-yu and Su Zhao-yu

Institute of Linguistics, Academia Sinica
Taipei, Taiwan
{cytling, morison}@sinica.edu.tw

Abstract

The aim of the present study is to better understand the temporal structure of discourse prosody through relative phonetic information, in particular, phrase final lengthening and discourse boundary discrimination using data of fluent Mandarin narrative speech. Two assumptions were tested, namely, by independent/single and by integrated/paired contributions to boundary discrimination. Five acoustic features (1.) boundary pause (BP), (2.) pre-boundary duration (PrDu), (3.) pre-boundary intensity (PrIn), (4.) duration contrast (DuCon) and (5.) syllable intensity contrast (InCon) were used as single factors to whether the identities of discourse boundaries can be discriminated. Subsequently, ten paired combinations from the above five features were generated as relative factors to test their respective discriminations toward boundary identities as well. The results demonstrated that single discrete cues were not as discriminative as paired ones. Among the paired combinations, the pre-boundary syllable duration and the following pause, PreDu+BP, is most discriminative, suggesting that boundary information is related to both cues. We further examined pre-boundary lengthening patterns by three discourse units the syllable, the prosodic word (PW) and the prosodic phrase (PPh) and found systematic global lengthening by pre-boundary phrase PPh is systematically related to discourse identities. We argue that temporal planning is constrained by discourse planning; higher level planning induces overall lengthening; and global lengthening reflects cognitive load. The results also suggest that tempo modulation across speech flow within the same peaking rate is default.

Keywords discourse boundary, boundary discrimination, final lengthening, relative phonetic information

1. Introduction

In previous work on narrative prosody, we have established a hierarchical discourse framework the HPG (Hierarchy of Prosodic Phrase Group) through corpus analysis [1]. The HPG framework specifies how prosodic units are constrained and governed by prosodic layers, and how theses units and layers contribute systematically and cumulatively to global output prosody [1][2][3]. Three major characteristics of the HPG that distinguish it from other prosody studies are (1.) it emphasizes the relative cross-unit prosodic association contained in fluent speech and specifies how such relative phonetic in the supra-segmental domain can be accounted for. (2.) Boundary breaks across fluent speech are treated as discourse units and bear discourse identities. (3.) An intonation phrase is a discourse unit subject to HPG specifications.

The HPG prosodic units from the bottom layer upward are the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath group (BG) and the multiple phrase group (PG). Two prosodic layers are higher than PPh in the hierarchy. The immediate higher node of the BG is the PG, which is the highest node in the HPG hierarchy and refers to breathing limit which corresponds to a compulsory physiological constraint. The highest node PG refers to a complete multiple-phrase speech paragraph and corresponds to the obligatory and ultimate cognitive constraint of speech. The hierarchical relationships among these nodes are SYL<PW<PPh<BG<PG. In correlation to the HPG units are respective discourse boundaries B1, B2, B3 B4 and B5 which bare the same hierarchical relationships and function as prosodic units. Hence, the relationships among the discourse boundaries are B1<B2<B3<B4<B5. The identities of these discourse boundaries are outcome of perceptual annotation by trained transcribers. The intra- and inter-transcriber consistency was over 93% [4]. Furthermore, by specifications of the HPG an intonation phrase is a discourse subunit PPh, and by default not an ultimate prosody unit. The discourse identity of a PPh is subject to three PG specifications the PG-initial, -medial or –final. As a result, output global discourse prosody must contain higher level PG information accordingly. In short, our previous work has shown that in fluent continuous speech, additional prosodic information exists in addition to tones, stress and phrase intonation in the supra-segments, and no prosodic information should be studied as discrete units; and relative associations must be accounted for.

In the following sections, we will present a study on the timing structure of discourse prosody through boundaries and lengthening to show how higher level temporal allocation is organized by discourse units and represents discourse-relative phonetic information .

2. Experiments

2.1. Speech material

Two types of Mandarin speech corpus in different speaking rates were used. Read speech of (1.) plain text of 26 discourse pieces (CNA, approximately 6700 syllables) by one male M051 and one female F051, and (2.) three rhyme formats of Chinese Classics (CL approximately 1600 syllables) by one male M056 and one female F054. The speech data were recorded in sound proof chambers. Pre-analysis annotation included automatically labeled segmental identities in the SAMPA-T notation using the HTK toolkit, and subsequent manual tagging by trained transcribers of perceived boundary breaks using the Sinica COSPRO Toolkit [5]. Annotated

segments were spot-checked by professional transcribers for identities and alignments. Table1 summarizes the speech material by corpus type, speaker, and the number of the HPG prosodic units and boundaries.

Table 1 *Summary of speech data by corpus type, speaker, and the HPG prosodic units and boundaries. The HPG prosodic units are the syllable (SYL), prosodic word (PW), prosodic phrase (PPh), breath group (BG) and phrase group (PG).Corresponding HPG boundaries following each of the prosodic units are B1, B2, B3 B4 and B5, respectively.*

corpus	speaker	SYL/B1	PW/B2	PPh/B3	BG/B4	PG/B5
CNA	F051	6583	3468	1092	297	151
	M051	6661	3332	1207	270	129
CL	F054	1444	599	290	135	58
	M056	1551	619	318	142	47

The mean syllable duration for speakers F051 and M052 is 199ms and 189 ms; the mean syllable duration for speakers F054 and M056 are 265ms and 202ms. Taken as a reference to speaking rate, we found a positive correlation by speech material than by speaker. The same materials were used for all three experiments in the present study.

2.1. Experiment 1

We have stated in Section 1 that discourse prosody is mainly about relative associative information manifested in the supra-segmental domain, and argued that using relative acoustic information would result in better generalized pattern and discrimination of discourse boundaries than discrete acoustic information. Three discourse boundaries, PPh boundary B3, BG boundary B4 and PG boundary B5, were selected as the categories of generalization and discrimination. Three discrete acoustic variables were chosen to test the generalization and discrimination. They are (1.) boundary pause (BP), (2.) pre-boundary syllable duration (PrDu) and (3.) pre-boundary syllable intensity (PrIn). The following two steps were employed to examine patterns of generalization and boundaries discrimination.

Procedures1. Whether a single acoustic factor is sufficient to generalize and discriminate discourse boundary identities

The procedure involved testing whether generalization and discrimination could be achieved by any single acoustic factor. The average values of specified acoustic feature for B3, B4 and B5 were derived from the speech materials by speaker and by speech type. These derived mean values across B3, B4 and B5 were plotted to denote the tendency among boundaries by speech data type and speaker. We then compared the trajectories among different speech data to look for whether the best single acoustic factor with most generalized pattern could be identified. We also tested whether discrimination of discourse boundary identities could be attributed to any one of these single discrete factors.

Procedure2. Whether a relative acoustic factor is sufficient to generalize and discriminate discourse boundary identities

The same rationale from Procedure 1 was utilized to test boundary generalization and discrimination, but using one relative acoustic factor at a time. Between-boundary duration contrast (BwDuCon) and between-boundary intensity contrast (BwInCon) were calculated and used as the contributing factor. Between-boundary duration contrasts were defined by subtracted outcome of cross-boundary syllables. The same subtraction was applied to derive the between-boundary intensity contrasts as well. Both duration and intensity contrasts specify cross-unit as well as cross-boundary relative acoustic information. The same averaging and comparison methods used in Procedure 1 were employed to see if any single relative factor is sufficient to discriminate the identities of discourse boundaries.

2.2. Experiment 2

We hypothesize that pairing of single factors would result in better generalization and discrimination than results from Experiment 1, and the discrimination varies by pair. We further hypothesize the discrimination varies by pair, thus specified pairing would result in better discrimination than single factors of the three discourse boundaries B3, B4 and B4.

The five acoustic features generated from Experiment 1, namely, (1.) boundary pause (BP), (2.) pre-boundary duration (PrDu), (3.) pre-boundary intensity (PrIn), (4.) duration contrast (DuCon) and (5.) syllable intensity contrast (InCon), were used as feature candidates to generate paired-combinations as variables for ANOVA. These five features were first normalized then paired. A total of ten paired combinations were selected. These 10 paired variables were calculated by ANOVA for discriminating categories B3, B4 and B5 from each other.

2.3. Experiment 3

We have previously established that temporal templates for each prosodic unit can be derived by the HPG framework [1][2], suggesting that default temporal patterns exists in each prosodic layer. Thus we hypothesize that final lengthening is unit/boundary specific and must be addressed with boundary pause information. In other words, boundary discrimination must include pre-boundary duration patterns by prosodic units and the following boundary pause to account for discourse effects, and final lengthening is not simply constrained by the intonation phrase. To test the hypothesis, we calculated pre-boundary duration patterns by the HPG prosodic units, namely, the syllable, the PW and the PPh, and compared their respective patterns to the speech data.

3. Results

3.1. Experiment 1

Results from Procedures 1 reveal that among the three single factors pause duration, pre-boundary syllable duration and pre-boundary syllable intensity, although pause duration can discriminate B3 from B4 and B5, B4 and B5 cannot be discriminated. Moreover, no identities of discourse boundary can be discriminated by either the pre-boundary syllable duration or the pre-boundary intensity as shown Figure 1. The results suggest that boundary discrimination cannot be attributed to any single factor Furthermore, pre-boundary lengthening is not a boundary feature by itself. The results have motivated further examination of the role of final lengthening in subsequent investigation.

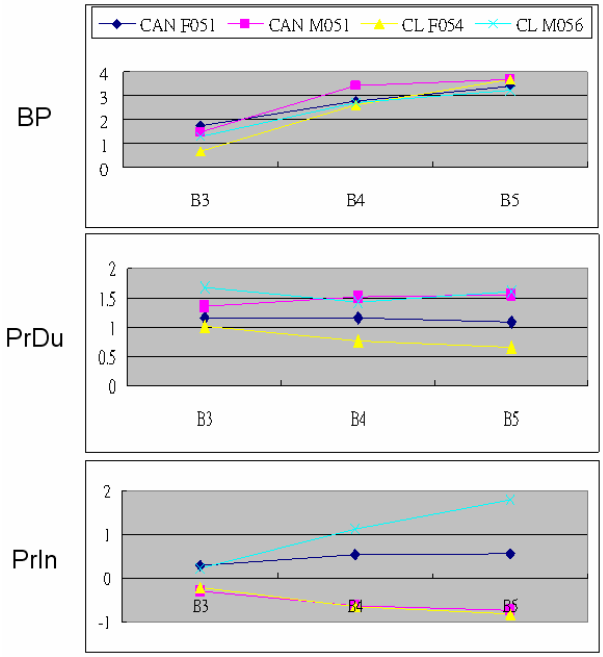

Fig 1: *Cross boundary discrimination by single acoustic features. Each panel denotes one specific acoustic feature. The horizontal axis represents the prosodic boundary indexes B3, B4 and B5. The vertical axis represents the coefficient of normalized values of boundary pause (BP), per-boundary duration (PrDu) and per-boundary intensity (PrIn), respectively. Zero at the vertical axis is defined as the mean of syllable duration.*

Results from Procedures 2 reveal that between the two extended single factors, namely, the contrasts between pre- and post- PPh boundary duration and intensity by one syllable, no significant discrimination of discourse boundaries could be achieved, either, as shown in Figure 2. However, regarding duration patterns, we note that between-PPh duration contrasts provided a generalization of speech data by type, as shown in the upper panel of Figure 2, which pre-boundary duration did not, as shown in the middle panel of Figure 1. The generalization corresponds to overall speaking rate by data type than by speaker, as reported in Sec. 2.1. Regarding intensity patterns, the between-boundary intensity contrasts by only one syllable also provided better generalization than pre-boundary intensity alone. In other words, although using relative information as single factors did not result in better discrimination, both factors resulted in better generalization of the speech data.

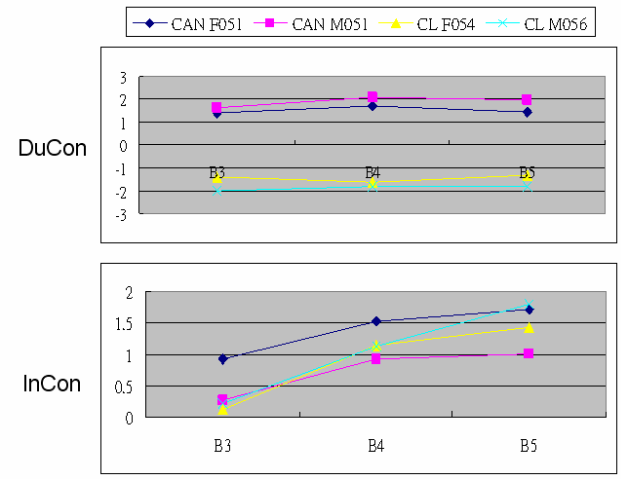

Fig 2: *Cross boundary discrimination by single contrastive factors. Each panel denotes one specific contrastive feature. The horizontal axis represents the prosodic boundary indexes B3 to B5. The vertical axis represents the coefficient of normalized values between boundary duration contrasts (DuCon) and between boundary intensity contrasts (InCon). Zero at the vertical axis is defined as the mean of syllable duration and intensity.*

3.2. Experiment 2

The results of Experiment 2 are summarized in Table 2. The within and between are two evaluation indicators for discrimination. Within by definition is the population variance of distribution of sample means and between is the distance between the sample means. The F-ratio (Between/Within) indicates distinctions among B3、B4 and B5.

The obtained results indicate that among the ten paired combinations, significance of boundary distinction was only found among two pairs, namely, PrDu+BP and PrIn+BP, where F (2, 40) =0.28387, P<0.05. That is, the PrDu+BP pair contributes most to boundary discrimination, followed by the PrIn+BP pair. It is obvious that when boundary pause is combined either with pre-boundary duration or intensity, discourse boundaries can be discriminated. In addition, the within is minimal for pairs PrIn+BP and PrDu+BP, indicating that when pre-boundary duration is combined with between-boundary intensity contrast, boundary discrimination is best, flowed by the combination of boundary pause and pre-boundary duration. The above results further indicate that the PrDu+BP combination, in short, PPh-final duration and pause make up the most discriminative relative cue for discourse boundary identities.

Table 2: *List of Within and Between and F-ratio for pairs of two acoustic features.*

Pairs of Acoustic features	PrIn+BP	PrDu+BP	BP+InCon
Between	2.394360811	2.117735421	0.930326811
Within	0.714117065	0.479096215	1.116294559
F-ratio	3.352896784	4.420271653	0.833406204

Pairs of Acoustic features	BP+DuCon	PrIn+InCon	PrDu+PrIn
Between	0.070391796	1.297194809	0.120075103
Within	1.810655131	0.912193354	0.875052214
F-ratio	0.038876424	1.422061237	0.137220501

Pairs of Acoustic features	PrIn+DuCon	PrDu+InCon	PrDu+DuCon
Between	0.353532872	1.020569418	0.076907482
Within	1.652913517	0.374550542	1.763574503
F-ratio	0.213884676	2.72478425	0.043608865

Pairs of Acoustic features	DuCon+InCon
Between	1.254027187
Within	1.736954223
F-ratio	0.721969048

Table 3 summarizes the averaged sum of PPh-final syllable duration and boundary pause duration in seconds, where constant pattern across boundaries can be observed.

Table 3: *A list of average sum of final syllable duration and pause (sec by speech data type and speaker)*

corpus	speaker	B3	B4	B5
CNA	F051	0.499738	0.607713	0.684998
	M051	0.519527	0.800465	0.880004
CL	F054	0.52102	0.833563	1.007355
	M056	0.456447	0.679508	0.774484

3.3. Experiment 3

Figure 3 shows the phrase final duration patterns by HPG prosodic units the syllable, the PW and the PPh across speech data and speaker. We note by analyzing the pre-boundary duration pattern of the final syllable alone reveals a pattern that consistent lengthening occurs before the B3 boundary, but not before higher boundaries B4 and B5. The result does not explain why discourse boundary identities could be consistently perceived across listeners. However, if the same inconsistency was found across all boundaries, which was case with the patterns found for the PW, then lengthening may not be a reliable boundary cue, and suggest that lengthening is related to the lower level phrase boundary only. Nevertheless, a cross-speaker and cross-data-type was found in the case when the duration patterns were extended to include the entire pre-boundary PPh, shown in the lower panel of Figure 3. The lengthening patterns of pre-boundary PPh were also consistent with respect to boundary identities. Furthermore, the consistent pattern also implies that pre-boundary lengthening at the higher level applies to higher and larger prosodic units, suggesting more cognitive load may result in overall slower speaking rates for such prosodic units.

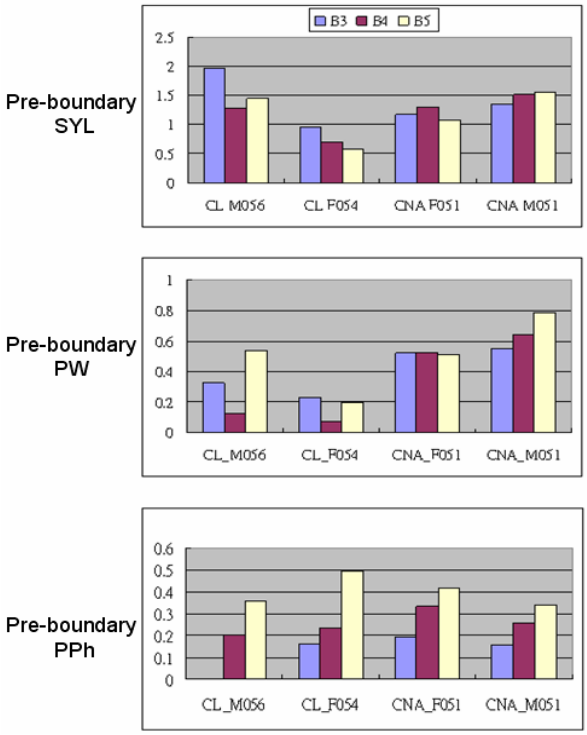

Fig 3: *Cross boundary comparison of duration patterns by prosodic units the syllable (SYL), the PW and the PPh. The horizontal axis represents indexes of the speech data and speaker. The vertical axis denotes normalized average duration of prosodic units.*

Figure 4 shows the results of average duration pattern by discourse boundary identities B3, B4 and B5. We found that discourse boundaries can be discriminated by the duration patterns of pre-boundary PPh across speaker and speech data type, as shown in the lower panel of Figure 4, but not by the patterns of pre-boundary SYL and PW, as shown in the top and middle panels. In other words, the identities of discourse boundaries are consistent with the respective lengthening patterns of the pre-boundary PPh.

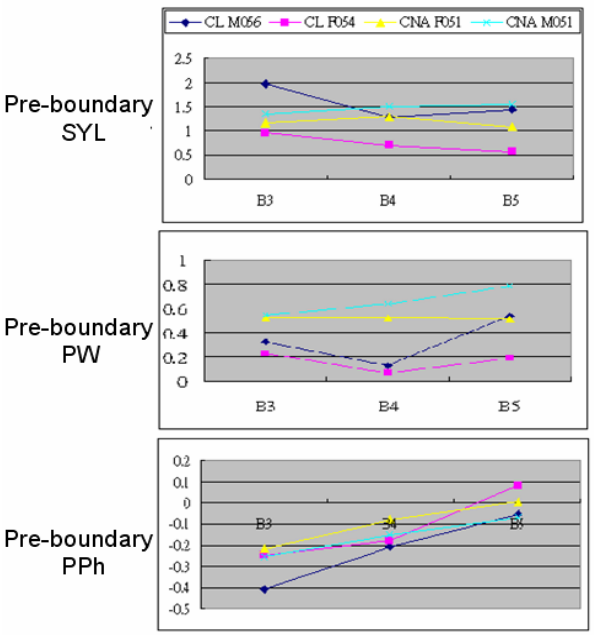

Fig 4: *Cross-boundary duration patter by boundary breaks. The panel denotes result of specific prosodic unit. Each curve denotes one of speech data. The horizontal-axis represents prosodic boundary index. The vertical-axis denotes the normalized average duration for specific prosodic unit.*

4. Discussion

One of the most difficult tasks of phonetic analysis is how to account for relative phonetic information. Both the pitch and temporal features in phonological structure are presented in abstract and timeless organization. Although timing structure has been studied extensively, little reference with respect to global patterns of continuous speech is available.

For example, one of the most well known previous acoustic studies on Mandarin duration patterns is how stress is related more to temporal modulation instead of F0 contours [6], referring to segmental duration at the lexical level. Many later studies on duration were studied in relation to prosody or as prosody units, but in units such as prosodic words or intonation phrase, and focused on segments and syllables. In other words, much less attention has been paid to relative phonetic information at the discourse level. Studies on boundary and lengthening were no exception. For example, a comprehensive investigation on Mandarin segment lengthening made important observations of how prosodic boundaries and paused occur between prosodic units instead of within units, and manifestation of pre-boundary lengthening bears prosodic functions to the phrase [7]. More recent studies reported on the role of lengthening with reference to prosodic boundary and its perceptual significance in continuous speech using the pre-boundary syllable [8][9], which inadvertently suggested the syllable as the default unit of lengthening. Another study reported that lengthening is complimented by pause duration at prosodic boundaries, but he units did not go beyond the intonation phrase[10]. Extensive perceptual studies revealed that although the sentence-final pause duration is significantly longer than prosodic-phrase final counterparts, the duration of sentence- and prosodic-phrase-final syllable was not significantly different [11]. Nevertheless, more recent studies reported how the degree of final lengthening is modulated by boundary types [12], and how segmental strengthening is relative to prosodic functions [13], but without systematic reference to discourse units and structure. In short, almost all of the previous studies have focused on modulation of segmental duration at the syllabic level. We noted also that in the case when the factor of the discourse was noted, comparatively little discourse account has been reported. In particular, the relative aspect of timing structure in discourse prosody, especially with respect to boundary features and boundary identities have not received much attention. We think that one reason of the oversight could be due to taking an intonation phrase as the ultimate prosodic unit without reference to the multiple-phrase speech paragraph

However, one of our recent study on PPh boundary B3 we studied the much varied B3 pause duration not by the pause duration themselves, but with respect to cross-unit contrastive patterns in the acoustic signals, because we hoped to find out why these within-PG phrase boundary identities were consistently perceived across listeners, including cases when there was no pause at identified boundary. We discovered that within PG phrase boundaries can be accounted for by boundary immediate contrastive patterns of duration and intensity without any pause information [14]. The findings enabled us to argue that the on-line perception of discourse boundaries in fluent speech makes cross reference between and across cues, and that relative information is crucial. The question then is in what domain and unit relative information exists. Therefore, in the present study, we further examined boundaries and lengthening with reference to discourse functions and overall temporal structure.

The results from Experiments 1 show that single factors are not discriminative of discourse boundaries. The results of Experiment 2 show that the identities of discourse boundaries can be discriminated when pre-boundary syllabic duration or intensity is combined with the following boundary pause. In other words, pre-boundary syllabic information by itself is not sufficient for the discrimination of boundary identities, but when coupled with the pause, the combined feature proved to be adequate. The results suggest that in limited context a little extra relative information goes a long way. We believe that more high-level relative information can be utilized to facilitate on-line processing.

The results from Experiment 3 are most interesting because it provided evidence of how global lengthening could be represented and what its discourse function is. The results make direct reference to overall timing modulation, lengthening could happen to the entire phrase that is located before a discourse boundary. Our result showed the reason why lengthening was not restricted to the final syllable only is because global temporal planning is also involved. Consistent perceptual identification of discourse boundary identities echoes the finding, because listeners must make use of global relative information to facilitate on-line processing. The same results also imply that overall modulation of temporal allocation is regulated by discourse prosodic organization, and interact with fixed or changed speaking rate. We believe that the implications of global lengthening have shed new lights to how speakers plan and process the temporal features across fluent speech. Default discourse temporal templates could be derived and modeled.

5. Conclusion

We have shown that (1.) overall temporal modulation within a fixed speaking rate involves the timing structure and temporal arrangement at the discourse level and result in overall lengthening of the pre-boundary phrase, (2.) how lengthening is in fact an integral part of boundary information by discourse units, and when coupled with boundary pause facilitates boundary identities to emerge, (3.) Lengthening is relative should be addressed with sufficient relative information, and (4.) global lengthening related to overall modulation of speaking rate shows that the timing structure of discourse prosody is subject to discourse organization and discourse association. In summary, we hope to show that relative phonetic information that exists in the speech events but usually outside the concern of phonology contributes significantly to speech production and speech processing. Such relative information would not emerge unless we adopt a discourse perspective of investigation and make use of methodological innovations.

6. Reference

[1] Tseng, Chiu-yu, Pin, Shao-huang and Lee, Yeh-lin 2004. Speech prosody: Issues, approaches and implications. in *From Traditional Phonology to Modern Speech Processing* (語音學與言語處理前沿), edited by Fant, G., Fujisaki, H., Cao, J. and Xu, Y., Foreign Language Teaching and Research Press (外語教學與研究出版社), 417-437, Beijing, China.

[2] Tseng, Chiu-yu, Pin, ShaoHuang and Lee, Yeh-lin, Wang, Hsin-min and Chen,Yong-cheng 2005. Fluent Speech Prosody: Framework and Modeling, *Speech Communication (Special Issue on Quantitative Prosody Modeling for Natural Speech Description and Generation)*, Vol. 46:3-4, 284-309.

[3] Tseng, Chiu-yu 2006. "Prosody Analysis" in *Advances in Chinese Spoken Language Processing*, edited by Chin-Hui Lee, Haizhou Li, Lin-shan Lee, Ren-Hua Wang, Qiang Huo, World Scientific Publishing, 57-76, Singapore.

[4] 鄭秋豫 2001. 語流中韻律結構的主要徵信。 *第6屆全國語音通訊學術會議(NCMMSC-6)*， (Nov. 19-24, 2001)，中國：深圳，169-172。

[5] Tseng, Chiu-yu, Cheng, Yun-ching and Chang, Chun-Hsiang 2005. Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech, *Oriental COCOSDA 2005*, (Dec. 6-8, 2005), Jakarata, Indonesia.

[6] 林燾 1983. 探討北京話輕音性質的初步實驗，*語言學論叢*, 第 10 輯，北京：商務印書館。

[7] 曹劍芬 1998. 漢語普通話語音節奏的初步研究。*中國社會科學院語言研究所語音研究報告 1998*。北京：中國社會科學院語言研究所。

[8] 祖漪清、陳肖霞 1999. 連續語流中的音節延長及其作用。*第四屆全國現代語音學學術會議*。中國：北京，58-63。

[9] Zu, Y. and Chen, X., 1999. Segmental duration and lengthened syllables." *ICPh99*`, San Francisco, 277-280.

[10] 錢瑤、初敏、潘悟雲 2001. 普通話韻律單元邊界的聲學分析。*第五屆全國現代語音學會議*。中國：北京，70-74。

[11] 王蓓、楊玉芳、呂士楠 2001. 漢語韻律層級邊界結構的聲學相關物。*第五屆全國現代語音學會議*。中國：北京，161-165。

[12] Fon, J. and Johnson, K., 2004. Syllable onset intervals as an indicator of discourse and syntactic boundaries in Taiwan Mandarin. *Language and Speech*, 47(1), 57-82.

[13] 曹劍芬 2005. 音段延長的不同類型及其韻律價值。*中國社會科學院語言研究所語音研究報告 2005*。北京：中國社會科學院語言研究所，5-12。

[14] Tseng, Chiu-yu and Chang, Chun-Hsiang , 2007. Pause or No Pause?—Phrase Boundaries Revisited. *The 9th National Conference on Man-Machine Speech Communication（NCMMSC 2007, 第九屆全國人機語音通訊會議）*, (October 21-24, 2007), 黃山, 中國.