Discourse Prosody Context—Global F0 and Tempo Modulations

Chiu-yu Tseng and Zhao-yu Su

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

Abstract

The present study is a corpus analysis of discourse prosodic information using two different types of fluent continuous Mandarin speech. Global F0 heights and duration patterns of within- and between-paragraph phrases were compared by discourse positions. Results showed that overall phrase-level F0 height was paragraph-initial>-medial>-final while the tempo pattern was paragraph-initial<-medial<-final. All of the differences were statistically significant across speakers and speech materials. The results suggest that discourse prosody context provides information of discourse planning, within-paragraph phrase association and between-paragraph topic change. We argue that global discourse prosody context is a crucial factor of speech communication, and can be applied to discourse segmentation, TTS synthesis naturalness and language pedagogy.

Keywords global discourse prosody, prosody context, F0 height, tempo pattern, duration modulations, discourse association, paragraph association, topic change.

1. Introduction

In previous work on narrative prosody, we have established a hierarchical discourse prosody framework called the HPG (Hierarchy of Prosodic Phrase Group) which accounts for discourse prosody organization and association [1]. The layered HPG prosodic units from bottom upward the hierarchy are the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath group (BG) and the multiple phrase group (PG). Corresponding to the HPG units are discourse boundaries B1, B2, B3 B4 and B5. Their relationships can also be expresses as SYL<PW<PPh<BG<PG and B1<B2<B3<B4<B5. Three characteristics distinguish the HPG from other discourse prosody studies: (1.) the HPG treats a multiple-phrase speech paragraph as a discourse unit and terms it the PG (Prosodic Phrase Group). (2.) The HPG boundary breaks are discourse specified prosody unit [1][2][3]. (3.) A PPh (Prosodic Phrase) is an intonation phrase and a subunit of the PG. A PG superimposes three alternative but relative discourse positions onto its subordinate PPhs, the PG-initial, -medial and –final, to designate the beginning, continuation and termination of a PG. The superposition requires the within-PG PPhs to adjust accordingly in order to achieve paragraph prosody. All of the above discourse prosodic units and boundaries are perceptually consistent at 93% across listeners. [4]

The aim of the present study is to show that in fluent continuous speech, discourse prosody is expressed through both cross-over as well as adjacent units adjustment and modulations to form the necessary prosody context. To illustrate the points, we will analyze the global patterns and adjustments of the PPh by two acoustic correlates, namely, overall phrase-level F0 height modulations and tempo alterations with respect to discourse organization. The global F0 contours are analyzed using the command-response model [5], and the tempo patterns using a more refined measurement of mean syllable durations. The results obtained will be discussed in relation to paragraph build-up, within-PG phrase association, topic change, topic association and their significance to discourse prosody and prosody context.

2. Speech material

Two types text were used to prepare Mandarin speech corpora. (1.) Plain text of 26 random discourse pieces (CNA, approximately 6700 syllables), and (2.) three rhyme formats of Chinese Classics (CL approximately 1600 syllables). Read speech of one male and one female for each text format was used, namely, M051 and F051 for CAN and M056 and F054 for CL. The data were microphone speech recorded at sound proof chambers. Automatic annotation of segmental labeling was performed by the HTK toolkit with the SAMPA-T notations. The HTK-annotated segments were spot-checked by professional transcribers for segmental alignments. Manual tagging was performed by trained transcribers using the Sinica COSPRO Toolkit of perceived prosodic units and boundary breaks. [5] The mean syllable duration of data type CNA is 199ms and 189 ms for F051 and M052; and the mean duration of CL is 265ms and 202ms for F054 and M056. A cross-speaker positive correlation of speaking rate by text formats was found. Rhymed formats were read in slower speaking rate than the unrhymed counterpart.

3. Analysis

3.1. The PPh as PG Specified Discourse Unit

By definition of the HPG, a speech paragraph PG is made of three relative positions the PG-initial, -medial and -final to designate the beginning, continuation and termination. These alternative specifications are superimposed onto the subordinate units the PPh whereby the PG-initial and –final PPh are singular phrases while the -medial PPh could be either singular or plural in number. In other words, the PPh is a discourse relevant prosodic unit instead of a discrete independent one, and is examined in relation to its discourse positions, identities and functions.

3.2. The PPh and PG position index

Annotated PPhs from the speech data were classified into three correlating classes by sequence index described below:

$$\text{PG-position} = \begin{cases} \text{PG-initial} & \text{when sequence index=1} \\ \text{PG-final} & \text{when sequence index=M} \\ \text{PG-medial} & \text{otherwise} \end{cases}$$

M= Number of PPh in PG

3.3. Extraction of prosodic features F0 height and tempo

Two supra-segmental acoustic features of the PPh were extracted, namely, the F0 height and duration patterns. These extracted features were compared by the PG sequence index to see if they exhibit different but consistent global patterns by speech type and by speaker. Statistical analyses were then performed to see if the differences among them were significant.

3.4. Extracting global PPh F0 using the command-response model

The command-response model, commonly known as the Fujisaki model, was used to extract the overall tendency of F0. [6] The model consists of base frequency, a Phrase command Ap indicating the magnitude of global phrase intonation and Accent command Aa indicating local humps. Aa is superimposed onto Ap to derive the ultimate output F0 contour. The model is defined as below.

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} Ap_i G_p(t - T_{0i}) + \sum_{j=1}^{J} Aa_j [G_a(t - T_{1j}) - G_a(t - T_{2j})]$$

i, j =Index of phrase command, Index of accent command
F_b =Base frequency
Ap_i =Phrase command magnitude
Aa_j =Accent command magnitude
$G_p(t - T_{0i})$ = Phrase command response function
$[G_a(t - T_{1j}) - G_a(t - T_{2j})]$ = Accent command response function

When the model is used to analyze Mandarin or other syllable-based tone languages, it has been commonplace that the Ap command is used to represent phrase intonation contour pattern, and the As command is used to represent the lexical tones. However, since the focus of the present study is on overall F0 patterns without reference to tones, only values of the Ap command were extracted. Auto extraction of the Ap command was based on Mixdorff [7] and modified to fit our data.

3.5. Deriving PPh tempo by syllable duration patterns

Mandarin Chinese is a syllable-time language, and averaged syllable duration is commonly used to represent speaking rate without. Nevertheless, extreme variations of syllable duration do occur and we have devised a more sophisticated definition to derive mean syllable duration. The maximum and minimum of syllable duration in the PPhs were first removed before the derivation. The remaining syllables were averaged to derive mean syllable duration defined and described below. *M* denotes the number of syllables in the PPh and *i* the index of syllable.

$$PPh\_Dur = \sum_{i}^{M} syl\_dur_i$$

$$PPh\_AveDur = PPh\_Dur / M$$

4. Results

4.1. Relative F0 heights Within- and Between-PG

Figure1 shows the plotting of an example of extracted Ap by speaker. Two adjacent PG's were plotted. We note a reset at the PG-initial PPh is followed by gradual lowering of F0 height until the end of the PG-final PPh. Similar lowering patterns are found for both PG's, and can be characterized as a down-stepping effect. The overall within-PG F0 down-stepping by PPh PG-Initial>-medial>-final characterizes within-paragraph association. The within-PG association can thus be specified by F0 modulations from high to mid to low by PG positions. We note also that PG adjacency can be specified by F0 modulations from low to high. The low-high constitutes a contrast that indicates topic change and between-PG association. In other words, within-PG association is expressed by F0 down-stepping and between-PG association by adjacent F0 contrast. Both features collectively account for paragraph make-up, topic change and higher-level discourse association.

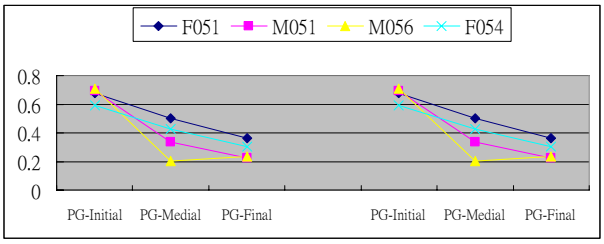

Figure1. An example of average Ap by PG-position of two adjacent PG's by speaker and by speech data type. The horizontal axis represents the PG-position index. The vertical axis represents the average Ap values.

4.2. Relative Tempo Modulations Within- and Between-PG

Figure2 shows the plotting of syllable duration patterns of the same example shown in Figure1. The within-PG tempo by PG positions shows a steady fast-to-slow pattern which can be characterized as PG-initial<-medial<-final. The PG-initial PPh is the fastest, the PG-medial the slower; and the PG-final the slowest. On the other hand, the between-PG duration patterns exhibits an adjacent long-short contrast that represent sudden change of tempo. These duration patterns show patterned tempo alterations with respect to within-paragraph association and higher-level discourse association, and are also consistent with F0 height modulations.

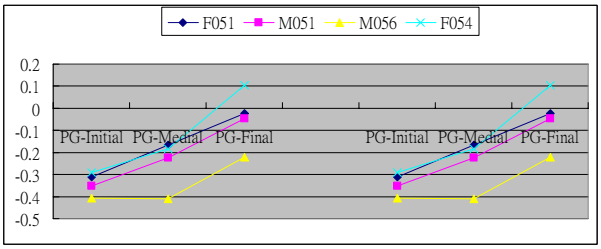

Figure2. The tempo of the same examples used in Figure1is plotted by speaker and by speech data type. The horizontal

axis represents the PG-position index. The vertical axis represents the mean syllable duration values.

4.3. Statistical Analysis

Statistical analysis of both the extracted Ap and duration differences using t-test is performed to test if the discrimination among the PG positions is significant. The test pairs are as follows.

$$\text{Group1/ Group2} = \begin{cases} \text{PG-Initial / PG-Medial} \\ \text{PG-Initial / PG-Final} \\ \text{PG-Medial / PG-Final} \end{cases}$$

The criterion for the t-test is as

$$\mu(\text{Group1}) - \mu(\text{Group2}) = 0 \quad (\text{Alpha}=0.01)$$

Alpha is set to 0.01. The results of the t-test for Ap are listed in Table1, significant differences are indicated by the asterisk ⋆

Table 1 Comparison of Ap by pairs of PG positions initial/medial, initial/final and medial/final and by speaker, the asterisk ⋆ denotes statistically significant differences.

Speaker	Initial-Medial		Initial-Final		Medial-Final	
F051	0.1764	⋆	0.313	⋆	0.1366	⋆
M051	0.35757	⋆	0.4674	⋆	0.1098	⋆
F054	0.162	⋆	0.286	⋆	0.124	⋆
M056	0.509	⋆	0.4739	⋆	-0.0352	

Comparison of extracted Ap values reveals that the three PG-positions are significantly different from each other, both between and across PG positions, and across the speakers and speech type. Only one speaker M056 showed no significant F0 lowering in one condition, namely, from PG-medial to –final. But the overall F0 patterns are consistent across speaker and speech type.

The results of the t-test for mean syllable duration are listed in Table 2; significant differences are indicated by the asterisk ⋆.

Table 2 Comparison of mean syllable duration by pairs of PG positions initial/medial, initial/final and medial/final and by speaker, the asterisk ⋆ denotes statistically significant differences.

Speaker	Initial-Medial		Initial-Final		Medial-Final	
F051	-0.1442	⋆	-0.286	⋆	-0.142	⋆
M051	-0.128	⋆	-0.307	⋆	-0.179	⋆
F054	-0.107		-0.394	⋆	-0.287	⋆
M056	0.0021		-0.1866	⋆	-0.1888	⋆

Comparison of mean syllable duration reveals that the tempo pattern of PG-Initial<-medial<-final are significantly different from each other, both between and across PG positions, and across the speakers and speech type. Two speakers F054 and M056 showed a slightly different pattern of no significant tempo difference between the PG-initial and –medial when $p \le 0.0001$. However, the overall tempo patterns are consistent across speaker and speech type.

5. Discussion

One of the most difficult tasks of prosody analysis is how to account for prosodic context through relative acoustic information. Although both the intonation patterns and timing structure have been studied extensively, little reference with respect to global patterns of continuous speech is available. Relative acoustic information such as F0, duration and intensity is often studied in discrete units and without reference to relevant context, and context is often taken as collocation or adjacency only without reference to higher level constraints. The most notable cases are intonation contours and phrase-final lengthening. For example, although the intonation phrase has been studied extensively, intonation variation in fluent continuous speech is often studied independent from discourse context. In the case of tone language such as Mandarin Chinese, much attention has been focused on tonal concatenation and tonal context while the default phrase intonation in isolation is adopted. As a result, relatively little attention has been paid to additional prosody information that also exists in the F0. Furthermore, since Mandarin is a syllable-time language where by default the syllable is assumed to be of no duration difference, relatively less attention has been paid to global tempo structures or temporal allocation across continuous speech. We believe that the oversight is largely due to taking the intonation phrase as the ultimate prosodic unit without reference to higher-level discourse information and discourse functions. Pre-boundary lengthening is no exception.

However, two of our recent studies on discourse boundaries have revealed interesting results that contradict the above assumptions. In one study of boundary pauses in read continuous speech, we discovered that within-PG phrase boundaries, or B3 by the HPG definition, can be specified by prosody context information such as boundary immediate long-short duration contrasts and decrease-increase intensity contrasts, both of which without reference to pause information. [8] Similar but distinct patterns were found for other prosodic boundaries, thus proving boundaries in fluent speech bear discourse identities. The findings further demonstrate that within-PG phrase boundaries can be realized by the prosodic context instead of the pause duration, and account for why the B3 pause durations are less significant boundary cues. Our preliminary perception experiment also showed that when boundary pauses were removed from the speech flow, identification of discourse boundaries was not impaired. Moreover, in another recent study on pre-boundary lengthening, we found that no single discrete acoustic factor such as pause duration, phrase-final vowel lengthening or phrase-final intensity decrease is as discriminative as any two adjacent factors combined. By pairing the pre-boundary syllable duration and boundary pause, the boundaries identities of B3, B4 and B5 can be discriminated. [9] Both studies suggested that discourse context provide significant prosody information. The question then is how discourse prosody context in the speech signal can be represented.

The results of the present study illustrate that it is possible to bring forth global prosodic context with respect to discourse planning threshold and discourse organization. From the analysis of relative PPh F0 height by PG positions, the overall within-PG F0 down-stepping by phrase showed how within-paragraph association is delivered by relative adjacent and cross-over association, as illustrated in Figure 1. Studying the

PPh as independent unit would most likely result in unaccountable variations. In addition, the global PG down-stepping of PG-Initial>-medial>-final is superimposed onto the intonation of subordinate PPhs and triggers relative F0 modulation across the board to form the necessary paragraph prosody context. Therefore, the results also illustrate how the speech paragraph context requires at least three relative positions rather than by immediate adjacency. In short, higher-level discourse prosody context above the intonation unit can only be taken into account by phrase association, and variations of intonation are by no means random. Moreover, the results of PG adjacency also shed lights on how between-paragraph discourse association can be specified by neighborhood contrastive patterns. That is, F0 reset is relevant with reference to relative mid and low. When the PG-final low is followed by a PG-initial high, the adjacent F0 contrast makes the change of topic more prominent to the ear. Such contrastive F0 adjacency forms the between-paragraph discourse prosody context that is distinct from within-PG associative patterns. In short, distinct discourse prosody context includes patterns such as overall F0 down-stepping (PG-initial>-medial>-final), within-PG cross-over association (PG-initial vs. PG–final), and between-PG contrast (PG-final vs. PG-initial). We believe such information reflects discourse planning and provide relative cues to on-line prosody look-ahead and global prosody processing.

The results of tempo modifications from within-PG analysis reveal relative fast-to-slow adjustments by PG positions, also, which can be seen as within-paragraph global temporal context of phrase association as well. The results of tempo modifications are consistent with F0 modulations and provide further evidence of discourse prosody context in the temporal domain.

The individual differences from these two analyses are also interesting. In the case of F0 modulations, speaker M056 completed the F0 down-stepping by PG-medial and held low thereafter. In the case of tempo modifications, speakers F051 and M051 kept phrase tempo constant from PG-initial to –medial, and achieved the tempo change from –medial to –final. Both the F0 and tempo differences are evidence how speaker variation can be realized by either acoustic correlate without changing the overall prosody context.

6. Conclusion

We have shown from the present study that discourse context information is systematic and global modulations are evident. The results also demonstrate how discourse prosody context can be extracted from the signal by treating relative prosodic information from perspectives of association and contrast because prosodic contextual information delivers within-paragraph phrase association as well as between-paragraph topic change. Higher level discourse association is also expressed from global prosodic context. Together with our recent findings from discourse boundary identities and discourse conditioned pre-boundary lengthening [8] [9]; we believe that the communicative functions of discourse prosody context merits more attention. In particular, in fluent speech the discourse prosody unit PPh, or an intonation unit, should be investigated in relation to prosody context by larger scale, and with respect to discourse organization.

In addition, discourse contextual information can be used in speech recognition for discourse segmentation, paragraph span and association as well as identification of topic change. Overall modulations of relative F0 height and tempo adjustments could also help improve the naturalness TTS output prosody. It can also be applied to language pedagogy to overall fluency drills as well as listening comprehension strategies. Future directions include implementing contextual prosodic information into an existing prosody model [1], and applying the model to speech synthesis and speech recognition.

7. 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*, edited by Fant, G., Fujisaki, H., Cao, J. and Xu, Y., Foreign Language Teaching and Research Press, 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] Tseng, Chiu-yu 2001. "On Major Features of Fluent Speech Prosody" (in Chinese) *The 6th National Conference on Man-Machine Speech Communication（NCMMSC 2001*,(Nov. 19-24, 2001，Shenzhen, China, 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] Fujisaki H, Hirose K. 1984. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *J.Acoust, J.Acoust. Soc. Jpn.(E)*, 1984; 5(4): 233-242,1984.

[7] Mixdorff, H., Hu, Y. and Chen, G. 2003. "Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin", *Proceedings of Eurospeech 2003*; 873-876.

[8] Tseng, Chiu-yu and Chang, Chun-Hsiang, 2007. Pause or No Pause?—Phrase Boundaries Revisited. *Tsinghua Science and Technology* (in press)

[9] Tseng, Chiu-yu and Su, Zhao-yu, 2008. "Boundary and Lengthening—On Relative Phonetic Information" *The 8th Phonetics Conference of China and the International Symposium on Phonetic Frontiers*, April 18-20, 2008, Bejing, China.