Mandarin Spontaneous Narrative Planning—Prosodic Evidence from National Taiwan University Lecture Corpus

Chiu-yu Tseng, Zhao-yu Su*and Lin-shan Lee***

*Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei, Taiwan
**Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
cytling@sinica.edu.tw

Abstract

This paper discusses discourse planning of pre-organized spontaneous narratives (SpnNS) in comparison with read speech (RS). F0 and tempo modulations are compared by speech paragraph size and discourse boundaries. The speaking rate of SpnNS from university classroom lecture is 2 to 3 times to that of RS by professionals; paragraph phrasing of SpnNS is 6 times that of RS. Patterns of paragraph association are distinct for SpnNS and RS. Sub-paragraph and paragraph units in RS are marked by distinct relative F0 resets and boundary pause duration, but by patterns of intensity contrasts in SpnNS instead. Consistent to both data sets is the finding that combined relative supra-segmental cues reflecting global prosodic properties are more discriminative to distinguish discourse boundaries than any fragments of singular cue, supporting higher-level discourse planning in the acoustic signals. We believe these findings can be directly applied to speech technology development.

Index Terms: discourse planning, spontaneous narrative, speech paragraph size, and discourse boundary discrimination

1. Introduction

Though there is considerable literature in the past decade on spontaneous dialogue speech (SpnDS), the focus has been features such as disfluency, repetition, repair, hesitation, fillers and filled pauses [1 for example]. Later reports of spontaneous narratives (SpnNS) of classroom lectures focus more on data processing, using alignments with presented Powerpoint slides as reference of segmentation [2][3][4]. So far there has been little report on the organization SpnNS which represent well-organized spontaneous discourse, especially with reference to paragraph phrasing and discourse planning. The aim of this paper is to investigate discourse planning characteristic to pre-organized SpnNS. Comparisons of F0 and tempo modulations will be made to read speech (RS) by speech paragraph size and discourse boundaries.

We have studied Mandarin fluent continuous RS extensively on perceived discourse segments with manual annotation, and found that the formation of output prosody is accountable with hierarchical associations specifying chunking and phrasing by discourse units. Paragraph prosody could be reconstructed by separating contributions from ranks of same-level smoothing of sister-constituent adjacency and higher-level global layering at the phrase lvel, and accounted for by their respective and cumulative contributions in the supra-segmental acoustic correlates [5][6][7], as specified by the HPG (Hierarchical Prosody phrase Grouping) framework. Quantitative evidence has shown how distributions of layered contribution by the HPG hierarchy is systematic [6][8], while levels of discourse boundaries and boundary related properties have to be taken into account as well to collectively make up discourse prosody [9][10][11][12].

The HPG discourse prosodic units are defined as those bearing variable templates in perceived pitch, tempo and loudness that receive specifications from relative discourse positioning, boundary assignments, change of breath and topic change. These prosodic units from the bottom rank upward are the syllable (SYL)/ boundary break B1, the prosodic word (PW)/B2, the prosodic phrase (PPh)/B4, the change of breath of breath group (BG)/B4 and the multiple-phrase prosodic group (PG)/B5. 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 and their respective boundaries B1, B2, B3 B4 and B5. Paragraphing is indicated by three PG relative positions, the PG-initial (PG-I), -medial (M) or –final (F). These positions trigger and cause respective PPh constrained to modify their acoustic patterns accordingly.

In the following sections, we will present F0 and tempo analysis in relation to discourse boundaries analyses with both data of SpnNS and RS to derive some of the most distinctive features of narrative planning.

2. Speech materials

Pieces from two kinds of speech data are used for comparative analysis: (1) the National Taiwan University (NTU) DSP Lecture Corpus (hence NDLC) for SpnNS (45 classroom microphone recording of a NTU (National Taiwan University) DSP course, 1 speaker LSL, 3.92GB in 45 waves) and (2) the Sinica COSPRO for RS (Sinica Mandarin Continuous Speech Prosody Corpora, 8 types of read speech, multiple speakers, 7.9GB). One hour of the NDLC (1 speaker 14,305 syllables total) and the complete set of CNA (26 random discourse pieces, 154 minutes by 1 male speaker M51 and 120 minutes by 1 female speaker F05, around 12,000 syllables per speaker) [8] were selected. All of speech data were processed manually by trained transcribers for perceived phrase and discourse boundaries using the HPG protocol and annotation. As a result, all segmented units are discourse prosody units. The mean number of syllable per BG and PG are listed in Table1 by data type and speaker.

Table1. *The mean number of syllable per BG and PG are listed in Table1 by data type and speaker.*

Corpus	NDLC	CNA	
Speaker	LSL	F051	M051
Mean (Syl/BG)	103.87	38.97	42.89
Mean (Syl/PG)	653.09	76.76	90.20
Number of BG per PG	6.29	1.97	2.10

The speaking rate of university classroom lecture NDLC is 2 to 3 times to that of RS. The size of paragraph phrasing of SpnNS is relatively much larger than RS: 654 syllables per PG to 77-90 syllables per PG, respectively. Average change of breath per PG is 6 times in SpnNS but only 2 times in CNA. Average number of syllables in each BG (breath group or breathing cycle) is around 40 for RS but 104 for SpnNS. The significant difference in paragraph size and speaking rate indicate that speakers would apply much larger scale to plan the delivery of well-organized narrative speech and would require much more cognitive resource, but would limit the scale to considerably smaller size for sight reading. The results also support the flexibility of planning scale and cognitive threshold.

3. Analysis

3.1. The Rationale and Hypothesis

In the following analysis, 6 higher-level HPG units were selected, namely, PPh, BG, PG and correlating boundary breaks B3, B4 and B5.Our motivation came from two recent previous investigations of boundary discrimination. In one study we analyzed PPh level F0 height modulations and found HPG discourse positions PG-I/-F is more contrastive than PG-I/-M or PG-M/-F, supporting the significance PG and hence global level cross-over contrast [12]. In another study we examined boundary pause duration, pre- and post-boundary SYL duration and pre- and post-boundary SYL intensity as singular cues and paired them as relative cues and compared the degree of discrimination. We found (1) combined cues by two are more discriminative than any single cue, and (2) the most discriminative cue was the pair of pre-boundary SYL duration and its following boundary pause [11]. These findings led us to hypothesize that patterns of pitch, tempo and loudness by the PPh and their contrastive pattern may play a significant role in boundary discrimination in fluent continuous speech, and expect the patterns may differ for SpnNS and RS. Therefore, in the following sections we will examine 4 acoustic features by the PPh, namely, global F0, tempo, average intensity and boundary pause duration in relation to boundary identities B3, B4 and B5 of SpnNS and RS.

3.2. Global PPh F0 Pattern for Mandarin Chinese Using the Fujisaki Command Ap

Mandarin is a well-known tone language, using distinct F0 contour patterns by the syllable to denote lexical difference. As a result, there is both tone and phrase information and modulations in the output F0 of continuous speech. To accurately represent F0 information other than tones in the output, it is essential to remove lexically contributed modulations. We have done so successfully by training a low pass filter to automatically segment the F0 contour of RS by unit of annotated PPh using the Ap command of the Fujisaki model. [13][14][15] To test if raw speech data of SpnNS without phrase boundary tagging could be segmented by the same extraction, we first manually adjusted 30sec from the 1-hr SpnNS data for PPh boundaries and then used the results as training data. The setup of this low pass filter derived from manual training is as follows:

$$\text{lowpass\_output} = \text{filter}(80, 1, \text{F0});$$

where sampling rate of F0 is 100Hz

After one application of the low pass filter, the unit was adopted as a PPh and optimized for Ap, its following boundary adopted as B3. We then analyzed Ap discrimination among adopted B3 with annotated B4 and B5 to observe possible difference between NDLC and CNA.

3.3. Pause Duration & PPh tempo

The duration of silent pause at annotated boundary positions and respective PPh boundary break identities B3, B4 and B5 was extracted as acoustic features. PPh tempo was derived by duration of PPh and number of syllable and defined as follows.

$$Tempo = \frac{PPh\_duration}{Syllable\_Nmu}$$

where i is the i-th frame of PPh and N is the numbers of frames.

3.4. PPh Average Intensity

To extract average intensity by the PPh, a 32ms-frame was used to derive the amplitude where i is the i-th frame of PPh and N is the numbers of frames.

$$Intensity = \frac{\sum_{i=1}^{N} Frame\_energy_i}{Frame\_Num}$$

3.5. Contrastive Feature

Contrasts are defined as disparity between two adjacent units. Except for pause, contrastive patterns of PPh tempo, PPh average intensity and Ap were derived, as defined below.

$$Contrastive\_feature = pre\_feature - post\_feature$$

3.6. Statistical Analysis

The F-ratio is adopted as the major indicator to evaluate the discrimination among B3, B4 and B5.

$$F - ratio = \frac{\text{Inter\_class's variance}}{\text{Intra\_class variance}}$$

As a result, the bigger the F-ratio is, the more discriminative the analyzed feature would be for higher-level discourse boundaries.

4. Results

4.1. Cross-Data and –Speaker Comparison of Discourse Boundary Discrimination

4.1.1. Ap

The three panels in Figure1 summarize PPh Ap by data type and speaker. The results showed that overall PPh F0 is consistently discriminative for RS but not for SpnNS. The mean Ap of preceding and following boundaries in RS (CNA) showed opposite tendencies: the higher prosodic boundary showed less pre-Ap mean, whereas the post-Ap means presented opposite tendency of pre-Ap B3, B4 and B5. We note that relative features are more discriminative than single features Pre and Post in RS (CNA), and bigger contrasts are found between higher level B5 and lower level B3. However, the same pattern was not found for SpnNS (NDLC), as shown in the lowest panel of Figure 1.

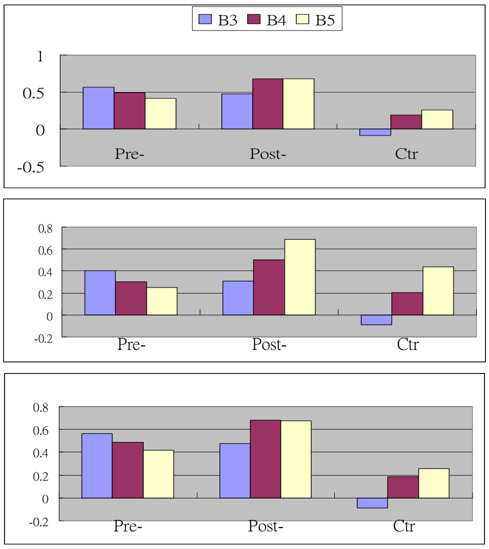

Figure1. *Cross-scale comparison of Ap mean patterns by corpus (from top panel to bottom are CNA_F051, CNA_M051 and NDLC). The horizontal axis represents indexes of feature type; the vertical axis denotes Ap mean values.*

4.1.2. Pause duration, Intensity and Tempo

Table2 lists the means of other prosodic features including pause duration, preceding Ap, preceding tempo, preceding intensity, contrastive Ap, contrastive tempo and contrastive intensity, cross-feature comparison is presented in 4.1.3 and plotted in Figure 2.

Table2. *Mean value of each prosodic feature in NDLC*

Break \ Feature	Pause	Pre-Ap	Pre-Tem	Pre-Int	ApCtr	TemCtr	IntCtr
B3	0.43	0.46	0.04	0.11	0.08	0.03	0.25
B4	0.74	0.27	-0.26	-0.71	-0.52	-0.13	-1.53
B5	1.15	0.23	-0.34	-0.65	-0.59	-0.52	-1.88

4.1.3. Cross-feature comparison by corpus

Results of cross-feature comparisons are presented in Figure2. The most discriminative feature for both CNA and NDCL to distinguish B3, B4 and B5 from each other is the combined cues of pause and relative intensity contrasts, showing how higher-level planning is reflected in the speech signal. Compared with CNA, discrimination by pause decreases in NDLC, but the discrimination of relative intensity is still obvious. Contrastive mean Ap is also discriminative to B3, B4 and B5 distinction. Similarly to our previous findings, singular prosodic features are less discriminative than contrastive features, thus provide further support to the significance of higher-level information in the prosodic domain.

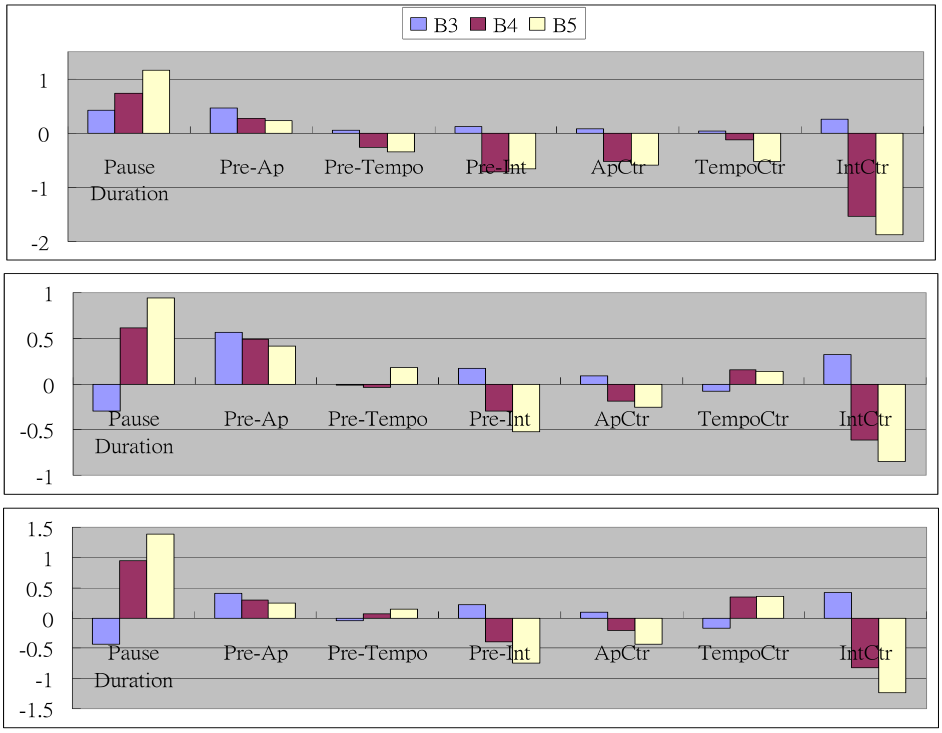

Figure2. *Cross-feature comparison of mean value by corpus (CNA_F051, CNA_M051 and NDLC from top to bottom; the horizontal axis represents indexes of feature type; the vertical axis denotes mean value of each feature.*

4.2. Discrimination Analysis

Table3. *Discrimination among B3, B4 and B5 (F-ratio) by prosodic feature and corpus*

Corpus \ Feature	Pause Duration	Pre-Ap	Pre-Tem	Pre-Int	ApCtr	TemCtr	IntCtr
NDLC	35.14	10.783	4.22	29.67	53.43	1.3527	132.59
CNA_f051p	204.5	18.786	3.0264	57.863	91.245	4.7692	98.26
CNA_m051p	720.95	35.51	3.5004	105.85	213.52	24.506	204.34

The above results also showed overall prosodic differences of SpnNS and RS. The most discriminative feature for boundary discrimination is boundary pause for RS (F-ratio=204.5 and 720.95, respectively), but relative intensity contrast for SpnNS, as evidenced in relative intensity (F-ratio=132.59). Interestingly, we found that the most discriminative contributions to both data sets are from 3 identical features, namely, pause duration, contrastive Ap and relative intensity, respectively. These 3 features are the most prominent cues to boundary discrimination; their respective distribution varies by corpus type, and their combined discrimination even more significant.

4.3. Discrimination by Relative Information

We further examined adjacency disparity to evaluate possible discrimination by contrastive patterns in relation to discourse boundaries B3, B4 and B5 whereby B3 is defined by HPG as within-paragraph boundaries while B4 and B5 are both between-paragraph boundaries by different degrees. Three conditions of pre- and post-unit contrasts were analyzed: Ap, intensity and tempo, as summarized in Table4, Table5 and Table6. The Table5 and Table6 showed that adjacent Ap and intensity contrasts distinguishes B3 from B4 and B5 across data type and speaker, implying pitch and loudness difference may differ PPh from BG and PG. However, the lower panel shows tempo contrasts only distinguishes the B5 from B3 and B4, implying PG is distinguishable from BG and PPh by tempo difference.

Table4. Distribution when post-Ap > pre-Ap

Break Corpus	NDLC	CNA_F051	CNA_M05
B3	34.98%	40.88%	37.93%
B4	86.59%	69.36%	76.67%
B5	91.67%	73.33%	86.72%

Table5. Distribution when post-intensity> pre-intensity

Break Corpus	NDLC	CNA_F051	CNA_M051
B3	36.52%	40..64%	35.46%
B4	97.56%	70.72%	78.44%
B5	100%	76.19%	90.52%

Table6. Distribution when post-tempo > pre-tempo

Break Corpus	NDLC	CNA_F051	CNA_M051
B3	50.68%	53.19%	55.79%
B4	56.10%	39.14%	36.53%
B5	91.67%	44.44%	34.48%

5. Discussion

From the Ap comparison between RS and SpnNS, results show the F0 contours and resets better discriminate discourse boundaries for RS, but not for SpnNS. This suggests F0 reset by PPh and declining contours are for RS. However, the large size of discourse unit and long stretch of discourse association in SpnNS has reduced and removed distinct PPh intonation patterns. In addition, contrastive intensity is more discriminative to SpnNS boundary identities whereas pause duration is more discriminative to RS. Lastly, pause, relative Ap and relative intensity are the most consistent cues for boundary discrimination across data type and speaker. These results collectively suggest that combined relative features are more discriminative than singular cues; higher level discourse planning can be derived in the speech signal.

6. Conclusions

Well-organized, highly communicative SpnNS is characterized by discourse planning scale, as evidenced in discourse unit size and non-distinctive pause durations. Patterns of paragraph phrasing and sub-paragraph association are marked by relative boundary intensity, instead of F0 resets and boundary pause duration found in RS. These patterns are a stark contrast to RS and spontaneous dialogue [4]. Consistent and significant to both SpnNS and RS is how combined global supra-segmental cues are more discriminative to discourse boundaries than any fragments of singular cue, and how higher-level discourse planning constrains and contribute to output prosody. These findings shed new lights on linguistic understanding of continuous speech and technology development of both speech processing and TTS.

7. References

[1] Nakamura, M., Furui, S., and I, Koji., "Acoustic and linguistic characterization of spontaneous speech", Proc. Symposium on Large-Scale Knowledge Resources (LKR2007), 163-168, Tokyo, Japan, 2007.

[2] Furui, S., "Recent progress in corpus-based spontaneous speech recognition", IEICE Transactions on Information and Systems, vol.E88-D, no.3, 366-375

[3] Furui, S., "Spontaneous speech recognition and summarization", The Second Baltic Conference on HUMAN LANGUAGE TECHNOLOGIES, 39-50

[4] Shriberg, E., "Spontaneous speech: how people really talk and why engineers should care", Proceedings of INTERSPEECH 2005, 1781-1784, Lisbon, Portugal, 2005.

[5] Watanabe, M., Hirose, K., Den, Y., Minematsu, N., "Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners", Speech Communication vol.50, 81－94, 2008

[6] Tseng, C., "Prosody Analysis", Advances in Chinese Spoken Language Processing, World Scientific Publishing, Singapore, 57-76, 2006.

[7] Tseng, C., Pin, S., Lee, Y., Wang, H., and Chen, Y., "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, 2005.

[8] Tseng, C., and Lee, Y., "Speech rate and prosody units: Evidence of interaction from Mandarin Chinese", Proceedings of the International Conference on Speech Prosody 2004, 251-254, 2004.

[9] Tseng, C., Cheng, Y., and Chang, C., Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech, Oriental COCOSDA 2005, Jakarata, Indonesia, 2005.

[10] Tseng, C., and Chang, C., "Pause or No Pause?—Phrase Boundaries Revisited". Tsinghua Science and Technology 13.4: 500-509, 2007.

[11] Tseng, C., and Su, Z., "Boundary and Lengthening—On Relative Phonetic Information.", The 8th Phonetics Conference of China and the International Symposium on Phonetic Frontiers, Beijing, China, 2008

[12] Tseng, C., and Su, Z., "Discourse Prosody Context--Global F0 and Tempo Modulations", Proceedings of INTERSPEECH 2008, 1200-1203, Brisbane, Australia, 2008.

[13] Fujisaki, H., and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", J.Acoust, J.Acoust. Soc. Jpn.(E), 1984; 5(4), 233-242, 1984.

[14] Mixdorff, H., "A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters", Proceedings of ICASSP 2000, vol. 3, 1281-1284, 2000.

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