Spontaneous Mandarin Speech Prosody—the NTU DSP Lecture Corpus

Chiu-yu Tseng*, Lin-shan Lee** and Zhao-yu Su*
*Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei, Taiwan
**Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
cytling@sinica.edu.tw
http://phslab.ling.sinica.edu.tw

Abstract

We report the collection of spontaneous Mandarin speech corpus the NTU DSP Lecture Corpus (NDLC) and preliminary results of comparative discourse prosody features between NDLC vs. read narratives CNA and CL from Sinica Mandarin Continuous Speech Prosody Corpora (COSPRO). Considerable differences were found in discourse scale, tempo modulations and discourse boundaries. Discourse topic segmentation by F0 contour features was also tested. The results suggest that large-scale discourse planning can be evidenced in large-size paragraph when communicative goals are structured by within paragraph cohesion and between paragraph associations. We will discuss the implications of lecture prosody to understanding spontaneous speech and to speech technology development.

Introduction

Using corpora of read speech (RS) recorded in soundproof chambers Tseng et al [1] [2] have studied narrative discourse prosody extensively and constructed a hierarchical framework the HPG (Hierarchy of Prosodic Phrase Group) to explain the generation of discourse prosody output with quantitative evidences of corpus analyses. The layered HPG prosodic units from bottom up 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 expressed as SYL<PW<PPh<BG<PG and B1<B2<B3<B4<B5. Three characteristics referring to discourse prosodic units and their functions distinguish the HPG from other prosody studies. (1.) The HPG specifies a PPh as an intonation unit (IU) that corresponds to a clause or a simple sentence within a complex sentence; a BG as a multiple-IU unit that corresponds to a complex sentence where a change of breath occurs at its end; and a PG a multiple-BG unit that corresponds to a speech paragraph where default change of breath also occurs. (2.) The HPG also specifies boundary breaks corresponding to the HPG prosodic units as prosodic units well [1, 2]. (3.) At the PG level, the HPG takes into account higher level discourse semantic association by three PG-positions by the PPh, namely, PG-initial, -medial and –final, thus specifying both linear and cross-over prosodic relationships of both a complex sentence and a speech paragraph. Both the PG-initial and –final PPh are major phrases while PG-medial minor phrase(s). But the prosodic functions of the PG-initial and –final PPh are distinct from each other and hence their prosodic properties are significantly different [1] [2]. The superposition of BG and PG onto the PPh's requires the within-PG PPhs to adjust respective intonation contour patterns, overall rate of speech and amplitude distributions systematically and by the HPG layers to achieve paragraph prosody.

Based on evidences obtained since 2005, Tseng and her group believe the HPG tree structure represents the base form of discourse prosody generation similar to the way the syntactic tree structure generates the sentence, used by the speaker as templates for both on-line speech planning and speech processing. If it is indeed the case, evidences could also be found from speech forms other than RS should also provide evidences

In the present study, we further test the HPG framework on spontaneous speech (SpnS) by analyzing both SpnS of university classroom lecture and two types of RS. Our aim is to look for patterns of consistency and divergence between read and SpnS, and at the same time to derive prosodic characteristics of spontaneous discourse prosody as well.

Reported features of classroom lectures are that lectures are not always practiced in advance; the same phrases are repeated many times for emphasis; the spontaneity of this kind of speech is much higher than other kinds of presentations; NTU DSP Lecture Corpus (NDLC) and the lecture speaking style is closer to that in dialogue because lecturers are always ready to be interrupted by questions from students [3]. We noted that the NDLC lectures are not interactive and do not possess dialogue features. Other reported features of classroom lectures are that the vocabulary may be very specific in classroom lectures and that lecture transcription is characterized by strong co-articulation effects, non-grammatical constructions, hesitations, repetitions, filled pauses, etc [4]. Preliminary observations only confirmed some of these features. However, since our focus is on prosody features, we will not discuss these features in the present paper.

Our rationale of choosing university classroom lectures is as follows: (1.) in comparison with RS, course lectures offer samples of well-planned and well-structured discourse in fluent narratives filled with distinct discourse functions such as topics, topic change, topic associations, parallel and parenthetical structures, discourse markers, and prosodic fillers. (2.) In comparison with other types of SpnS such as dialogue, it provides clearer prosodic patterns reflecting higher discourse level planning.

In the following sections, we will present three analyses to examine and compare (1.) scales of discourse planning across speech type, (2.) PPh tempo modulations by discourse boundaries, and (3.) topic change by overall PPh F0 contour patterns.

Speech materials

A small portion of speech data from the National Taiwan University (NTU) DSP Lecture Corpus (hence NDLC) was used for the present study. The corpus consists of a total of 45 classroom lectures of a NTU (National Taiwan University) DSP course taught by one of the co-authors Professor Lin-shan Lee in three formats: audio data (3.92GB in 45 waves of approximately 90 minutes apiece), video data and the Powerpoint slides used. The speech data is transcribed into text using NTU developed Mandarin ASR and a language model. Each of the 45 lectures is also segmented into sections in alignment with slides presented. All three formats can be accessed on-line at http://speech.ee.ntu.edu.tw/~duidae/interface/mytesttube/index.php. Tseng and her group acquired the data to explore the discourse prosody features of spontaneous Mandarin lecturing and their linguistic significance.

For SpnS, one of the 90-minute 45 lectures was used, totaling around 41,000 syllables. Two highly experienced transcribers manually tagged into three discourse prosody units by the HPG protocol: PG, BG and PPh and correlating discourse boundary breaks B3, B4 and B5. The transcribers' initial impression of the NDLC includes the following: (1.) that the speaker's articulation is clear and narrative presentation lucid, (2.) that less paralinguistic and non-linguistic elements occurred than observed spontaneous dialogue speech, (3.) that relatively large discourse units and topic change are clearly perceived and (4.) that the speaker makes abundant use of discourse marker, parallel structure, prosodic fillers as well as focus and prominence.

For RS, two types of data from the Sinica COSPRO (Sinica Mandarin *Co*ntinuous *S*peech *Pr*osody Corpora) were used to represent two different types of reading style, namely, CNA and CL. CAN stands for reading of plain text of 26 random discourse pieces totaling around 12,000 syllables. CL stands for reading of Chinese Classics differing in rhyming structure from regular, semi-regular to irregular, totaling around 3,500 syllables. Speech data of one male and one female for each type was selected, namely, M051 and F051 for CNA and M056 and F054 for CL. These RS 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.

Analysis Method

One index for estimating the planning scale of discourse unit in NDLC and two features for tempo modulation modeling and topic segmentation are defined for the following three analyses.

Analysis1. Discourse Planning Scale

By the HPG definition, one BG corresponds to a multiple-phrase complex sentence, and one PG a multiple-sentence speech paragraph. Therefore, the numbers of BG's within one PG are good indicators of the size of the PG. We define R as the indicator of discourse planning scale by the average number of BG's in one PG. In other words,

R = Number of B4 / Number of B5

Analysis2. PPh Tempo Modulations by HPG Boundaries

Averaged syllable duration is used to represent speaking rate. Each PPh preceding annotated HPG boundary breaks B3, B4 and B5 are extracted as indicators of overall phrase level tempo modulation. We then compared the derived PPh duration by the above three boundary breaks.[6][7] [8]. We noted that extreme variations of syllable duration exist in the data sets, and devised a more refined definition to derive mean syllable duration. That is, token syllables with the maximum and minimum of durations in the PPhs were 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..

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

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

Subsequently, PPh duration average was compared by HPG boundary breaks B3, B4 and B5.

Analysis3. Global F0 Contour Patterns by the PPh

Global F0 contour patterns by the HPG prosodic unit the PPh are used as indicators of possible change of topic. To extract global PPh F0 contour patterns, we adopt the filter-based method by Mixdorff [9][10] but modified the parameters of low pass filtering into two layers to fit the lecture corpus. The two layered parameter setups are described below.

1st layer low pass filter parameters setup

Para1 = ones(1,600)/600

Output1 = filter(Para1, 1, F0_value)

2nd layer low pass filter parameters setup

Para2 = ones(1,500)/500

Output2 = filter(Para2, 1, output1)

To smooth the output of the 1st layer lowpass filtering, the 2nd layer lowpass filtering was implemented to derive the overall tendency of F0 contour patterns. .

Results

Analysis1. Discourse Planning Scale

Table 1 summarizes the speech data used for the present study for the present study. The scale of discourse planning is analyzed by HPG layers. The mean length of BG and PG in seconds is summarized in Tables 2 and 3; and mean number of syllables per BG and PG summarized in Tables 4 and 5.

Table 1 List of syllable numbers by data type and speaker

Corpus	Lecture	CNA		CL	
Speaker	LSL	F051	M051	F054	M056
SYL #	41108	11592	11600	3502	3510

Table 2 mean length per BG by data type and speaker

Corpus	NDLC	CNA		CL	
Speaker	LSL	F051	M051	F054	M056
Mean(Sec/BG)	17.82305	5.785467	6.3121	3.8982	3.37951
StaDv(Sec/BG)	11.24641	3.474952	4.0115	1.3348	1.362523

Table 3 mean length per PG by data type and speaker

Corpus	NDLC	CNA		CL	
Speaker	LSL	F051	M051	F054	M056
Mean(Sec/PG)	178.793	7.01032	6.8516	4.404	3.51766
StaDv(Sec/PG)	137.466	6.39666	6.2803	1.588	1.517

As shown in Tables 2 and 3, one characteristic that distinguishes the NDLC spontaneous lecture and RS CAN and CL is how long the discourse units BG and PG could be. The mean length of BG is approximately 18 seconds for NDLC, 6 for CAN and 3/4 CL, respectively; while the mean length of PG is approximately 180 seconds (or 3 minutes) for NDLC, 7 for CAN and 4 for CL, respectively. The results show that in SpnS, a complex sentence BG could be 3 to 4.5 times longer than that of text reading, while a speech paragraph 26 to 40 times longer. Moreover, Tables 4 and 5 show the size per BG and PG in mean number of syllables.

Table 4 mean number of syllables of BG by data type and speaker

Corpus	NDLC	CNA		CL	
Speaker	LSL	F051	M051	F054	M056
Mean(Syl/BG)	103.87	38.966	42.89	18.64	18.781
StaDv(Syl/BG	71.2424	28.031	32.6	10.14	11.682

Table 5 mean number of syllables of PG by data type and speaker

Corpus	NDLC	CNA		CL	
Speaker	LSL	F051	M051	F054	M056
Mean(Syl/PG)	653.0926	76.761	90.195	43.55	56.957
StaDv(Syl/PG)	406.7074	45.260	66.723	20.06	29.672

Table 4 shows the mean syllable numbers of BG is approximately 110 for NDLC, 40/43 for CNA and 19 for CL, respectively; while syllables numbers of PG is approximately 653 for NDLC, 77/90 for CNA and 44/57 for CL, respectively. In other words, the mean syllable numbers within one BG in NDLC are about 2.5 and 5.5 times to those in CNA and CL; the mean syllable numbers within one PG in NDLC are about 8 and 12 times to those in CNA and CL, respectively. Furthermore, the average number of BG within one PG is derived and shown as R in Table 6. The results show that there are equivalence of 6 complex sentences in one speech paragraph for NDLC and anywhere between 2 to 3 for CNA and CL.

Table 6 mean number R of BG within one PG by data type and speaker

Corpus	NDLC	CNA		CL	
Speaker	LSL	F051	M051	F054	M056
R	6.2875	1.969958	2.102948	2.33631	3.03260

In summary, the above results show that the scale of discourse planning unit is relatively much larger than reading of text.

Analysis 2 Tempo Modulations by-B3, B4 and B5

The results of the comparison of overall pre-boundary PPh tempo modifications are plotted in Figure 1. Three classes of PPh by discourse boundaries were compared for mean syllable duration, i.e, Pre-B3, Pre-B4 and Pre-B5. All mean values were normalized for cross-speaker comparisons. We note that the speaking rate of pre-boundary PPh for both types of RS (CNA and CL) exhibit similar overall patterns across the 4 speakers, namely, Pre-B3<Pre-B4<Pre-B5, denoting that for RS the higher the level of discourse boundary, the slower the speaking rate for the preceding PPh. Or, overall slowing corresponds of higher-level ending is systematic. However, the patterns derived from the NDLC exhibit a reverse pattern between the Pre-B3 and Pre-B4 PPh, causing an overall change of pattern to Pre-B4<Pre-B3<Pre-B5. The latter results suggest that when physiological constraint calls for a change of breath, on-going effect is signaled with faster speaking rate. The speaker allowed overall PPh slow-down only in the termination of a long speech paragraph at the Pre-B5 positions. We believe this is a distinct feature of spontaneous lecture speech (SpnLS).

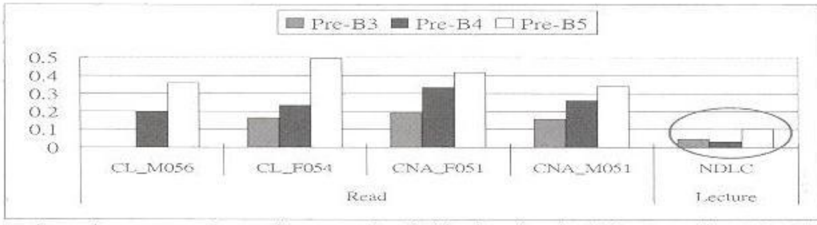

Figure1. Cross boundary comparison of averaged syllable duration by PPh preceding B3, B4 and B5. The horizontal axis represents index of the speech data and speaker. The vertical axis denotes normalized mean syllable duration of PPh preceding boundaries B3, B4 and B5.

Analysis 3 Topic Segmentation by Overall PPh F0 Contour Pattern

Figure2 presents an example of speech segment from the NDLC where a PG, followed by a B5 and a portion of the another PG is shown. The original F0 contour was plotted in red while outputs of the 1st and 2nd layer of lowpass filtering are plotted in blue and green respectively. Thus, the overall tendency of a series of F0 contours within and cross-over a PG can be observed. In this example, the PG end and boundary B5 appear at the relatively low values of the 2nd layer lowpass filter shown in green.

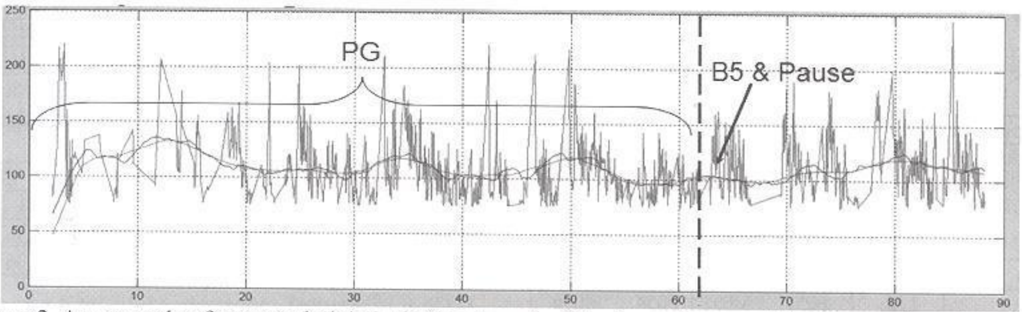

Figure2. An example of extracted global of F0 contour by PPh shows the overall tendency of a series of F0 contours within and across a PG. The horizontal and vertical axes represent time and F0 values respectively. The red plotting denotes original F0 contours by PPh. The blue and green plotting represent the output of the 1st and 2nd layer lowpass filters, respectively.

Discussion and Conclusions

Using discourse prosody specifications by the HPG framework, we analyzed and compared discourse prosodic features between spontaneous classroom lecture and reading of text in three measurements to test whether the RS based HPG discourse framework also applies to spontaneous lecture and what major prosodic characteristics distinguish spontaneous lecture from RS. Three measurements were used, namely, scales of discourse planning, phrase level speaking rate modulations in relation to discourse boundaries, and overall PPh F0 contour patterns in relation to topic change. We found that SpnLS is characterized by relatively large-size speech paragraphs consisting of more complex sentences, indicating that well-planned SpnLS with clearly structured themes of information is distinctly different from RS, whereby the latter consists of relatively small-size paragraph of lesser number of complex sentences and less complex sentences. In all the parameters analyzed, systematic modulation of phrase level speaking rate and distinct paragraph boundary properties are found in which association of multiple complex sentences to form paragraph is characterized by faster speaking rate between sentences to indicate the on-going effect. Paragraph associations are evidenced in overall F0 ending patterns with the following paragraph breaks which collectively indicate topic change.

From the above analyses, we show that the HPG prosody framework is sufficient to account for discourse planning for both SpnS and RS. Differences between RS and SpnLS can be attributed to different distribution patterns by the HPG levels and units. The HPG framework is further proven to be the base form of discourse prosody planning. We believe the above results could also be readily applied to technology development such as TTS to improve overall output naturalness.

Future works include more comparative analysis of NDLC and other types of RS from the Sinica COSPRO in discourse prosody in order to construct a more comprehensive account of SpnLS and RS, key word spotting by prosodic features and also prosodic template fitting of discourse units.

Reference

[1] Tseng, C-Y., Pin, S-H., Lee, Y-L., Wang, H-M. and Chen, Y-C., 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.

[2] Tseng, C-Y. 2006. "Prosody Analysis" *Advances in Chinese Spoken Language Processing,* Lee, C., Li, H-Z., Lee, S-S., Wang, R.-H and Huo, Q. (eds) World Scientific Publishing, 57-76, Singapore.

[3] Yamazaki, H., Iwano, K., Shinoda, K., Furui, S. and Yokota, H. 2007. "Dynamic Language Model Adaptation Using Presentation Slides for Lecture Speech Recognition." In *INTERSPEECH 2007,* August 27-31, Antwerp, Belgium.

[4] Trancoso, I., Nunes, R., Neves, L., Viana, C., Moniz, H. and Diamantino Caseiro, D. and Mata, A. I., 2006. "Recognition of Classroom Lectures in European Portuguese". In *INTERSPEECH 2006*, Pittsburgh, PA, USA.

[5] Tseng, C-Y., Cheng, Y-C., and Chang, C-H., "2005. Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech". In *Oriental COCOSDA 2005,* Dec. 6-8, 2005, Jakarata, Indonesia.

[6] Tseng, C-Y. and Su, Z-Y., 2008. "Boundary and Lengthening—On Relative Phonetic Information". In *The 8th Phonetics Conference of China and the International Symposium on Phonetic Frontiers*, April 18-20, 2008, Bejing, China.

[7] Tseng, C-Y., and Su, Z-Y., 2008. "Discourse Prosody and Context – Global F0 and Tempo Modulations". In *Interspeech 2008,* 1200-1203. Brisbane, Australia.

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

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

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