Prosodic Patterns of Information Structure in Spoken Discourse—a Preliminary Study of Mandarin Spontaneous Lecture vs. Read Speech

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*

Abstract

The aim of the study is to explore the prosodic patterns spontaneous lecture speech vs. read speech to show where and how these monologues differ and why by analyzing perceived emphasis and its acoustic features within and between speech paragraphs. Systematic but distinct patterns are found for both speech types in emphasis distribution across speech, overall and local tempo modulations. Read speech is characterized by discourse coherence while spontaneous information structure in addition. Intricate tempo modulations characterizing information structure are discussed.

1. Introduction

In this paper, we present a preliminary comparative prosodic analysis of spontaneous lecture speech (SpnL) vs. read speech (RS) to show how these monologues differ from each other prosodic organization and patterns and what the differences found imply. Previously, we have been studying discourse coherence patterns of RS from the perspective of chunking and paragraph phrasing, focusing on the prosodic associative patterns constrained by discourse coherence; and was able to derive correlative prosodic templates. [1] We adopt the same perspective in subsequent investigation of SpnL and thus diverge ourselves from linguistic features of disfluency, repetition, repair, hesitation, fillers and filled pauses in most reported works. [2] The rationale is that both SpnL and RS are monologues that involve strategic discourse planning. RS planning is passive, focusing mainly on thematic and semantic coherence; while SpnL planning is more active, focusing on massage delivery. Our recent perceptual analysis of discourse association showed that the same patterns of paragraph phrasing and discourse association apply to both RS and SpnL, but in temporal terms the average paragraph size of SpnL is approximately 3 times that of RS; and in number of syllables 8 times that of RS [3]. In other words, themes of significance are delivered in much larger units, and the speaker must utilize prosodic means to sustain listeners' attention. In the sections below, we will present analysis of perceived emphasis in RS and SpnL as a reference of information allocation and planning. Issues under examination include emphasis distribution in relation occurrence, position in phrase and paragraph; while acoustic analysis include tempo/rate modulations, pitch and intensity patterns.

2. Speech and Materials and Pre-processing

The speech materials used are from the following two sources: (1) the NTU DSP Lecture Corpus (hence NDLC) which contains 45 hours of microphone speech from 15 3-hr sessions of classroom lectures on Digital Speech Processing (DSP) from National Taiwan University (NTU), 1 speaker, 3.92GB in total, and (2) the Sinica COSPRO (Sinica Mandarin *Co*ntinuous *S*peech *Pr*osody Corpora) which contains read speech of 8 types of text designed to represent a variety of styles, multiple speakers, 7.9GB in total. One hour of the NDLC (14,305 syllables total) and 26 random discourse pieces of CNA (120 minutes around 12,000 syllables) female speech (F051) [3] were selected as samples of SpnL and RS.

The selected speech data were manually processed by trained transcribers for perceived discourse boundaries and phrasing units using the HPG (Hierarchical Prosodic phrase Grouping) platform and annotation protocol [4]. The HPG framework allows extraction of prosodic features by layer and quantitative account of contribution by cumulative account. The prosodic units from the bottom up the hierarchy are the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath-group (BG) which denotes a physio-linguistic unit of breathing arrangement during speaking, and the multiple-phrase prosodic group (PG) which denotes a speech paragraph. The relationships among these units are SYL<PW<PPh<BG<PG; their respective boundaries B1, B2, B3 B4 and B5. A PG is further specified by three relative positions that denote paragraph coherence, namely, the PG-initial (PG-I), -medial (-M) and –final (F). In addition to the HPG annotations, perceived emphases were manually tagged by trained transcribers for the present study. The rationale of is to include both phrasal prominence and any other stressed and accentuated sections, most of them are keywords occurred as PWs, across the speech flow as reference of essential information. Pre-processing results showed that in temporal terms the average length of BG (change of breath during speaking) is approximately 18 seconds in SpnL and 6 seconds in RS; while the average length of PG with multiple changes of breath is approximately 180 seconds (or 3 minutes) in SpnL and 7 seconds in RS, respectively. The results entail that in SpnL a coherent speech paragraph PG is 26 times that of passive reading, in the paragraph each breathing cycle BG 3 times the duration of text reading. By syllable number the average of BG is around 110 for SpnL and 43 for RS, respectively; while the average syllable number of PG is approximately 653 for SpnL and 90 for RS, respectively. This implies that average syllable numbers spoken in one BG in SpnL are about 2.5 times to RS while a coherent speech paragraph PG may take 8 times of change of breath to that of RS, respectively. [3] The sheer size of the speech paragraph in SpnL motivated us to apply the same concept of associative positions (–I, -M and –F) to the PPh level to accommodate better account of emphatic portions of the speech data.

3. Analysis

3.1. Rationale and hypothesis

We hypothesize that the planning of SpnL is more complex than RS and involves more elaborate planning of IS in addition to discourse structure (DS). While RS is passive planning of the speaker who reads out loud to express mostly discourse coherence (DC); SpnL requires more elaborate planning of information structure (IS) in addition to DC. By IS we adopt a broad view to mean roughly structural and semantic properties of utterances relating to the discourse content, the actual and attributed attention states of the discourse participants, and the participants' attitudes, thus notions like focus, presupposition, given vs. new, theme vs. rheme and the various dichotomies such as topic vs. comment or focus, ground or background vs. focus, etc. are subsumed. [5] Our goal is to derive IS related prosodic patterns through perceptual and acoustic analysis of emphasis. Since patterns of pitch reset, duration modulation and loudness control are directly related to perceptual contrasts; their respective acoustic correlates the F0, duration and amplitude patterns will be examined. We further hypothesize that keywords and their occurrence in speech is syntax and discourse governed as well as speaker intended, thus reflecting more complex interaction of phrase-level (syntactic) and higher-level (discourse) planning of IS. The following analyses are thus two-fold, one analysis aims to compare the similarity and diversity between RS and SpnL for information distribution; another to derive acoustic and prosodic patterns that are directly related to allocation of information.

Figure1 shows the distribution of perceived emphasis in RS and SpnL by 3 relative positions at the phrase level, namely, at the PPh-Initial, -Medial and –Final; as well as in relation to same-level prosodic boundaries B3 and higher-level boundaries B4 and B5.

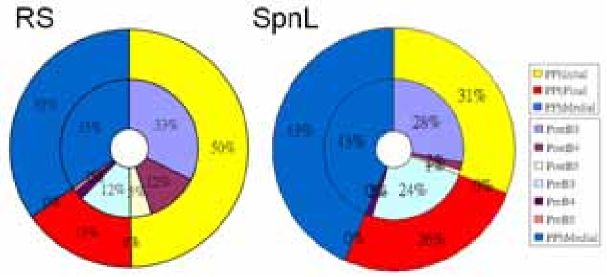


Figure 1 .The distribution of emphasis by discourse associative position, boundary location and speech data type. The outer circle shows the distribution of emphasis by associative positions PPh-Initial,-Medial and Final. The inner circle shows the distribution of emphasis before and after PPh- local boundary B3 and higher-level boundaries B4 and B5, respectively.

In the following sections, we will focus on acoustic analysis of emphasis with respect to duration and tempo patterns by phrase and by speech paragraph, but less on F0 and intensity patterns.

3.2. Normalized position of emphasis

Due to different sizes of both the PPhs and BGs in the data sets, we first normalized the position of emphasis and plotted the distribution of number of emphasis in PPhs and BGs analyzed (Figure2). The normalization results enable us to better examine the location as well as allocation of emphasis in various PPhs/BGs. The equation of normalization is as follows.

$$NEmP = (EmP - BGS)/BGD \qquad (1)$$

where *Emp* and *NEmp* denotes original and converted emphasis position. *BGS* and *BGD* represent the onset time of *BG* and the duration of BG, respectively.

3.3. Tempo modulation of emphasis regarding discourse information

By tempo modulation we mean overall change of speaking rate by phrase in relation to each change of breathing cycle. Tempo modulation was examined by each phrase and by the number of emphases contained, length of the phrase, phrase position in a breathing cycle, and the duration of the one breath. A linear regression (LR) model of syllable duration was adopted to extract duration pattern by phrase [1], and parameters were modified to accommodate phrase level features. The rate by PPh is extracted and compared with the number of emphasis contained, PPh length, PPh position within BG and BG length. Below is the LR model for PPh tempo features.

$$T = f(EMN, PPhLen, BGP, BGL) + res \qquad (2)$$

where *f* denotes linear regression by multiple variables, *T* denotes the regression values for tempo feature of a current PPh and *res* denotes error in comparison with original values; *EMN, PPhLen* and *BGP BGL* denote the number of emphasis contained in the current PPh, the length of current PPh, the position in current BG and the length of current BG, respectively.

3.4. Tempo of emphasis regarding rate of phrase

In addition to tempo feature of emphasis defined in relation to discourse information (3.3.), a relative tempo feature of the emphasis itself is also defined by measuring the normalized rate of emphasis against the overall rate of its embedding phrase. The proposed relative measurements have been proven to better account the contrastive nature of supra-segmental features [6] and provide clearer picture of the speech data.

$$RTPEM = TPEM - TPPPh \qquad (3)$$

where *TPEM* and *RTPEM* denote the original tempo (derived in 3.3.) and relative tempo for emphasis, respectively. *TPPPH* is the original tempo feature of PPh in which the emphasis is embedded in.

4. Results

4.1. Emphasis (perceived keyword) distribution in BGs

Patterns of emphasis distribution between RS and SpnL are derived and shown in Figure 2. In RS, maximum distribution of emphasis is at the BG onset of BG and descends with BG positions, with least emphasis at the offset. In SpnL, the distribution of emphasis in SpnL assumes a pattern similar to Gaussian mixture model, where minimum emphasis occurs at the BG onset, two peak distributions in BG medium positions, and maximum distribution at the BG offset. In other words, two distinct patterns are found: emphasis in RS is at the beginning of the paragraph, and never at the end. Whereas in SpnL the pattern is almost reverse where emphasis never occurs at the beginning but with two high occurrences in the middle, spread across the speech paragraph, and marks the paragraph end. These results are interpreted as indication of key information distribution.

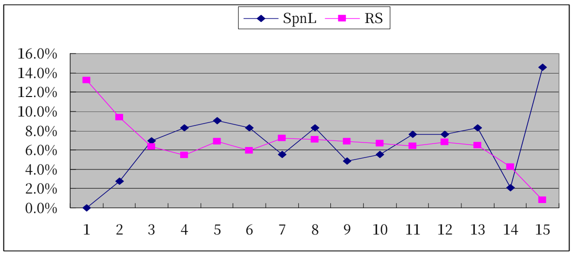


Figure 2 .The distribution of emphasis (perceived keywords) by BG-position and speech data type. The horizontal axis represents relative position in BGs. The vertical axis represents the percentage of number of emphasis in current position.

4.2. Emphasis (perceived keyword) distribution in PPhs

The same patterns of emphasis distribution are also derived by the PPh for both data sets, as shown in Figure 3. At the phrase level, though emphasis occurs at the onset for both SpnL and RS, the distribution differs. Nearly 45% of the phrases of SpnL begin with emphasis while only about 25% of RS phrases assume the same pattern. This implies that well-organized lecture speech has higher probability to emphasize at the beginning of PPhs.

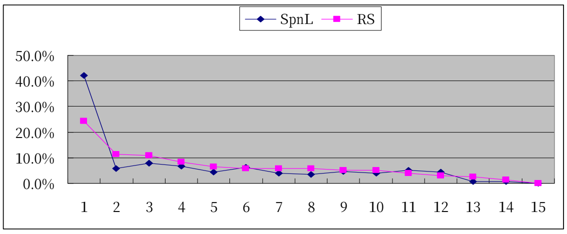
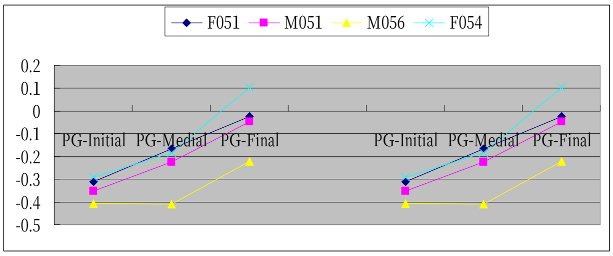


Figure 3 .The distribution of emphasis (perceived keywords) by PPh-position and speech data type. The horizontal axis represents relative position in PPhs. The vertical axis represents the percentage containing emphasis in current position.

4.3. Tempo patterns of the paragraph unit BG

Figure 4 shows tempo modulation within and between the paragraph unit BG. For RS, the overall tempo pattern of the speech paragraph is to start fast and gradually slow down until the end. This fast-to-slow continuum also creates the sharpest slow-to-fast contrast between paragraph boundaries. For SpnL, the overall paragraph is different, the fast-slow contrast is not a continuum, but hill shaped with the paragraph beginning in medium rate, slowing down until before the middle of the paragraph; then accelerating to end the paragraph at the fastest rate. The fast-to-medium rate contrast between paragraph boundaries is also sharp.

SpnL



RS

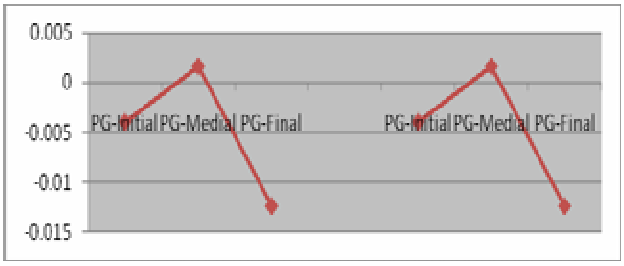


Figure 4.Tempo allocation patterns by BG-position and speech data type. The upper and lower panels denote tempo allocations of BGs for RS and SpnL respectively. The horizontal axis represents the relative position in BGs. The vertical axis represents the normalized mean value of PPh tempo by BG-position.

We further examined tempo modulation of shorter paragraphs in SpnL for more detailed information and as a reference of BG length. Figure 5 shows the tempo patterns of BGs of 7 or less PPhs. The results show that the slowest rate, also occur in the middle of the paragraph.

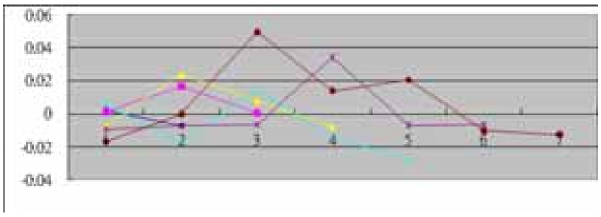


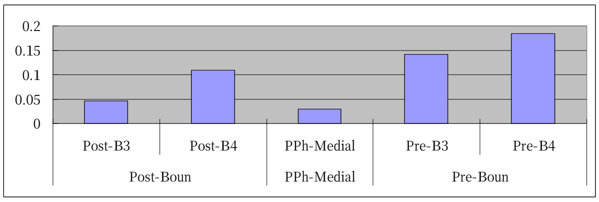
Figure5. Tempo allocation patterns of BG consisting of 7 or less phrases by BG-position in SpnL. The horizontal axis represents the relative position in BGs. The vertical axis represents the normalized mean value of PPh tempo by BG-position.

In other words, tempo modulation of RS and SpnL differs distinctly. For RS the slowest rate implies paragraph ending while for SpnL it implies a dividing point.

4.4. Tempo patterns of emphasis by location and boundary type

Since boundary properties, commonly referred to as lengthening, are discourse constrained [6]; the rates of emphasis in both RS and SpnL are analyzed with respect to the tempo to the current PPh where the emphasis occurs, and with regard to boundary type (Figure 6). The results show that emphasis in SpnL is slower than the current PPh tempo regardless of boundary type and positions. However, no consistent tempo pattern of emphasis is found in RS: it (emphasis) is faster than the current PPh tempo in all positions and longer before boundaries. The post-boundary slowing down is clearly attributed to boundary lengthening. These results illustrate that the tempo modulation is a more important acoustic cues for emphasis in SpnL than in RS.

SpnL



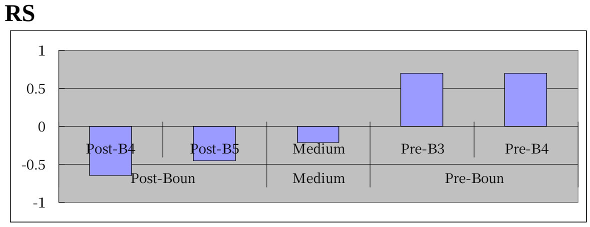


Figure6. The relative tempo of emphasis by position in PPh and discourse boundary type. The upper and lower panels denote tempo relative tempo of emphasis for SpnL and RS respectively. The horizontal axis represents position PPh and boundary type. The vertical axis represents the relative tempo of emphasis and zero means the tempo of emphasis is equal to current PPh tempo.

4.5. The distribution of F0 and intensity of BGs in SpnL

In relation to results of tempo analysis of SpnL (see 4.3 and lower panel of Figure 4) in which the slowest rate implies a dividing point corresponding to two high occurrences of perceived emphasis distribution (4.1), the distribution of mean F0 and intensity are also analyzed in search of corresponding dividing points (Figure 7). However, the results showed only overall declination of F0 and intensity by relative BG position, while no correlating patterns are found. These results suggest that perceived emphasis in Mandarin monologue is related mostly to tempo modulations rather than to F0 or intensity settings.

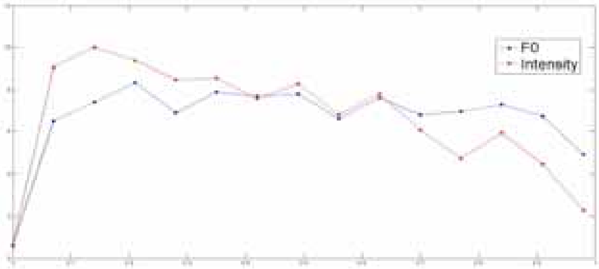


Figure7. Distributions of mean F0 and intensity by BG positions. The horizontal axis represents relative position in BGs. The vertical axis represents the percentage of values larger than mean in current position.

5. Discussion

The patterns of emphasis in RS are simpler and similar at both the emphasis-local and higher layers. Most of the emphasis coincides with paragraph prominence at the onset a discourse unit. Prosodically phrasal prominence is the most explicit slot by default, and phrase end is the least explicit slot where only less than 25% occurred due most likely to syntactic and/or semantic make-up. RS tempo modulations are the same with or without account of emphatic portions; indicating default emphasis does not trigger prosodic alternations in addition to discourse coherence. In turn, emphasis itself is not marked by tempo modulation. On the contrary, emphasis in SpnL speech exhibited distinctly different patterns. Instead of occurring at the prosodically most explicit position by default, it occurs from the mid-paragraph and across the board, and marks the paragraph end by its high distribution at the paragraph end. In other words, the occurrences mark the most explicit prosodic expressions at the least explicit locations; making the (the occurrences) stand out by the sharper contrasts they create. The highest occurrence at the pre-boundary (terminal) position is the prosodic highlight to reiterate most explicitly the most important information at the least explicit slot; hence overrides boundary lengthening regardless of phrase or paragraph ending. Much more complicated tempo modulations accompany such prosodic highlights, both by the emphasis itself and by the phrases that embeds them. An emphasis always assumes the slowest rate; the entire embedding phrase also slows down. This suggests the speaker's intended loading and weighting of information, expressed via prosodic explicit means, is a direct reflection of IS [7] planning on top of DS. Our results clearly suggest that systematic prosodic manipulations to signal the implicit/explicit, given vs. new, theme vs. rheme contrasts can be located in the speech signal. Furthermore, in acoustic terms, information structuring in the prosodic domain appears to have most to do with tempo modulations, and much less with F0 and intensity.

6. Conclusions

Different tempo patterns of prominence, represented by perceived emphasis, are found for RS and SpnL by tempo, pitch and intensity features, both in the emphasis-local PPh level and the higher-level BG level. Acoustic patterns of emphatic portions in RS and SpnL monologues are analyzed in relation to discourse structure. In RS most of the emphasis coincides with phrasal prominence with no noted tempo change; discourse coherence remains the predominant feature. Contrastively, SpnL is featured by more intricate tempo modulation patterns at various levels to signal explicitly and unambiguously at the more implicit locations, most notably at the paragraph end, done at the trade-off of unit-final lengthening. These results showed that IS and speaker's intensions would override DS. We believe these findings have furthered our understanding of the organization of SpnL in the prosodic domain. In particular, the fact that tempo patterns extracted at various prosodic domains could be applied to keyword spotting and topic change. Future work will focus on more detailed parallel analysis of prosodic patterns and IS.

7. References

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

[3] Tseng, C. Su, Z. and Lee, L. 2009. Mandarin Spontaneous Narrative Planning—Prosodic Evidence from National Taiwan University Lecture Corpus . Proc. Interspeech2009, Brighton, UK. 2943-2946. .

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

[5] Kruijff-Korbayava, I. and Steedman, M. "Discourse and Information Structure", J. of Logic, Language and Information 12:249-259, 2003

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

[7] La mbrecht, K. Information Structure and Sentence Form—Topic, focus and the mental representations of discourse referents. Cambridge Studies in Linguistics, 1994.