Information Structure by Way of Discourse Prosody

Chiu-yu Tseng

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

Abstract

Our previous research findings have shown that global paragraph or discourse prosody reflects overall between and cross phrase association. This paper presents evidence of how additional focal information is layered over global association, in the form of prominence, in order to reflect additional message from information structure as well as speaker intension while maintaining discourse coherence. Data of read and spontaneous Mandarin speech are analyzed to illustrate (1) how focus or prominence represented tagged as perceived emphasis would result in difference in tempo patterns, (2) how perceived emphasis are genre conditioned, (3) what rhythm, pitch and intensity features are in the emphasis-local phrase level and higher discourse level; (4) how acoustic patterns of emphasized portions in output speech could be better analyzed to reveal the existence and influence of discourse structure, (5) how in the acoustic domain, the assigned focal chunks are highlighted to signal redundant information through contrasts with their immediate neighborhood; and how the degree of contrast reflects information weighting on the one hand while making the

highlighted chunks perceptually more salient on the other. Results also show that placement of prosodic highlights is systematic while discourse coherence is always maintained. It is argued here that the information structure in global discourse prosody directly reflects speech planning and delivers the communicative focal points in the speech signal. The planning itself is of course systematic but highly flexible.

I. Why Discourse Prosody?

It is well accepted that utterances are phrased into constituents and hierarchically organized into various domains at different levels of the prosodic organization (Shattuck-Hufnagel & Turk 1996, Gussenhoven 1997, Selkirk 2000, Cutler & Butterfield 1992). The focus has been on phrase intonation making simple sentences or short phrases the default unit of study. We noted that evidence of larger scale prosody planning by phrase groups were found not through investigation of individual intonation but through pause and timing structure in continuous speech (Keller et al. 1993, Zellner 1994). Our goal of studying Mandarin Chinese was to test whether global discourse prosody, most common in narrative speech, could be tapped from speech output and whether systematic account of why and how tones and phrase intonations adjust could be derived from acoustic analysis of output speech. If successful, then output speech is not at all random variations, canonical forms remain

phonological, and phonetic investigations should go beyond acoustic measurements by face value.

We constructed a perception-based hierarchical framework of discourse prosody the HPG (Hierarchy of Prosodic Phrase Group) featuring the organization of global prosody from a top-down perspective and focusing on multi-phrase speech paragraph as a discourse prosody unit (Tseng et al. 2004, 2005a, Tseng 2008). By default the HPG framework assumes that individual syllable tones are constrained by the initiation, continuation and termination of higher level units the prosodic words and phrase intonation they are embedded in, at the same time individual phrase intonations are constrained by the initiation, continuation and termination from higher-level paragraph in which they are embedded as well. As a result, adjustments are necessary interactions from multiple layers of higher level information; both syllable tones and phrase intonations could not keep their canonical form in continuous speech.

The framework consists of 5 levels of perceived boundary breaks B1 through B5 while prosodic units are defined by corresponding chunks located inside each level of manually-labeled perceived boundary breaks. A schematic representation of the hierarchy is shown in Figure 1(a).

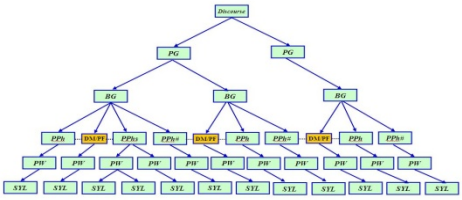


Figure 1(a) a schematic representation of the HPG (Hierarchy of Prosodic Phrase Group) and its components the Syllable (SYL), Prosodic Word (PW), Prosodic Phrase (PPh), Breath Group (BG) and Prosodic Phrase Group (PG).

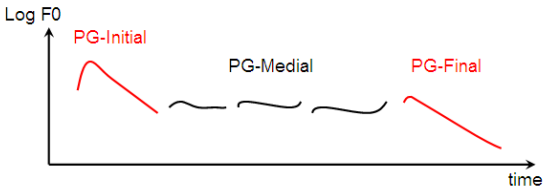


Figure 1(b) global trajectories of perceived pitch contours of a 5-PPh PG

The layered HPG prosodic units from the lowest level are the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath group (BG) and the multiple phrase group (PG). BG refers to a physio-linguistic unit BG correlating to an change of breath during speaking (Lieberman 1967, Tseng 2002) while PG corresponds to a speech paragraph. Corresponding to the HPG units but not shown in Figure 1(a) are the 5 perceived discourse boundary breaks B1/SYL, B2/PG, B3/PPh, B4/BG and B5/PG. The relationship of these prosodic units and boundary breaks are paragraph and discourse specified, which can be expressed as SYL<PW<PPh<BG<PG and B1<B2<B3<B4<B5. The hierarchy makes clear that discourse prosody context is both single-unit neighborhood concatenation and cross-over association.

Patterns of perceived global modulation are illustrated in Figure 1(b) where PG-initial, -Medial and –Final specifies the initiation, continuation and termination of

a speech paragraph, respectively. (Tseng et al. 2005a, Tseng 2006)

The graphic illustration in Figure 1(b) suggests that listeners perceive relative but different pitch heights and slopes in relation to an ultimate unit beyond individual tones and intonation. A speech paragraph is therefore one unit which associates different degrees of F0 reset, tilting, flattening and declination to new/given topical information, continuation and terminating echo rather than taking one IU/PU (Chafe 1987, Du Bois et al. 1993, Liu and Tseng 2009) at a time and refrained by linear relationship only. The global picture also implies that the planning and processing of discourse units exceeds well above simple or complex syntactic sentences while the changes made by individual sub-units are all for a reason.

Quantitative account of layered contribution and constitution by acoustic correlates the fundamental frequency (F0), duration and amplitude using a step-wise linear regression technique was adopted for the HPG framework (Tseng et al. 2004; 2005a, Tseng 2008). Contributions obtained from each layer are added up to derive the ultimate prediction accuracy, at the same time accounted for both local and higher level information. In short, there is discourse structure (DS) in the speech signal.

II. Why Information Weighting in Relation to Discourse Prosody?

Based on the results of DS, we further hypothesize that the planning of

spontaneous SpnLture speech (SpnL) is more complex than read speech (RS) since it involves more elaborate planning of information structure (IS) in addition to discourse structure (DS). While RS is more passive planning refSpnLting largely the speaker's expression of discourse coherence (DC); SpnL requires more explicit 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 (Kruijff-Korbayava & Steedman 2003). 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 assume that prosodic highlights in speech can be syntax and discourse governed as well as speaker intended, thus refSpnLting more complex interaction of phrase-level (syntactic) and higher-level (discourse) planning of IS as well as speaking intension at the same time. The following analyses are thus two-fold, one analysis aims to compare the similarity and diversity of information distribution between RS and SpnL; another analysis to

derive acoustic and prosodic patterns that are directly related to allocation of information in the speech signal.

III. Speech Materials and Preprocessing

Speech data of read narratives (RS) and spontaneous classroom SpnLtures (SpnL) are used to examine discourse prosody instead of short phrases produced in isolation. Read speech includes two types of Mandarin speech recorded in sound proof chambers: (1) plain text of 26 discourse pieces from Sinica COSPRO (Tseng et al. 2005b) (approximately 6700 syllables, produced by 1 male and 1 female radio announcers), coded as CNA, (2) simulation of weather broadcast (WB) (approximately 7,000 syllables, produced by 1 male and 1 female untrained speakers). All of the text was designed to illustrate discourse speech prosody. Spontaneous speech is university classroom SpnLtures (approximately 90 min or 41,000 syllables, produced by one Mandarin male speaker), coded as SpnL.

III.1. Tagging discourse units

The rationale of the tagging is to make paragraph and discourse specifications inherent (Tseng et al. 2005b) and subsequent examination and extraction of discourse prosody possible. The tagging notations followed the ToBI convention (Silverman et al. 1992) to divide speech strings into various sizes of prosodic units by boundary

breaks rather than identifying single prosodic units only. The prosodic units, however, are HPG specified and are therefore different from singularly defined prosodic units PU or ToBI defined prosodic units that are bound to syntactic units. Five levels of perceived boundary breaks, B1 to B5 are manually tagged of across the flow of fluent speech and checked for both intra- and inter-transcriber consistency.

The annotation is designed in accordance with the underlying principle of the framework rather than text based, and purposely removes itself from possible connotations from other levels of linguistic information. Segmental identities are automatically labeled, followed by manual spot checking of alignments. Trained transcribers then listen to the speech data from headsets and manually tag 5 levels of perceived boundary breaks using the Sinica COSPRO Toolkit (Tseng et al. 2005b). Cross-transcriber consistency is checked, and only consistently transcribed data are used for analysis.

III.2. Tagging perceived emphasis

Emphasis is defined by perceived degree of accentuation and independent of discourse organization and tagged by trained transcribers. The definitions of perceived emphases are as follows (Tseng et al. 2011):

- E0-unstressed portions marked by reduced pitch, volume and/or segment contractions

- E1-normal pitch, volume with no segmental contractions
- E2-higher pitch or louder volume irrespective of speaker's tone of voice
- E3-higher pitch or louder volume marked by speaker's tone of voice

In other words, E2 relates to perceived focus due to syntactic or structural information whereas E3 relates to speaker intended focus and tone of voice.

Discourse units and perceived emphasis were tagged independently to make possible examination of any possible prosodic interaction between perceived prosodic highlights with respect to paragraph/discourse structure.

IV. Methodology

IV.1. Information weighting by emphasis category

To analyze the distribution patterns of tagged emphases, three steps of tailored quantization are developed. 1). The first step aims to obtain the relative positions of emphasis/no-emphasis portions in every PPh/BG, due to different sizes of both the PPhs and BGs in the data sets. The normalization results enable us to better examine the location as well as allocation of emphasis in various PPhs/BGs. The equation of normalization for BG is as follows.

$$NEmP = (EmP - BGS) / BGD \tag{1}$$

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.

2). The second step aims to plot the distribution of emphasis by histograms in which the probability Pro is described as

$$\Pr o(t_e) = n(t_e) / Nu(e) \tag{2}$$

where *e* and *te* represent emphasis categories and relative PPh positions given *e*, respectively. *Nu* and *n* denotes the number count of *e* and *te*, respectively. The distribution of emphasis is plotted first by $e = \text{E2} \cup \text{E3}$, then further broken down by emphasis status E2 and E3. In addition, to normalize the effects from emphases, the same weight is assigned to each of the tagged portion of E1/E2/E3, and canonical distribution ($\text{E1} \cup \text{E2} \cup \text{E3}$) was plotted. The third step aims to model possible information attributed weighting of perceived emphases whereby degrees of emphasis are defined by the three tags as shown in (4) below while the sum of information weighting by PPh/PG position is defined in (5) below.

$$Score(t_n) = \begin{cases} 1, \text{if label} = \text{E1} \\ 2, \text{if label} = \text{E2} \\ 3, \text{if label} = \text{E3} \end{cases} \tag{3}$$

$$S(t_n) = \sum_{n=1}^{N} Score(t_n) / N \tag{4}$$

in which *S* and *tn* represent weighting sum and position index given n-th phrase respectively.

3). To observe further interaction between discourse positions and the perceived

emphases, the PPhs from the speech data were quantized by nine relative PPh positions and further classified into three correlating HPG paragraph position PG-initial, -medial and –final.

$$\text{PG position} = \begin{cases} \text{PG-initial when sequence index=1} \\ \text{PG-final when sequence index=M} \\ \text{PG-Medial otherwise} \end{cases}$$
$$\text{M=Number of PPh in PG} \tag{5}$$

IV.2. 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 (Tseng et al. 2005a), 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.

$$T = f(EMN, PPhLen, BGP, BGL) + res \tag{6}$$

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.

IV.3. Tempo of emphasis regarding rate of phrase

In addition to tempo feature of emphasis defined in relation to discourse information (Sec. IV.2.), 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 (Tseng & Su 2008) and provide clearer picture of the speech data.

$$RTPEM = TPEM - TPPPh \tag{7}$$

where TPEM and RTPEM denote the original tempo (derived in IV.2.) and relative tempo for emphasis, respectively. TPPPH is the original tempo feature of PPh in which the emphasis is embedded in.

V. Perceived Highlights and Information Structure in Spoken Discourse??

V.1. Distribution of perceived emphasis by genre

V.1.1. *Results*

In order to see whether the perceived emphasis is genre related, the distributions

of emphasis by genres are compared and plotted in Figure2. The left panel shows distribution patterns when E2 and E3 are collapsed into one category (E2∪E3); speech genre appears to have no correlation with the distribution of emphasis. However, by further breaking down the emphases by degrees E2 and E3, the E3/E2 ratio shows that SpnL (0.24) is distinctly different from CNA (0.05) and WB (0.02), and marked by more tone-of-voice type of emphases E3.

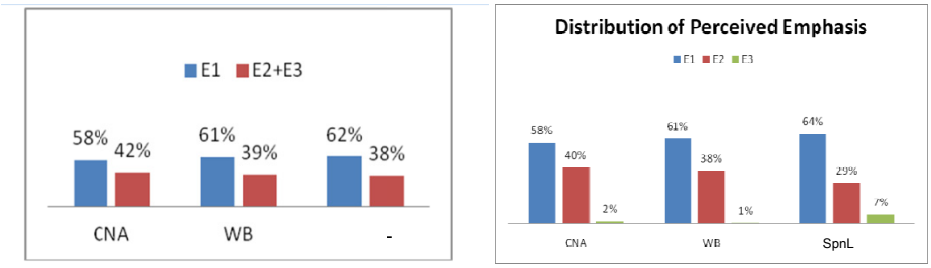


Figure 2 Distribution of Perceived Emphasis by speech genres CNA, WB and SpnL. The left panel show the emphasis/no emphasis distribution; the right panel shows distribution of E1, E2 and E3.

V.1.2. *Discussion*

The above results suggest that prosodic highlighting is genre related. SpnL (spontaneous classroom lecture speech) is distinctly different from RS (read speech), most notably marked by more occurrence of speaker intended emphasis (i.e. E3), and is clearly more expressive and communicative.

V.2. Distribution of perceived emphasis by discourse structure

V.2.1. *Results*

A. *Emphasis distribution by prosodic boundary*

Figure 3 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.

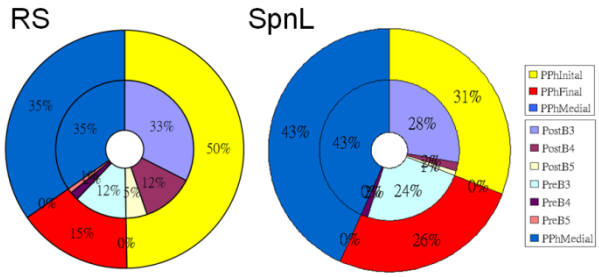


Figure 3 The distribution of emphasis by discourse associative position, boundary location and speech data type. The outer circle shows the distribution of emphasis by associative positions PPh-Initial,-Medial and Final. The inner circle shows the distribution of emphasis before and after PPh-local boundary B3 and higher-level boundaries B4 and B5, respectively.

B. *Emphasis (perceived keyword) distribution in BGs*

Patterns of emphasis distribution between RS and SpnL are derived and shown in Figure 4. 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.

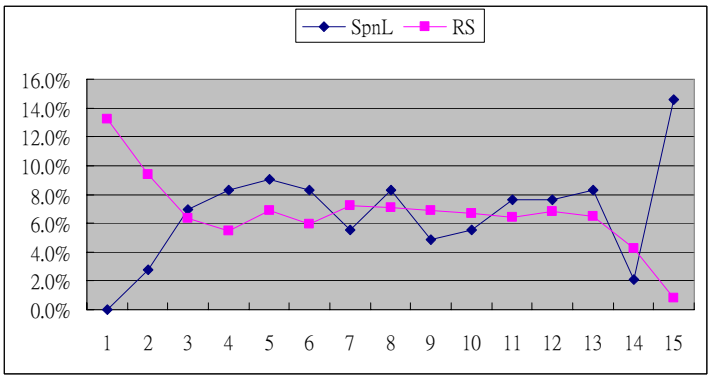


Figure 4 The distribution of emphasis (perceived keywords) by BG-position and speech data type. The horizontal axis represents relative position in BGs. The vertical axis represents the percentage of number of emphasis in current position.

C. 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 5. 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 SpnLture speech has higher probability to emphasize at the beginning of PPhs.

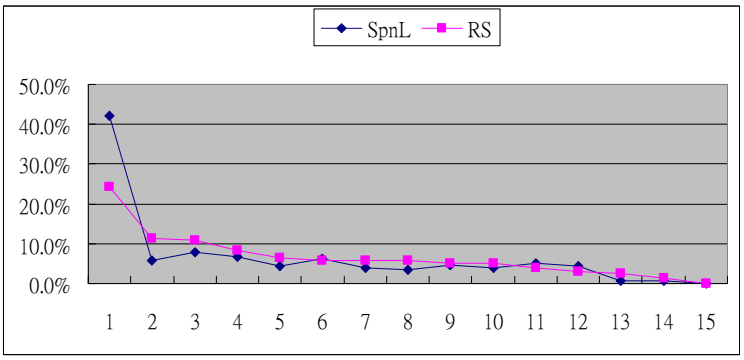


Figure 5 The distribution of emphasis (perceived keywords) by PPh-position and speech data type. The horizontal axis represents relative position in PPhs. The vertical axis represents the percentage containing emphasis in current position.

V.2.2. *Discussion*

The patterns of emphasis in RS are similar at phrase layer but distinct at higher BG layer. At the phrase layer, 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. At higher BG layer, 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 and marks the paragraph end. These results are interpreted as indication of key information distribution.

V.3. Distribution of perceived emphasis by genres and paragraph positions

V.3.1. *Results*

In order to compare the distribution of emphasis with respect to paragraph positions PG-initial, -Medial and –Final, the same kind of comparison was plotted in Figure 6. The left panel shows the difference between canonical distributions (E1∪E2∪E3), namely, the distribution of all emphasis/no-emphasis portions while the right panel shows the distribution of emphases by degree (E2∪E3). As found in Sec. V.1., the canonical distributions are similar across genres and PG positions. The effect

of PG-positions is similar across all three genres. Emphasis distribution is phrase initial>final> medial>; the number of phrase initial and final emphases are almost the same. In other words, discourse effect is almost identical. However, the distribution of E2∪E3 by PG positions is distinctly different. CNA is marked by phrase initial emphasis; WB marked by phrase final emphasis while SpnL marked by phrase initial and phrase final emphases. Discourse effect is different.

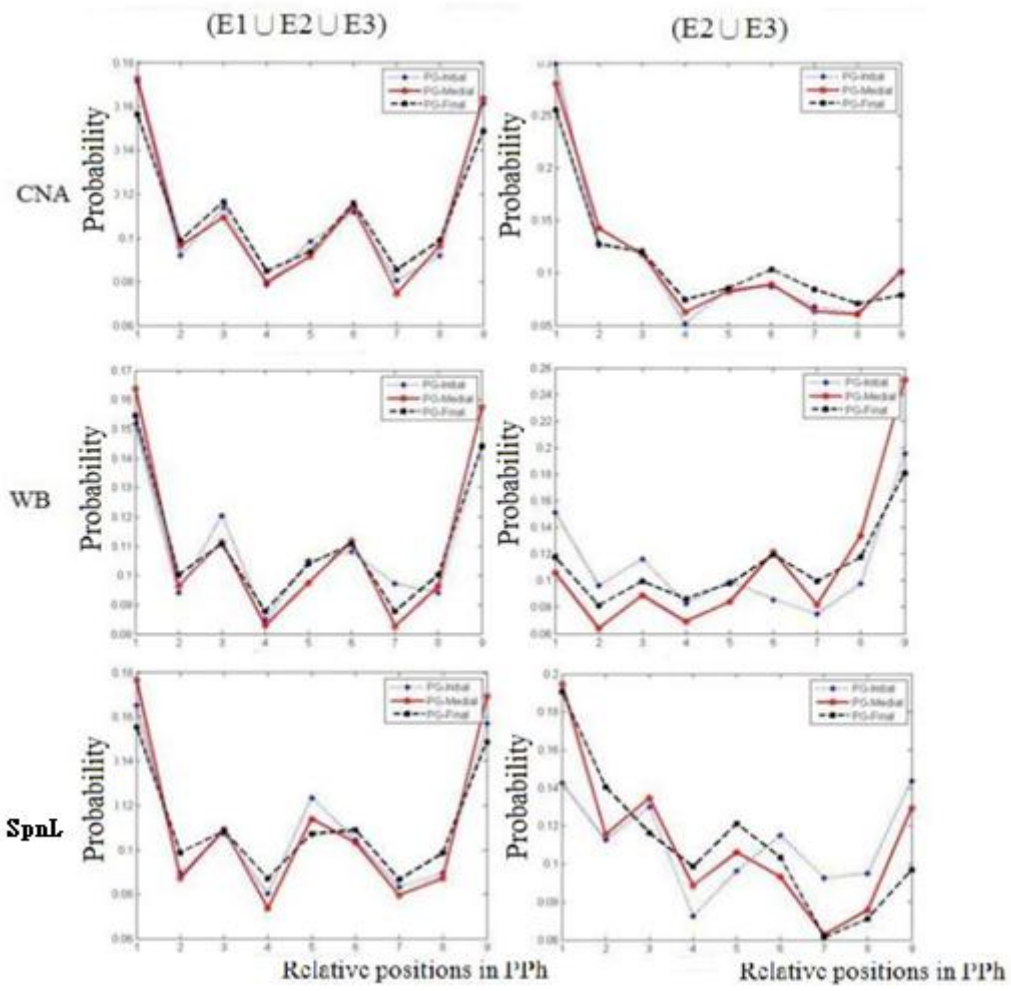


Figure 6 Distribution of (E1∪E2∪E3) and (E2∪E3) by genres CNA, WB, SpnL and discourse structure PG-initial, -medial and -final

V.4. Information weighting by emphasis category and discourse structure

V.4.1. *Results*

The following analysis aims to further model whether the weighting of information is related to genre, discourse structure and position inside a PPh; information weighting is defined by tag and sum as described in (3) and (4) of step 2 (Sec. IV.1.). The results are plotted in Figure 7: the left panel shows information weighting by discourse positions; the right panel inside PPh, respectively. The results of information weighting by discourse positions show the SpnL model predicts more emphasis in PG-initial than PG-Medial/PG-Final positions while the predictions of the CNA and WB models are similar where no discourse effect is found. The prediction by PPh yielded different results: In CNA and SpnL, the number of emphasis decreases by PPh position while the reverse is found for WB. In other words, CNA and SpnL are marked by PPh initial emphasis while WB by PPh final emphasis. More difference is found for unit PPh than for discourse positions.

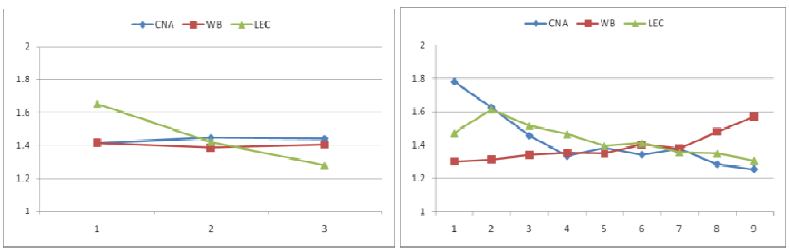


Figure 7 *Information models by perceived emphasis for PG by discourse positions and by PPh.*

V.3. & V.4. *Discussion*

Similar to results found in V.1., the distribution of perceived emphasis also varies by genre: reading prose is the most passive mode, marked by phrase initial emphasis which would be default, while simulating weather forecast is marked by phrase final emphasis and SpnL marked by both phrase initial and phrase final emphases (Sec.V.3.). The phrase initial ones would be default as RS while the phrase final ones would be speaker intended. More difference is found for unit PPh than for discourse positions (Sec.V.4.).

V.5. Acoustic characteristics of perceived emphasis (highlight)

V.5.1. *Results*

A. *Contrastive analysis of perceived emphasis and by acoustic correlates and speech genre*

Having found genre specific attributes by emphasis distribution, we are interested to know whether systematic and genre-specific acoustic patterns could be obtained from the speech data. This is particularly significant because the emphases are identified perceptually. Figure 8 shows the contrastive patterns between sections of identified emphasis in the speech signal and sections without emphasis by duration, average F0, F0 range and intensity and by speech genres. The results show that the acoustic contrasts are most pronounced for SpnL. Results of two-way ANOVA shows

significant differences are found for all four acoustic features in SpnL ($p<0.0001$), while the most discriminative features are average F0 and intensity (F-ratio=846, 873). Similar but less pronounced patterns of average F0 and intensity are found in CNA (F-ratio=492, 364). However, in WB, the two most discriminative features are intensity and duration (F-ratio=196, 170) instead.

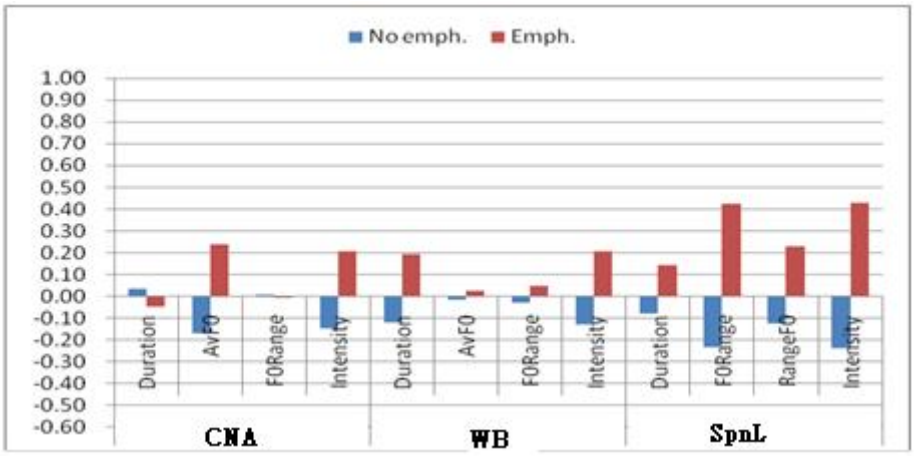


Figure 8 Mean values of acoustic correlates by emphasis/no emphasis and genres

B. *Contrastive analysis of perceived emphasis by PPh-positions*

To further find out how distinct patterns of information weighting by unit PPh instead of by discourse positions (Sec. V.4, Figure 7) also apply to acoustic patterns, the same analysis of Sec. V.5.1.A. is performed by PPh positions. The results are plotted in Figure 9. Results of two-way ANOVA shows that in SpnL the same significant difference is found for in all four features duration, average F0, F0 range and intensity and across all PPh positions ($p<0.001$). For CNA, significant difference across PPh position is found in average F0 and intensity ($p<0.001$) only. For WB,

significant difference is found in intensity by all PPh positions and duration in PPh-Final position only ($p<0.0001$). Moreover, the two most discriminative features in SpnL are the average F0 and intensity in PPh-Final positions (F-ratio=410, 287).

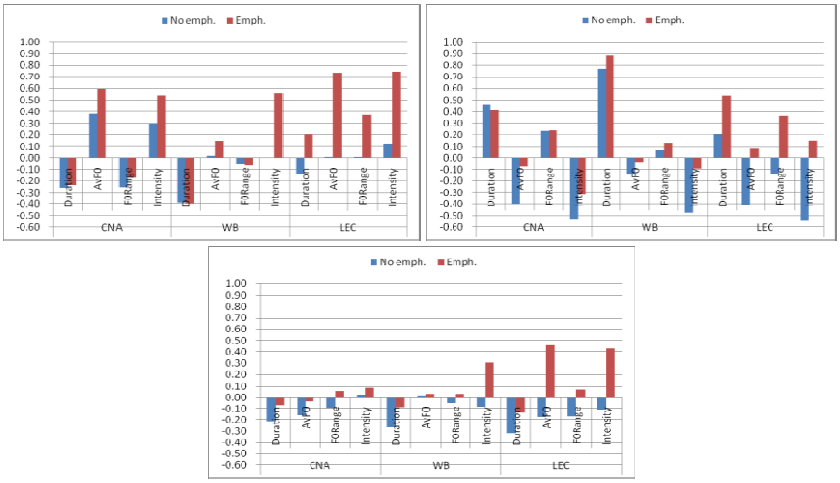


Figure 9 Mean values of acoustic features by emphasis/non-emphasis, genre and PPh position. The panels from the top denote the PPh-Initial, PPh-Medial and PPh-Final positions, respectively.

V.5.2. Discussion

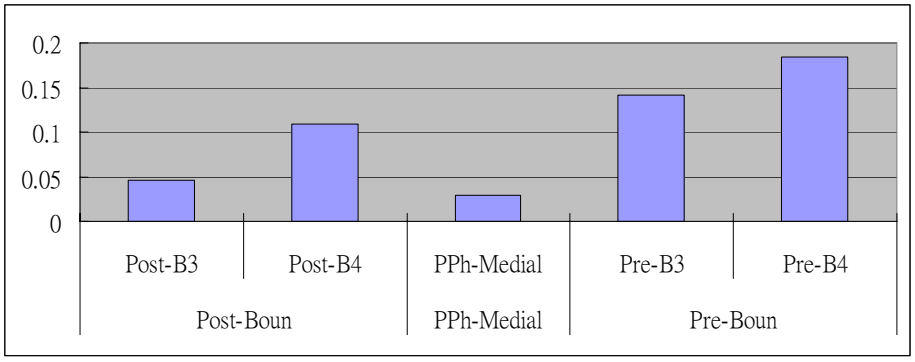
Correlative analyses between the perceptually identified emphases and their acoustic characteristics showed that SpnL is different from prose reading CNA and simulated weather broadcast WB in every acoustic feature examined. Significant differences of acoustic contrasts are found for all four acoustic features, namely, duration, average F0, F0 range and intensity. Nevertheless, emphases in passive reading (CNA) are realized by contrasts in average F0 and intensity while the style of WB is realized through contrasts in duration and intensity. These results suggest that prosodic highlighting can be realized in different combinations of acoustic features to signal different degree of contrastive strength.

V.6. Tempo patterns of perceived emphasis by location and boundary type

V.6.1. *Results*

Since boundary properties, commonly referred to as lengthening, are discourse constrained (Tseng & Su 2008); 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 10). 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.

SpnL



RS

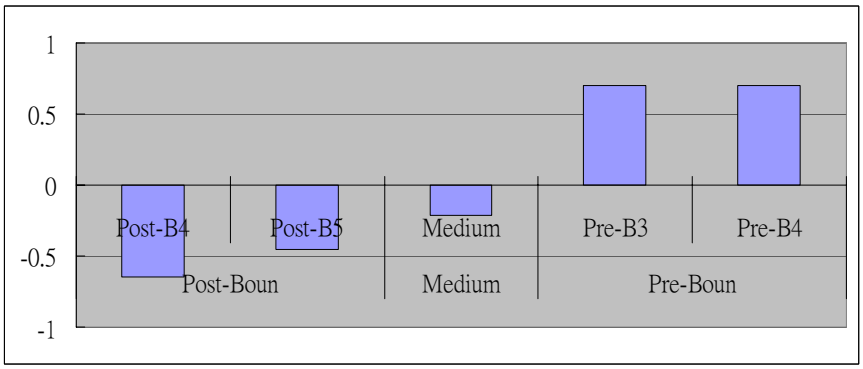


Figure 10 The relative tempo of emphasis by position in PPh and discourse boundary type. The upper and lower panels denote tempo relative tempo of emphasis for SpnL and RS respectively. The horizontal axis represents position PPh and boundary type. The vertical axis represents the relative tempo of emphasis and zero means the tempo of emphasis is equal to current PPh tempo.

V.6.2. Discussion

RS tempo modulations are the same with or without account of emphatic portions, beginning at the fastest rate at the onset and gradually slowing down to the end. Emphasis in RS is not marked by tempo modulation, either by the emphatic portions itself or by overall tempo to change. On the other hand, emphasis in SpnL speech exhibited distinctly different patterns. Instead of the default position of paragraph prominence, it occurs from the mid-paragraph across the board, with the highest distribution at the paragraph end, occurring only at the unexpected positions to make the implicit explicit. And when emphasis occurred, they assumed more phrasal prominence (over 40%) than their counterpart in RS.

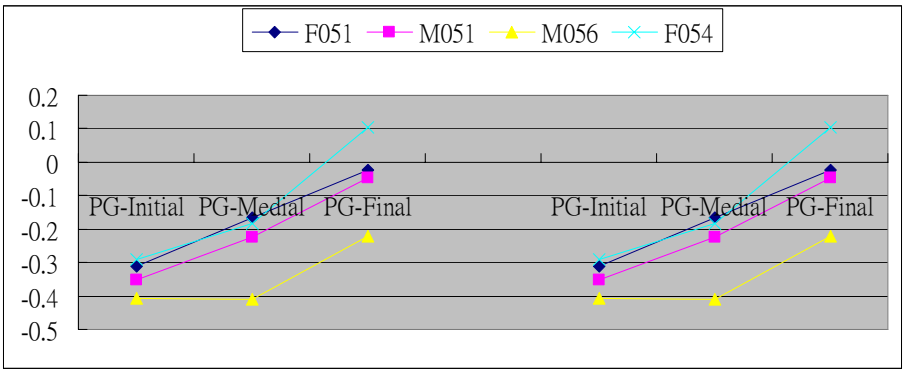
V.7. Tempo patterns of the paragraph sub-unit BG

V.7.1. *Results*

Figure 11 shows tempo modulation within and between the paragraph sub-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.

RS



SpnL

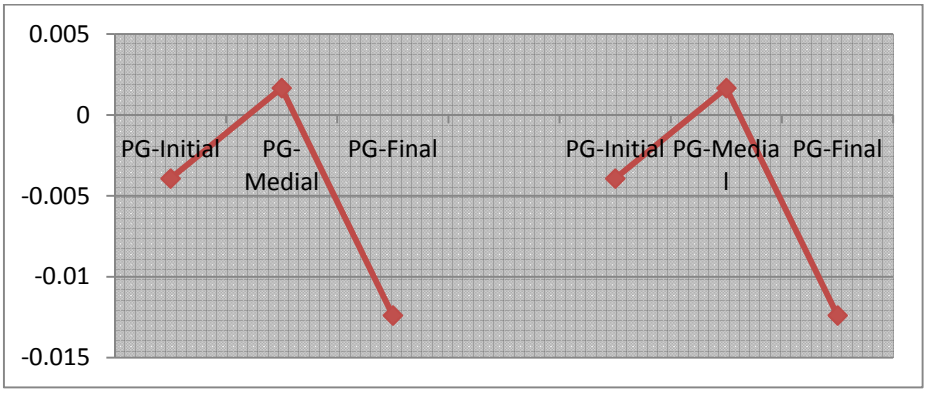


Figure 11 Tempo allocation patterns by BG-position and speech data type. The upper and lower panels denote tempo allocations of BGs for RS and SpnL respectively. The horizontal axis represents the relative position in BGs. The vertical axis represents the normalized mean value of PPh tempo by BG-position.

We further examined tempo modulation of shorter paragraphs in SpnL for more

detailed information and as a reference of BG length. Figure 12 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.

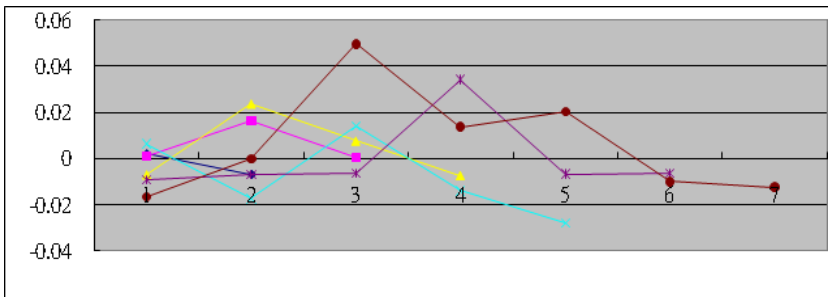


Figure 12 Tempo allocation patterns of BG consisting of 7 or less phrases by BG-position in SpnL. The horizontal axis represents the relative position in BGs. The vertical axis represents the normalized mean value of PPh tempo by BG-position.

V.7.2. Discussion

The highest occurrence of highlighting the paragraph end in SpnL implies the speaker's intention to reiterate the most important information, even overriding boundary lengthening regardless of phrase or paragraph ending. Much more complicated tempo modulations of emphasis are also found, both by the emphasis itself and globally by its embedding the phrase. An emphasis always assumes the slowest rate, and the entire embedding phrase also slows down. This suggests the speaker's intended loading and weighting of information, reflecting active planning of IS (Lambrecht 1994) on top of DS, where systematic prosodic manipulations to signal the implicit/explicit, given vs. new, theme vs. rheme dichotomies can be traced in the speech signal.

V.8. Distribution of F0 and intensity of BG in SpnL

V.8.1. *Results*

In relation to results of tempo analysis of SpnL (see Sec.V.7. and lower panel of Figure 11) in which the slowest rate implies a dividing point corresponding to two high occurrences of perceived emphasis distribution, the distribution of mean F0 and intensity are also analyzed in search of corresponding dividing points (Figure 13). 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.

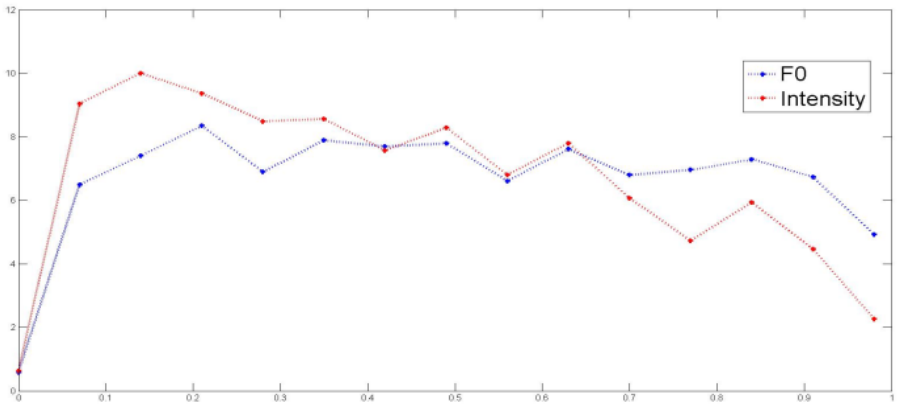


Figure 13 Distributions of mean F0 and intensity by BG positions. The horizontal axis represents relative position in BGs. The vertical axis represents the percentage of values larger than mean in current position.

V.8.2. Discussion

Furthermore, information structuring in the prosodic domain appears to have

most to do with tempo modulations, and much less with F0 and intensity. Since our data are Mandarin only, this could be language dependent and merits further investigations of other language(s).

V.9. Could emphasis be analyzed as additional information layered over discourse structure?

V.9.1. *Results*

In order to test whether emphases could be analyzed as an extra layer of prosody specifications over the canonical prosody patterns by discourse positions (Tseng et al. 2006), perceived prosodic highlights are normalized from the acoustic signal and compared with sections of speech signal where no prosodic highlights are identified. Figure 14 shows the discourse patterns by acoustic features in which perceived emphases are removed. In addition, speech sections without emphases representing the canonical discourse pattern are also plotted for comparison. The results show an almost complete overlap between the canonical discourse patterns and the emphases removed patterns, suggesting that the discourse pattern can be seen as underlying structure (or base form) while prosodic highlighting layering over.

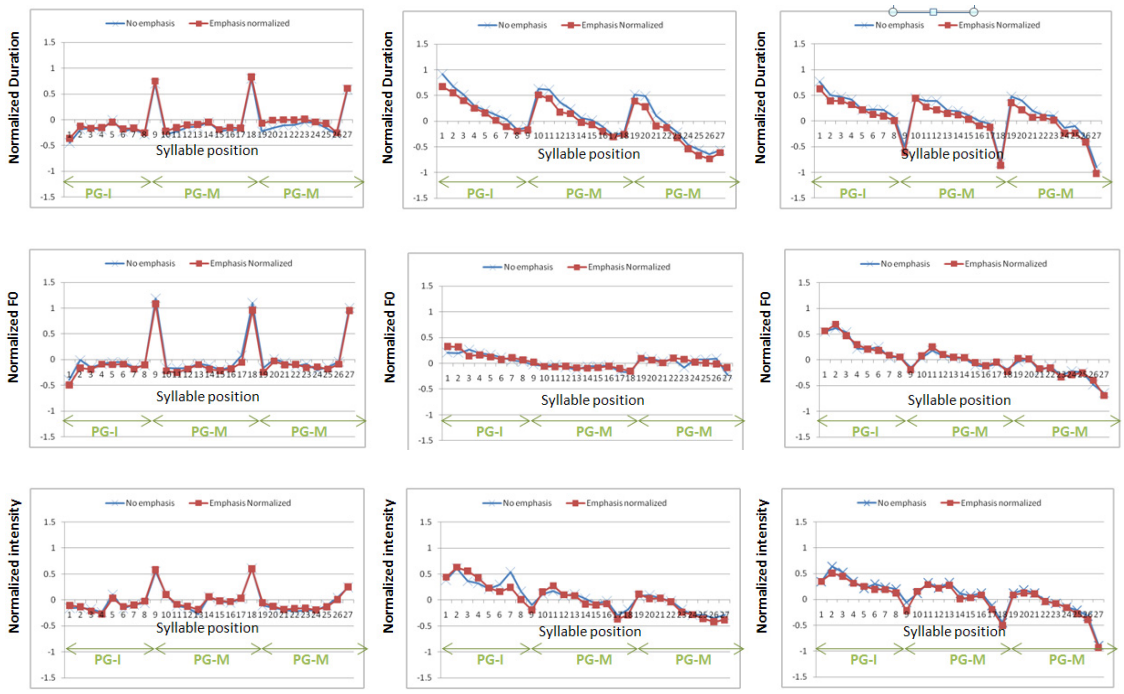


Figure 14 Perceived emphases are normalized and compared with units without emphases in relation to discourse structure. Acoustic patterns by discourse associative positions PG-Initial, -medial and –final and speech genres prose reading CNA, simulating reading of weather broadcast WB and spontaneous university classroom SpnLture SpnL are derived and plotted. Unit of analysis is PPh.

V.9.2. Discussion

A simple procedure that normalized the identified prosodic highlights from the speech signal revealed an underlying pattern that is almost identical to canonical discourse prosody patterns. The results imply that the surface prosodic twists and turns caused by different locations and needs of emphatic expressions in no way interferes with the underlying discourse structure which is obligatory to deliver core linguistic content.

VI. General Discussion

The above analyses can be summarized into two categories, namely distribution

analyses (Sec. V.1. to V.4.) and acoustic patterns (Sec. V.5. to V.8.). Results from distribution analysis showed that prosodic highlights, perceived as emphasis and focus, are genre conditioned. In RS most of the placements of prosodic highlights are at the default phrase-initial positions. However, in SpnL 75% of the prosodic highlights occur at the phrase-medial and –final positions. The distinction shows that production planning of RS requires less planning from the speaker while processing load of SpnL should be greater from the listener. Results of acoustic patterns showed that even for RS, prose reading is realized through contrasts in pitch and loudness while simulating weather forecast is realized through contrasts in tempo and loudness. These results lead us to further investigate tempo features in RS and SpnL in relation to boundary properties (1) pre-boundary lengthening and (2) post-boundary shortening. We noted that per-boundary lengthening effects remain while post-boundary duration reduction was only found for RS. As a result, the overall tempo of RS and SpnL is different; SpnL may be more syllabic in rhythm and more staccato sounding. More explicit planning of emphasis is found in SpnL than in RS, evidenced in overall as well as emphasis-local tempo modulation. We argue that placements of highlights and corresponding tempo modulations are directly related to active planning of where and how to express information chunks in the speech flow. On the basis of tempo evidence, we note that in our Mandarin data the weighting of

information is expressed not through pitch and loudness features but instead through tempo features. Finally, we also noted through the interaction between DS and IS (Sec. V.9) it is clear that DS remain intact regardless of IS, hence discourse coherence is the underlying structure while IS could be regarded as different degrees of added weights layered over DS. The results of cross-genre patterns also demonstrate that prosodic highlighting is indeed genre related; distribution of key information can be attributed to both linguistic content and communicative needs. Prosodic highlighting can be analyzed as an extra layer over discourse structure, the former signals key information while the latter underlying linguistic association. Therefore, the prosodic realization of output continuous speech may appear to be strewed with emphases and highly different from canonical forms. However, beneath the acoustic deviations and discrepancies, the underlying structure in fact remains intact. Future work will focus on more detailed analysis of information structure and its prosodic realization.

VII. Conclusion

The goal of this paper is to examine how discourse structure interacts with information structure. We examined in both RS and SpnL the following: (1) perceived prosodic highlights in three genres of fluent continuous Mandarin; (2) different tempo distribution of prominence, represented by perceived emphasis; (3) tempo, pitch and

loudness features at both emphasis-local and higher-level units the PPh level and BG level, respectively and (4) acoustic patterns of emphatic portions. In RS the predominant feature remains to be discourse coherence. Nevertheless, SpnL is featured by more intricate tempo modulation patterns at various levels that are caused by distribution and placement of key information, and most notably at the paragraph end as a highlight of key information. As a result, boundary properties appear distorted without unit-final lengthening. Our account above showed that how IS of the speaker override DS and delivered through the prosodic domain on the surface but how DS remain constant underneath the seemingly highly varied output surface. It is therefore clear for us that prosodic highlights should not be investigated independent of discourse structure; just as post-focus compression should not be examined independent from its embedding unit. We believe these findings have furthered our understanding of the organization of SpnL in the prosodic domain. In particular, the tempo patterns could be of use to technological development such as keyword spotting, topic change and information weighting. Future work will be on more detailed analysis of prosodic patterns with respect to IS.

VIII. Reference

Chafe, Wallace. 1987. Cognitive constraints on information flow. In R. Tomlin, ed.,

Coherence and grounding in discourse. Amsterdam: John Benjamins.

Cutler, A. & Butterfield, S. 1992. Rhythmic cues to speech segmentation: Evidence from Juncture misperception. *Journal of Memory and Language*, 31:218-236.

Du Bois, J.W., Schuetze-Coburn, S., Cumming, S. and Paolino, D. 1993. Outline of Discourse Transcription. In: Edwards, J., Lampert, M. (eds.) *Talking Data: transcription and coding in discourse research*, Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey, Hove and London. 45-90.

Gussenhoven, C. 1997. Types of focus in English? In Daniel Buring, Matthew Gordon and Chungming Lee (eds.) *Topic and Focus: Intonation and Meaning: Theoretical and Crosslinguistic Perspectives*. Dordrecht: Kluwer.

Keller, E., Zellner, B., Werner, S., and Blanchoud, N. 1993. The prediction of prosodic timing: Rules for final syllable lengthening in French. *Proceedings of ESCA Workshop on Prosody*, Lund, Sweden, 212-215.

Kruijff-Korbayava, I. and Steedman, M. 2003. Discourse and Information Structure. *Journal of Logic, Language and Information* 12:249-259.

Lambrecht, K. 1994. *Information Structure and Sentence Form—Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Liu, Y. and Tseng, S. 2009. Linguistic Patterns Detected through a Prosodic

Segmentation in Spontaneous Taiwan Mandarin Speech. *Linguistic Patterns in Spontaneous Speech*, Institute of Linguistics, Academia Sinica. 147-166.

Lieberman, Philip. 1967. *Intonation, perception, and language. Cambridge.* M.I.T. Press.

Selkirk, E. 2000. The interaction of constraints on prosodic phrasing. In Merle Horne (ed.) *Prosody: Theory and Experiment*. Dordrecht: Kluwer, 231-262.

Shattuck-Hufnagel, S., Turk, A. 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguist Research*, 25(2): 193.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. 1992. ToBI: A standard for labeling English prosody. *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP 92)*, (Oct. 12-16, 1992), Alberta, Canada, 867-870.

Tseng, C. 2002. The prosodic status of breaks in running speech: Examination and Evaluation. *Proceedings of the 1st International Conference on Speech Prosody 2002*, (Apr. 11-13, 2002), Aix-en-Provence, France, 667-670.

Tseng, C. 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.

Tseng, C. 2008. Corpus Phonetic Investigations of Discourse Prosody and Higher

Level Information (in Chinese). *Language and Linguistics,* 9(3): 659-719.

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.

Tseng, C., Pin, S. and Lee, Y. 2004. Speech prosody: issues, approaches and implications. In Fant, G., H. Fujisaki, J. Cao and Y. Xu Eds. *From Traditional Phonology to Mandarin Speech Processing, Foreign Language Teaching and Research Process,* 417-438.

Tseng, C., Pin, S., Lee, Y., Wang, H. and Chen, C. 2005a. 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.

Tseng, C., Cheng, Y. and Chang, C. 2005b. Sinica COSPRO and Toolkit—Corpora and platform of Mandarin Chinese fluent speech. *Proceedings of Oriental COCOSDA 2005*, (Dec. 6-8, 2005), Jakarata, Indonesia, 23-28.

Tseng, C., Su, Z., Chang, C. and Tai, C. 2006. Prosodic filers and discourse markers—Discourse prosody and text prediction. *TAL 2006 (The Second International Symposium on Tonal Aspects of Languages)*, (April 27-29, 2006), La Rochelle, France.

Tseng, C., Su, C. and Huang, C. 2011. Prosodic Highlights in Mandarin Continuous Speech—Cross-Genre Attributes and Implications. *Interspeech 2011(The 12th Annual Conference of the International Speech Communication Association),* (Aug 27-31, 2011), Florence, Italy. 4 pages.

Zellner, B. 1994. Pauses and the temporal structure of speech. In E. Keller (ed.) *Fundamentals of Speech Synthesis and Speech Recognition*, Chichester: John Wiley, 41-62.