



Prosodic highlights in Mandarin continuous speech—Cross-genre attributes and implications

Chiu-yu Tseng¹, Chao-yu Su² and Chi-Feng Huang¹

¹ Phonetics Lab, Institute of Linguistics, Academia Sinica Taipei, Taiwan

² Taiwan International Graduate Program (TIGP), Academia Sinica Taipei, Taiwan

cytling@sinica.edu.tw

Abstract

The present study examines perceived prosodic highlights in three genres of fluent continuous Mandarin to test (1) whether prosodic highlights are genre related, (2) how they interact with discourse structure, (3) how they signal information status, (4) whether systematic acoustic patterns could be obtained from speech data analysis, and (5) whether prosodic highlights is layered over to discourse structure. Results demonstrate that prosodic highlighting is genre related; distribution of key information can be attributed to linguistic content and communicative needs. Prosodic highlighting is an extra layer over discourse structure, the former signals key information while the latter underlying linguistic association.

Index Terms: prosodic highlights, perceived emphasis, discourse structure, information weighing, acoustic features

1. Introduction

The present study examines perceived prosodic highlights in three genres of fluent continuous Mandarin to test (1) whether prosodic highlights are genre related, (2) how they interact with discourse structure, (3) how they signal information status, (4) whether systematic acoustic patterns could be obtained from speech data analysis, and (5) whether prosodic highlights is layered over discourse structure. Perceived prosodic highlights are defined as auditory perceptible prominent words across output continuous speech signaling emphasis, focus, contrastive stress or accentuation. Thus the identified highlights could be caused by linguistic factors such as syntactic, semantic and functional specifications as well as speaker intended stress, emphasis and focus as long as they are perceptually prominent. By discourse structure, our focus is on the prosodic formation and chunking/phrasing association of multi-phrase topic structure [1] most clearly evidenced in speech paragraphs. In the following sections, we will analyze and discuss manually tagged perceived prosodic highlights (emphases) of three speech genres, namely, reading prose, simulating weather forecast and spontaneous university classroom lectures in accordance with the questions raised.

The paper is organized as follows: Sec. 2 describes speech materials used and annotation rationale. Sec. 3 describes methodology. Sec. 4 presents distribution of perceived highlights by genre, discourse structure and information weighting. Sec. 5 presents acoustic characteristics of perceived highlights. Sec. 6 and 7 are discussion and conclusions.

2. Speech materials and annotation rationale

2.1. Speech Materials

The three genres of Mandarin speech data consists of two

types of read speech and one spontaneous speech. The read speech is recorded in sound proof chambers and consists of (1) reading of plain text of 26 discourse pieces from Sinica COSPRO [2] (45 min/7,000 syllables/85MB produced by 1 male and 1 female radio announcers), coded as CNA, (2) simulating weather broadcast (WB) (approximately 45 min/6,700 syllables/50MB, produced by 1 male and 1 female untrained speakers). All of the text was designed to illustrate discourse speech prosody. Spontaneous speech is microphone speech of university classroom lectures (LEC) (approximately 26 min/7200 syllables/49 MB produced by one L1 Mandarin male speaker).

2.2. Annotation and rationale

The selected speech data were manually tagged by trained transcribers. 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.

2.1.1. Tagging discourse units

For discourse structure, it is essential to examine multi-phrase speech paragraphs as discourse units in addition to individual words, phrases and sentences. For this reason a perception based hierarchical discourse prosody framework the HPG (Hierarchy of Prosodic Phrase Group) [3] is adopted. The framework identifies and requires manual tagging of perceived multi-phrase speech paragraphs; and specifies the associative prosodic patterns of the phrases inside each paragraph [4, 5]. Five levels of perceived boundary breaks B1 through B5 across the flow of fluent speech following the ToBI notations are used to divide speech strings while prosodic units are defined by corresponding chunks located inside each level of boundary breaks. The 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) which corresponds to a speech paragraph. A physio-linguistic unit BG correlating to an audible and complete change of breath is included [6, 7] to accommodate change of breath during the production of continuous speech. Corresponding 5 discourse boundary breaks B1/SYL, B2/PG, B3/PPh, B4/BG and B5/PG. In turn, the relationship of the prosodic units and boundary breaks can be expressed as SYL<PW<PPh<BG<PG and B1<B2<B3<B4<B5 whereas paragraph and discourse specifications are inherent [2]. We note that by default the top-down perspective also specifies how discourse prosody context is both single-unit neighborhood concatenation as well as cross-unit association.

2.1.2. Tagging perceived emphasis

Perceived emphasis is defined as follows and tagged by trained transcribers independent of discourse tagging:

- 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.

3. Methodology

To analyze the distribution patterns of tagged emphases, three steps of tailored quantization are developed. The first step aims to obtain the relative positions of emphasis/no-emphasis portions in every PPh; quantization is adopted by nine relative positions. The second step aims to plot the distribution of emphasis by histograms (Figure 1, Sec. 4.2.) in which the probability Pro is described as

$$Pro(t_e) = n(t_e) / Nu(e) \quad (1)$$

where e and t_e represent emphasis categories and relative PPh positions given e , respectively. Nu and n denotes the number count of e and t_e , respectively. The distribution of emphasis is plotted first by $e = E2 \cup E3$, 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 ($E1 \cup E2 \cup E3$) 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 (3) below while the sum of information weighting by PPh/PG position is defined in (4) below.

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

$$S(t_n) = \sum_{n=1}^N Score(t_n) / N \quad (3)$$

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

To observe possible interaction between discourse positions and the perceived emphases, the PPhs from the speech data were further classified into three correlating HPG paragraph position PG-initial, -medial and -final.

$$PG\text{-position} = \begin{cases} PG\text{-initial} & \text{when sequence index}=1 \\ PG\text{-final} & \text{when sequence index}=M \\ PG\text{-medial} & \text{otherwise} \end{cases} \quad (4)$$

$M = \text{Number of PPh in PG}$

4. Distribution of perceived highlights

4.1. Emphasis distribution by genres

In order to see whether the perceived emphasis is genre related, the distributions of emphasis by genres are compared and plotted in Figure1. The left panel shows distribution patterns when E2 and E3 are collapsed into one category ($E2 \cup E3$); 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 LEC (0.24) is distinctly different from CNA (0.05) and WB (0.02), and marked by more tone-of-voice type of emphases E3.

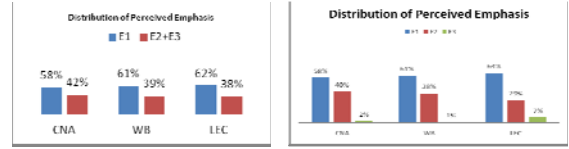


Figure 1: Distribution of Perceived Emphasis by speech genres CNA, WB and LEC. The left panel show the emphasis/no emphasis distribution; the right panel shows the distribution of E1, E2 and E3.

4.2. Emphasis distribution by discourse structure

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 2. The left panel shows the difference between canonical distributions ($E1 \cup E2 \cup E3$), namely, the distribution of all emphasis/no-emphasis portions while the right panel shows the distribution of emphases by degree ($E2 \cup E3$). As found in Sec. 4.1., 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 vs. E3 by PG positions is distinctly different. CNA is marked by phrase initial emphasis; WB marked by phrase final emphasis while LEC marked by phrase initial and phrase final emphases. Discourse effect is different.

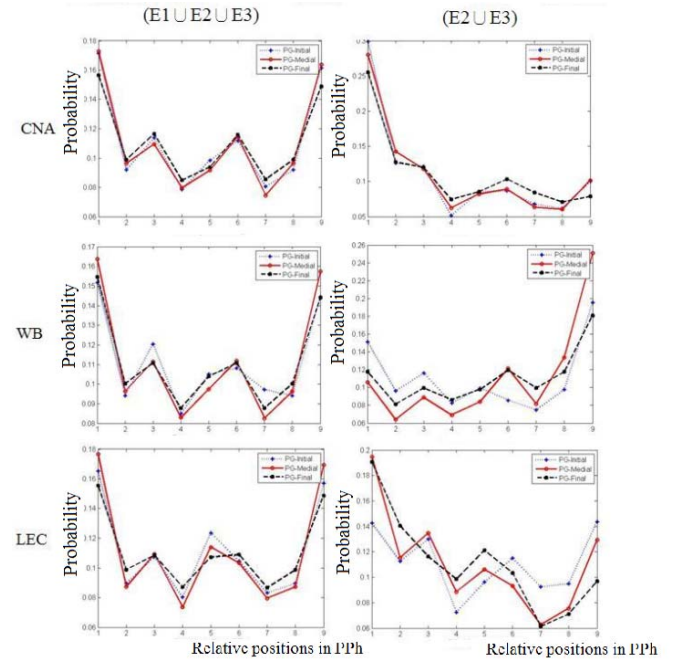


Figure 2: Distribution of ($E1 \cup E2 \cup E3$) and ($E2 \cup E3$) by genres CNA, WB, LEC and discourse structure PG-initial, -medial and -final.

4.3. Information weighting by emphasis category

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 (2) and (3) of step 3 (Sec3). The results are plotted in Figure 3: 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 LEC 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 LEC, the number of emphasis decreases by PPh position while the reverse is found for WB. In other words, CAN and LEC are marked by PPh initial emphasis while WB by PPh final emphasis. More difference is found for unit PPh than for discourse positions.

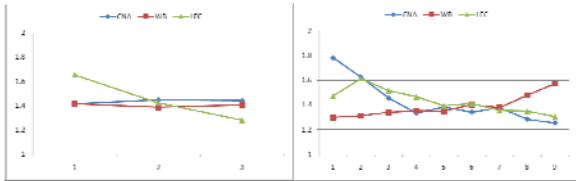


Figure 3: Information models by perceived emphasis for PG by discourse positions and by PPh. The left panel presents emphasis distribution in three PG positions, -Initial, PG-Medial and PG-Final. The right panel presents emphasis distribution by discourse sub-unit PPh.

5. Acoustic characteristics of perceived highlights

5.1. 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 4 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 LEC. Results of two-way ANOVA shows significant differences are found for all four acoustic features in LEC ($p < 0.0001$), 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.

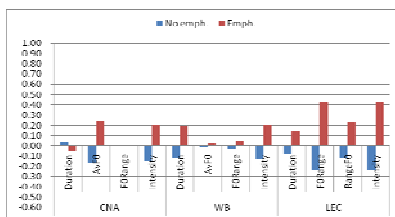


Figure 4: Mean values of acoustic correlates by emphasis/no emphasis and genres

5.2. 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. 4.3, Figure 3) also apply to acoustic patterns, the same analysis of Sec. 5.1 is performed by PPh positions. The results are plotted in Figure 5. Results of two-way ANOVA shows that in LEC 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 CAN, 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 LEC are the average F0 and intensity in PPh-Final positions (F-ratio=410, 287).

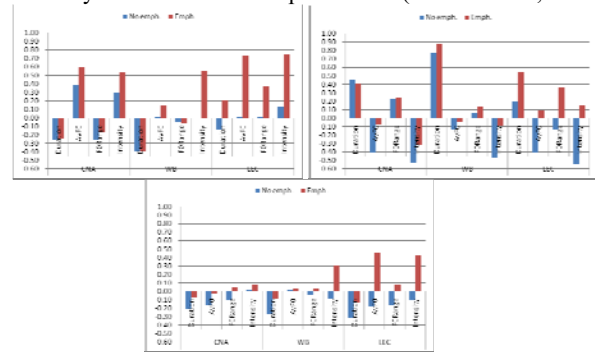


Figure 5: Mean values of acoustic features by emphasis/non-emphasis, genre and PPh position. The upper left, upper right, lower panels denote the PPh-Initial, PPh-Medial and PPh-Final positions, respectively.

5.3. Layering emphasis over discourse structure?

In order to test whether emphases could be analyzed as an extra layer of prosody specifications over the canonical prosody patterns by discourse positions [8], perceived prosodic highlights are normalized from the acoustic signal and compared with sections of speech signal where no prosodic highlights are identified. Figure 6 and Figure 7 show 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.

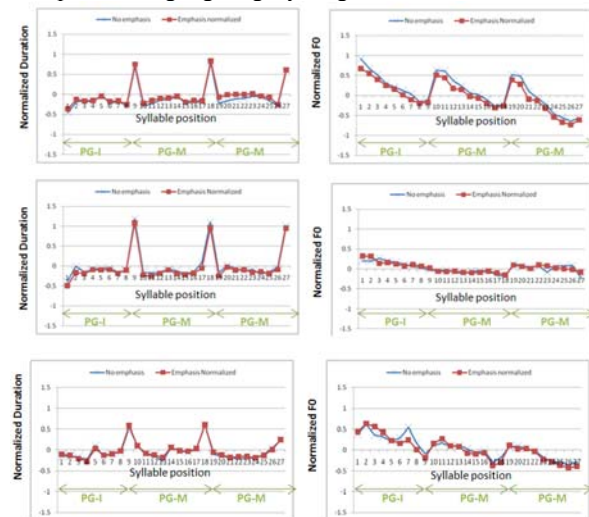


Figure 6: Perceived emphases are normalized and compared with units without emphases in relation to discourse structure. Duration and F0 patterns by discourse associative positions PG-Initial, -medial and -final and speech genres prose reading CNA, simulating reading of weather broadcast WB and spontaneous university classroom lecture LEC are derived and plotted. Unit of analysis is PPh.

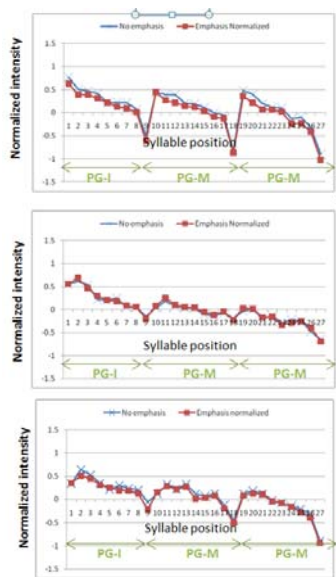


Figure 7: Perceived emphases are normalized and compared with units without emphases in relation to discourse structure. Intensity patterns by discourse positions PG-Initial, -medial and -final and speech genres prose reading CNA, simulating reading of weather broadcast WB and spontaneous university classroom lecture LEC are derived and plotted. Unit of analysis is PPh..

6. Discussion

The above results collectively suggest that prosodic highlighting is genre related. Spontaneous classroom lecture is distinctly different from read speech, most notably marked by more occurrence of speaker intended emphasis, and is clearly more expressive and communicative (Sec. 4.1). The distribution of perceived emphasis also varies by genre: reading prose is the most passive mode, marked by phrase initial emphasis, while simulating weather forecast is marked by phrase final emphasis and LEC marked by both phrase initial and phrase final emphases (Sec.4.2). More difference is found for unit PPh than for discourse positions (Sec.4.3). Correlative analyses between the perceptually identified emphases and their acoustic characteristics showed that LEC is different from CNA and 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 (CAN) are realized by contrasts in average F0 and intensity while the style of WB is realized through contrasts in duration and intensity (Sec. 5.1, 5.2.). These results suggest that prosodic highlighting can be realized in different combinations of acoustic features. Finally, 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 (Sec. 5.3). 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.

7. Conclusions

The goal of this paper is to examine perceived prosodic highlights in three genres of fluent continuous Mandarin and see how they can be explained by systematic patterns by genre,

discourse structure, information weighting acoustic manifestations. Results of cross-genre patterns 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 strewn 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.

8. References

- [1] Keen, E and Schieffelin, B. 1976. Topic as a discourse notion. in C. Li and S.Tompson Eds, *Subject and Topic*. New York: Academic Press, pp. 335-84.
- [2] Tseng, C., Cheng, Y., and Chang, C., 2005. Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech, Oriental COCODA 2005, Jakarta, Indonesia, 2005.
- [3] Tseng, C., Pin, S., 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*, pp. 417-438.
- [4] Tseng, C., Pin, S., Lee, Y., Wang, H. and Chen, 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, pp. 284-309.
- [5] Tseng, C. 2008. Corpus Phonetic Investigations of Discourse Prosody and Higher Level Information (in Chinese). *Language and Linguistics*, 9(3): 659-719.
- [6] Lieberman, Philip. 1967. *Intonation, perception, and language*. Cambridge: M.I.T. Press.
- [7] 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, pp. 667-670.
- [8] Tseng, C., Su, Zh., Chang, C. and Tai, C. 2006. Prosodic files 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.