The Convergence of Perceived Prosodic Highlight for Discourse Prosody

- A cross-speech genre analysis

Helen Kai-yun Chen, Wei-te Fang, Chiu-yu Tseng

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

Abstract

Drawing on a parallel mechanism of motivic similarity in melody perception of music study, this research explores the convergence of perceived prosodic highlight allocations from lower-level prosodic units in reflecting higher-level discourse prosody. Based on the assumption that discourse prosody in cross-genre Mandarin speech corresponds to a **coarsely graded** distinction of prosodic contrastiveness that can be realized in limited numbers of variations (variation as in the sense of music study), this study attempts to test the feasibility of such convergence. It is demonstrated that numbers of prosodic variations at discourse-levels can be successfully narrowed down after merging. While the convergence for higher-level discourse prosody is achievable, the study further unveils the source for the divergent realizations of prosodic variations. The cross speech-genre analyses show that the divergence is directly associated with the chunking size of discourse-prosodic units in that the larger the planning size is, the more divergence exhibited in the numbers of non-mergeable prosodic patterns. The study thus offers an alternative insight to the commonly shared view toward the limited number of intonation variations found at higher levels of prosody discourse.

Index Terms: discourse prosody, convergence of prosodic highlights, perceived prominence and prosodic contrastiveness, cross speech genre comparison

1. Introduction

The objective of the study is to examine the convergence of perceived prosodic highlight allocations for discourse prosody in continuous Mandarin speech from 4 diverse styles. Our basic assumption is that the allocation of perceived prosodic highlight in continuous speech can be converged or merged into limited patterns of variations (in the sense of music study, i.e. [1]) in reflecting discourse-based prosodic realizations. It is held that such convergence reflects a 'coarsely graded' distinction of prosodic contrastiveness from the prosody processing at higher discourse levels. In other words, we adopt a viewpoint similar to the melody perception in music study (i.e. [1]) in that the motivic similarity can be identified between different parts of the overall music melody. As explained, listeners recognize the motivic similarity in music without identity, as similarity is a *graded* feature influenced by many factors within music perception [1].

While drawing on a parallel mechanism from music perception for the perception of speech, it is further suggested that a language has only 'limited numbers of linguistically distinctive intonation contours' [1, pp. 197]. Supporting evidences have been drawn from the claim including [2] about the possibility of grouping large set of pitch contours based on small number of distinct pitch contours. Also relevant perceptual research (i.e. [3]) has shown that limited number of basic intonation patterns can be identified based on specific types of rises and falls within pitch contours. Nevertheless, we note that these studies have been focusing mostly on Indo-European languages such as English and Dutch. Actually much of the discussion regarding discourse prosody has been framed by sentence-level prosody only and the concept of 'supra-declination' (i.e. [4]) that cannot faithfully mirror the discourse-level prosody. As demonstrated recently by [5], nevertheless, the prosodic contrastiveness in the higher level of intonation variations of Mandarin speech is not as distinct as the English data. Instead a sharper F0 contrast is required for the lower level prosodic-word or -phrase units, comparing to a much flatter F0 contour at the discourse level from Mandarin data [5]. These findings in reflecting language-specific prosodic realizations at higher discourse prosodic levels thus call for an alternative treatment towards the 'coarsely graded' nature of discourse prosody.

It is based on the suggestion that there exists limited number of intonation contours for discourse prosody that we propose the convergence of perceived prosodic highlight allocations should be achievable. The method of converging patterns based on the relative up- and down-stepping of prosodic highlight allocations by the lower discourse prosodic levels is attempted to capture the nature of discourse prosody in Mandarin speech. Specific research questions include: 1). Can perceived prosodic highlight allocations from lower-level discourse prosodic units be converged into limited numbers of prosodic variations in reflecting discourse prosody; 2). What are the factors influencing the successful convergence and narrowing down of prosodic highlight patterns, as well as factors for the divergence found among prosodic variations that cannot be further merged. As will be shown, while the results demonstrate the possibility of narrowing down numbers of prosodic patterns after the convergence (at least for the discourse level of breathing group), the divergence is found to be related to the chunking size of discourse-prosodic units. In particular, it is suggested that the larger the chunking size is, the more divergence found in the prosodic variations that cannot be merged across speech genres. Thus through the cross-speech genre comparison the study sheds light on the **coarsely graded** nature of prosodic variations for discourse prosody from Mandarin speech.

2. Speech data and annotations

2.1. Speech data

Two types of read and spontaneous speech respectively are incorporated for current analyses. The read speech includes

data produced via tasks of prose reading (CNA) and weather broadcast simulation (WB). As for spontaneous speech, one type is a university classroom lecture (SpnL) and the other a spontaneous conversational interaction (SpnC). The following Table 1 summarizes the amount of data from each speech style incorporated in this study.

Table 1. *Summary of total time and number of syllable of the data from 4 speech genres.*

Corpora	Total time (min)	Total number of Syl
CNA	50	22988
WB	28	14083
SpnL	145	33306
SpnC	54	10756

2.2. Preprocessing and annotations

For preprocessing, the speech data first underwent force alignments by the HTK Toolkit, and the output was then manually checked by trained transcribers. Afterwards the data have undergone labor-intensive annotations in separate layers for at least the following information.

2.2.1. Annotation for discourse-prosodic unit

Following the hierarchical prosodic phrase grouping (HPG) proposed by [6], [7], and [8], 5 levels of discourse-prosodic units (DPU) were annotated for all speech data. The 5 levels are marked from B1 through B5, corresponding respectively to syllable (SYL), prosodic word (PW), prosodic phrase (PPh), breath group (BG, a physio-linguistic unit constrained by change of breath while speaking continuously) and multiple phrase speech paragraph (PG) [6]. By default the boundary breaks, prosodic units and their relationship within the HPG framework could be stated as: SYL/B1<PW/B2<PPh/B3<BG/B4<PG/B5 [8].

2.2.2. Annotation for perceived prosodic highlight

The same speech data were manually tagged by trained annotators, in a separate layer, into a string of perceived emphasis/non-emphasis tokens (ETs) according to 4 relative degrees of perceived strength of prominences, following the definitions:

- E0 -- reduced pitch, lowered volume, and/or contracted segments
- E1 -- normal pitch, normal volume and clearly produced segments
- E2 -- raised pitch, louder volume and irrespective of the speaker's tone of voice
- E3 -- higher raised pitch, louder volume and with the speaker's change of tone of voice

By such annotation, we stress the fact that degrees of prominences can be perceived consistently by only limited number of contrastive levels. In addition, note that among the 4 styles of speech data only SpnL and SpnC were annotated for reduced level E0, as it is assumed that speakers rarely carry out reduction given the paradigms of the reading tasks.

3. Methodology

To test the feasibility of converging the prosodic highlight allocations from lower-level DPUs, we follow the rationale from the previous study [5], which concentrates on the allocation of perceived prosodic highlights at PPh level of the HPG framework. In [5], patterns of perceived prosodic highlights are examined by adopting the high/low (H/L) concept in phonology (i.e., [9]). In other words, the patterns derived do not represent the independent occurrence of a certain allocation pattern. Rather, each pattern reflects the up- or down-stepping of the relative prosodic prominence within the lower-level DPU. In this study similar rationale also holds for merging of prosodic highlight allocations in correspondence to the DPU level of **BG**. We adhere to the following steps for the convergence process (also illustrated in Fig. 1):

- Step 1: The first PPh within each BG is set as 1, serving as the initial anchor point. In the case when the following PPh corresponds to the same emphasis token (ET) pattern as its preceding PPh, a number 0 is assigned; otherwise the number 1 will be given. As result, the up- and down-stepping based on the relative ET patterns within the BG can be transcribed into a number string consisting of 0 and 1, as shown in the layer below the ET pattern layer in Fig. 1.
- Step 2: As the F0 contour within each BG has been transcribed into the number string by 0/1, we arbitrarily break it up into subgroups, whenever a change from 0 to 1 occurs. The purpose is to further merge the original patterns in reflecting the allocation of perceived prosodic highlights.
- Step 3: Given that the F0 contour transcribed by 0/1 statues can only observe the relationship from ET patterns of adjacent PPhs, again we transform the number string into an alphabet order, following these rules: a. The first PPh without any relative value is always set as 'A'; b. If the following PPhs correspond to an ET pattern that is different from any of the previous patterns, a new alphabet will be assigned; otherwise the PPhs of the same ET pattern will be labeled by the same alphabet (e.g. as demonstrated by the alphabet assignment layer in Fig. 1).
- Step 4: Finally, to optimally capture the prosodic variations, the alphabet sequence is further narrowed down and the final output is derived via collapsing two or more occurrences of the same alphabet into one.

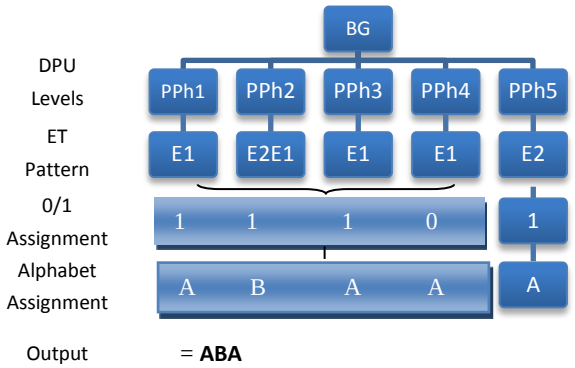

Figure 1: *An illustration of steps taken for the process of convergence.*

4. Results and discussions

4.1. Chunking size of DPU

First of all, a summary of the chunking size of the DPUs at each level of the HPG framework is provided in Table 2. The chunking size by mean syllables has been calculated, and the results serve as the reference point of the following analyses. As shown, the chunking size of lower levels PW and PPh is quite consistent across the 4 genres; as for the upper discourse-prosodic levels (i.e. BG & PG) the chunking size varies drastically, especially for the spontaneous speech data SpnL and SpnC.

Table 2. *Chunking size by mean syllables*. (Std in parenthesis)

DPU / Genres	PW	PPh	BG	PG
CNA	2.2 (0.7)	7.5 (4.1)	24 (13)	78 (49)
WB	2.3 (0.5)	9.6 (5.2)	38 (24)	95 (66)
SpnL	2.4 (0.9)	7.3 (5.7)	121 (99)	629 (516)
SpnC	2.2 (0.9)	7.8 (7.0)	36 (29)	1160 (595)

4.2. The convergence of perceived emphasis patterns at BG

Following the methodology from Section 3, we start out by testing the feasibility of converging the ET patterns from PPh in reflecting the relative prosodic variations at the DPU level of **BG**. In Table 3 it summarizes numbers of perceived emphasis patterns before and after the convergence. The distribution of converged emphasis variations by BG is otherwise presented in Figure 2.

Table 3. *Summary of numbers of perceived ET patterns before and after convergence.*

Genre / Pattern	CNA	WB	SpnL	SpnC
# Before converging	373	201	454	181
# After converging	56	52	205	52

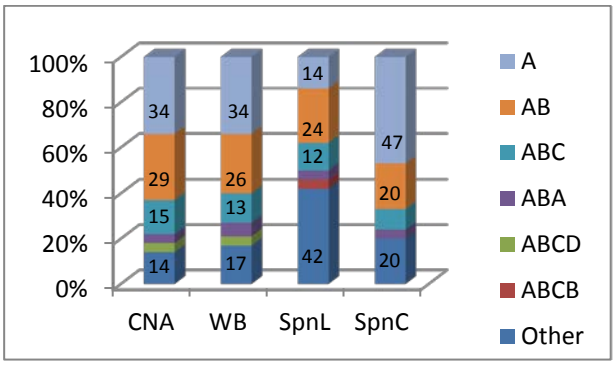

Figure 2: *Distribution of prosodic variations after convergence at **BG*** (numbers under 10% are excluded).

4.2.1. Discussion

From Table 3, the results demonstrate that after merging, numbers of perceived emphasis variations narrow down considerably except for the speech style of SpnL. With the other three speech styles, we are able to narrow down around 70-85% of the perceived emphasis patterns out of the convergence and the numbers of variations are down to around 50. Moreover, based on Fig. 2 we find that these 50 or so emphasis patterns from the 3 styles CNA/WB/SpnC are distributed among limited numbers of prosodic variations. Actually, leaving the category 'other' aside, at least 80% of the emphasis patterns in these 3 speech styles can be merged into 4 major variations in common (i.e. A/AB/ABC/ABA). The result thus exhibits the successful convergence of perceived prosodic highlights for upper-level discourse prosody following the proposed methodology.

As for the spontaneous speech of classroom lecture (SpnL), surprisingly it is the only speech style demonstrating the discrepancy and we are able to narrow down about only half of the emphasis patterns (i.e. Table 3). Turning to the distribution of converged emphasis variations, there exists at least 40% from the category 'other' in which the emphasis patterns cannot be further merged, though the rest can still be account for by those 4 major variations identified as in the other 3 styles. In order to uncover why there exists such incongruity from SpnL, we consider the possibilities that: a). the annotation of reduction E0 for perceived prominences could be a contributing factor; b). the high divergence of non-merge-able emphasis patterns from SpnL is related to the chunking size at the DPU levels. We will explore these two plausible explanations next.

4.3. Considering reduction (E0) in the convergence

Here we set forth the test to find out if the E0 of the perceived prosodic highlight labeling might be responsible for the divergence in emphasis patterns that cannot be merged from SpnL. In fact, one reason to take E0 into account is that, as has been shown in [10], the addition of one level of reduction (E0) to the same set of spontaneous speech data does result in an increase in the varieties of emphasis token patterns at different DPU levels. Thus we follow it up by testing if the merging of the tags E0 and E1 in SpnL and SpnC can contribute to a better result after the convergence. In other words, we try to find out if the emphasis patterns can be further combined after narrowing down the levels of perceived emphases. This test is based on the assumption that the contrast degree between E0 and E1 should be of minimal influence towards the overall layout of intonation contour. The result is presented in Figure 3, after we execute the convergence procedures following the combination of perceived prosodic highlight levels E0 and E1:

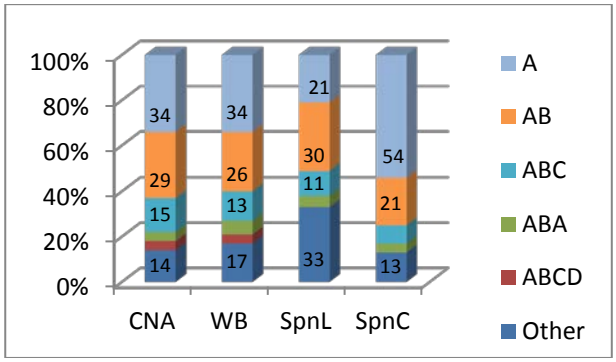

Figure 3: *Distribution of prosodic variations after convergence based on combining E1 and E0* (numbers under 10% are excluded from the Fig).

4.3.1. Discussion

After combining E1 and E0 into one emphasis level, followed by the merging of emphasis patterns, we find the numbers of

the category 'other' in SpnL and SpnC indeed can be narrowed down. Comparing to Fig. 2, however, the result from combining E1 and E0 is not as effective as expected, i.e. the proportion of the category 'other' can only be narrowed down by less than 10% (SpnL: 42% down to 33%; SpnC: 20% to 13%). In other words, although decreasing levels of perceived prominences does improve the results from the convergence, it does not pose as a major factor changing the distribution among the converged variations across speech genres. Mostly, the overall distribution does not change and this still does not offer a reasonable explanation why SpnL would be the style with the most divergent 'other' category across genres. It is thus concluded that the decreasing of emphasis levels would not result in much significant impact on the convergence of perceived emphasis patterns for the higher-level discourse prosody.

4.4. The convergence of perceived emphasis patterns at PG

The second test involves examining if the more divergent category 'other' found in SpnL data might be related to the planning size. Following the same methodology proposed in Section 3, we attempt the feasibility of narrowing down the variations of relative prosodic highlight patterns at the top discourse level of **PG**. The results by the convergence are presented in the following Fig. 4 (note the exact number in percentage is excluded from the Fig. if under 5%).

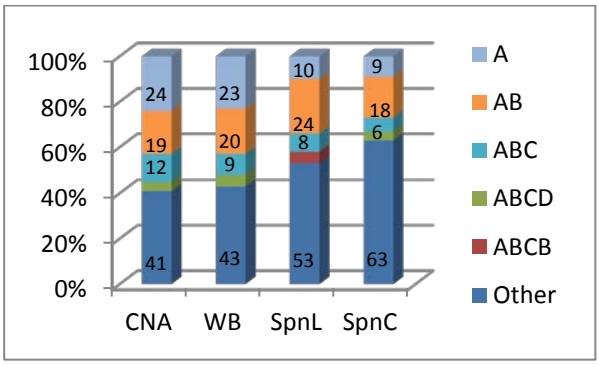

Figure 4: *Distribution of prosodic variations after the convergence at **PG***

4.4.1. Discussion

Fig. 4 demonstrates that, after converging the perceived emphasis patterns by the PG level, the variations can still be narrowed down to 3 major ones, namely **A, AB,** and **ABC** across the 4 speech styles. Note that, however, none of the variations takes up more than 25% of the overall distribution within each speech style. Instead, the category 'other' within each genre increases significantly. The increased proportion of the category 'other' otherwise implies that after converging, patterns derived from the perceived prosodic highlights are simply too diverged to be collapsed into one unique variation.

Moreover, based on results of cross-genre comparison, it has shown clearly that at PG level the SpnC style displays the largest proportion of the category 'other' after converging, i.e. 63% of the emphasis patterns are just too diverged from one another and cannot be further merged into a unique variation. Interestingly, if we turn to the DPU chunking size summarized in Table 2 earlier, SpnC indeed corresponds to the largest planning unit at PG level, as for BG level it is the SpnL that reflects the largest size of speech planning. The findings, therefore, illustrate most notably that the larger the chunking size is, the more diverged variations we would arrive at after the convergence taking place. In summary, one of the key factors influencing the successful convergence of perceived prosodic highlights is hence the *chunking size*, i.e. how speakers plan for the unit size at higher DPU levels that are discourse based. When the unit size of upper-level DPU gets larger, it allows for more room to allocate the ups and downs of the intonation contours. This in turn explains why we found in our SpnL data the highest percentage of the category 'other' at BG level.

5. General discussion and summary

In this study we concentrate on the convergence of perceived prosodic highlights from the lower-level DPU into prosodic variations that correspond to the upper-level discourse prosody. Our assumption is based on that a general tendency of **coarsely graded** prosodic contrastiveness can be found in discourse prosody. From results of cross speech genres comparison, it is demonstrated that prosodic variations composed of perceived emphasis token patterns from lower DPU levels can be successfully narrowed down after the convergence taking place. Most of all, in exploring the factors influencing the successful merging of perceived prosodic highlights, it is further identified that the planning size of DPU is directly associated with the divergence found in the category 'other', i.e. patterns that cannot be further merged. This is substantiated by the convergence results at both BG and PG, which reveals that the larger the chunking size is at these levels; the more divergent the category 'other' turns out to be.

The study also tests if merging emphasis levels (E1 &E0) may contribute to the further convergence of prosodic patterns. The result, however, indicates that such level combination does not lead to much significant improvement after the convergence. Interestingly, as suggested previously in [10] that by adding the emphasis level of reduction (E0) it results in an increase of perceived emphasis token patterns. Thus we preliminarily conclude that the reduction in levels of perceived prominences may play a much more significant role in the allocations of prosodic highlights at lower-levels of DPU, but less so at the higher-level discourse prosody.

Parallel to the motivic similarity found in the music perception, therefore, the current study showcases that perceived prosodic highlight allocations can be converged in reflecting the **coarsely graded** nature of the overall discourse prosody. Most of all, our discussion of discourse prosody has been focused on the granularity of intonational variations grounded in prosodic contrastiveness, which is analogous to the graded feature of motivic similarity in the music study. Eventually the current findings contribute the substantiation of the coarsely graded variations in the realization of discourse prosody from diverse Mandarin speech genres. For the direction of future research, we plan to explore further how the coarsely defined prosodic convergence interacts with the allocation of information status and information content.

6. References

[1] A. D. Patel, *Music, Language, and the Brain.* New York: Oxford University Press. 2008.

[2] M. A. K. Halliday. *A course in Spoken English: Intonation.* London: Oxford University Press. 1970.

[3] J. 't Hart, R. Collier, and A. Cohen. *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge, UK: Cambridge University Press. 1990.
[4] A. Wichmann. *Intonation in Text and Discourse: Beginnings, Middles and Ends*. London: Routledge. 2014.
[5] C. Tseng and C. Su. "Where and How to Make an Emphasis? – L2 Distinct Prosody and Why." in *ISCSLP 2014 - 9th International Symposium on Chinese Spoken Language Processing, September 12-14, Singapore, Proceedings*, 2014, pp. 633-637.
[6] C. Tseng, S. Pin, Y. Lee, H. Wang, and Y. Chen. "Fluent speech prosody: Framework and modeling," *Speech Communication*, Vol. 46, No. 3-4, pp. 284-309, 2008.
[7] C. Tseng and C. Su. "Discourse prosody and context – Global F0 and tempo modulations," in *INTERSPEECH 2008 – 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, Proceeding*, 2008, pp.1200-1203.
[8] C. Tseng. "An F0 analysis of discourse construction and global information in realized narrative prosody," *Language and Linguistics*, Vol. 11, No. 2, pp. 183-218. 2010.
[9] http://www.ling.ohio-state.edu/~tobi/
[10] H. Chen, W. Fang, and C. Tseng. "Information content, weighting, and distribution in continuous speech prosody – A cross-genre comparison. In *Oriental-COCCOSDA 2015, October 27-30, Shanghai, China, Proceedings*, 2015, pp. 117-122.