Advance Prosodic Indexing — Acoustic realization of prompted information projection in continuous speeches and discourses

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

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

Abstract

This study aims at exploring the mechanism of information planning in continuous speeches and discourses. The main focus is on how information planning is signaled via advance prosodic prompting and indexing. Recently it has been identified that a perceivable prosodic *projector* with its intended *projection* (as opposing to prosodic highlight associated key information) form a crucial information unit that signals up-coming focal information ahead of time in order to facilitate prediction. Here F0 realization of such *projector* **PJR** plus the immediately following *projection* **PJN** across four speech genres demonstrate that a general high-falling contour can be identified in most of the data. Further removal of intonation effect and calculation of down-stepping degree reveal a *positive correlation* in that the bigger the *projection* trajectory, the larger the *down-stepping* degree found between the beginning and ending of **PJR-PJN** unit. Thus the study sheds lights on how advance prosodic prompting indexes upcoming information; current results offer strong evidences of information planning and arrangement at discourse levels and help facilitate better account of context prosody.

Index Terms: continuous speeches and discourses, perceived prosodic prompted projector and projection, information planning, advance prosodic indexing, context prosody

1. Introduction

In continuous speeches and discourses, speakers plan ahead of time allocating information in order to effectively and optimally make the communication across. One crucial aspect of information planning concerns how speakers allocate key information for interlocutors to pinpoint the most salient part from the speech flow. Traditionally it has been suggested that the focal and most salient information is directly associated with prominences in speech prosody [1], [2]. What's been less discussed until of late is other functions of prosodic highlight (or perceived emphasis), i.e. to index 'specific parts of discourse' [3, P.8]. It is held here in particular that functions of *advance* indexing should at least take into consideration the ability to generate anticipated expectations and predictions of the *projection* for up-coming information during discourse planning [4], [5], [6]. After all, the processing of information throughout the discourse involves much of complex prediction-making from interlocutors and their constant attempts to minimize errors from such predictions [7], [8].

The present study hence focuses on the *projecting* function of perceived prominences annotated across 4 genres of continuous speeches and discourses in diversity. As have been identified recently [5], almost 70% of annotated perceived emphasis tokens within continuous speeches correspond to either key information, or *projector* indexing *in advance* the up-coming key information. It has been further noted that the use of prosodic prompted *projector* outnumbers that of prominence-corresponding key information by about 15-25% [5]. Such result reflects that the *advance indexing* function of *projector* and its intended *projection* of information planning during discourse exchanges deem more attention (cf. [6]). It is thus held that the mechanism of on-line information planning in continuous speeches and discourses is reflected in the allocation of perceived prosodic highlight corresponding to not merely focal information, but most of all *advance indexing* of soon-to-arrive key information via its *projection*.

Currently prosodic prompted *projector* (henceforth **PJR**) is defined as perceived prominence that functions to create in advance expectations for up-coming key information, following [6]. It has been shown that **PJR**, together with its perceived trajectory of *projection* (hereafter **PJN**), collectively form a unit of information planning **PJR-PJN** during continuous speeches and discourses. As explicated in [6], **PJR-PJN** unit has been defined as carrying at least a piece of focal information within its *projection* trajectory. Actually, the concept of projection can be held in similar spirit as the discussion regarding *projection principle* from theoretical syntax [9], [10]: such as in Mandarin it is expected that a numeral plus a classifier together *project* a NP in the immediate following. We note, however, that syntactically defined projection is merely accountable for phrase-/sentence-level information planning but often does not translate well to the planning beyond sentences in continuous discourses. Indeed it has been shown how **PJR-PJN** as a unit of information planning in continuous speeches includes not only the local but most of all global predictions about the allocation of up-coming information [6]. Their interactions cause same-level as well as hierarchical trade-off and compensation to generate context prosody that may appear complex on the surface, but is in fact systematic and predictable.

This study aims at the acoustic features of prosodic highlight prompted information unit **PJR-PJN** in 4 diverse speech and discourse genres. Specifically, we focus on the realization of F0 across the annotated **PJR-PJN** unit. By examining F0 feature with and without intonation effect from higher-level discourse units, the main purpose lies in the identification of a general tendency towards the prosodic realization of the named unit. In particular we take into consideration of both local (hence more syntactic oriented) and global (therefore more towards discourse levels) projection in constituting **PJR-PJN** unit. As will be shown, results draw a general pattern of high-to-low falling contour across the unit from most data, inclusive or exclusive of higher-level intonation. Additional supportive evidences are provided by calculating *down-stepping degree* between the starting and end points of PJR-PJN unit, which otherwise reflects a *positive* correlation between down-stepping degree and projection trajectory size. Eventually results based on cross speech genre comparison is oriented towards a solid account for the mechanism behind information planning by *context prosody* in continuous speeches and discourses: with the objective of deriving underlying acoustic patterns in terms of prosody indexed information projection out of surface variations and realizations within speech signals.

2. Speech data and annotations

2.1. Speech data

For present analyses we incorporate Mandarin speech data from 4 diverse genres, in which 2 are spontaneous speeches and 2 read speeches. For spontaneous speeches, one is a university classroom lecture (**SpnL**) and the other a spontaneous face-to-face interaction (**SpnC**). Read speeches include data from tasks of prose reading (**CNA**) and weather broadcast simulation (**WB**). Note that we incorporate data of read speeches for the purpose of comparing with features belonging to spontaneous discourses. Table 1 summarizes the total amount of speech data from each genre.

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

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

2.2. Preprocessing and annotations

All above speech data have first undergone preprocessing of force alignments by HTK Toolkit. The output was then manually checked and adjusted by trained transcribers. Next the data have been tagged, via labor-intensive annotations, in separate layers for the following information.

2.2.1. Annotations for discourse-prosodic unit (DPU)

First of all, the data have been annotated for levels of discourse-prosodic units (DPU), adhering to the hierarchical prosodic phrase grouping framework (HPG) proposed by [11], [12], and [13]. In HPG framework it includes 5 DPU levels, marked from B1 through B5 that correspond respectively to: *syllable* (**SYL**), *prosodic word* (**PW**), *prosodic phrase* (**PPh**), *breath group* (**BG**, a physio-constrained unit defined by change of breathe while speaking continuously) and *multiple phrase speech paragraph* (PG) [11]. By default the relationship between prosodic units and boundary breaks could be accounted for by: SYL/B1 <PW/B2 <PPh/B3 <BG/B4 <PG/B5 [13].

2.2.2. Annotations for perceived prosodic highlight

In a separated layer we manually tagged the same data into a string of perceived emphasis/non-emphasis tokens. The annotation is based on perceived strength of prominences in 4 relative degrees including:

- 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

Note among the 4 speech genres, only **SpnL** and **SpnC** were annotated for E0. This is based on the assumption that speakers rarely carry out reductions in reading tasks.

*2.2.3. Annotations for information unit **PJR-PJN***

The annotation for information unit **PJR-PJN** is done in yet a separate layer. The identification of *projector* **PJR** is by perceived prosodic highlights already annotated in speech data (and by *prosodic word* **PW** corresponding to possible candidates such as modifiers, conjunction, as well as verbs that take objects including clausal ones, i.e. [5], [6]). Following the definition from above and also [14], the respective *projection* **PJN** to each **PJR** is identified as the anticipated moment of syntactic and/or semantic completion, occasionally coincides with prosodic completion and covers at least a piece of focal information. In addition, it is noted that *projection* trajectory can be of different size, from the local to the global one, as shown in the following:

- '那也是**最早的一篇**文章' (**SpnL**: *local* projection) (1)
- '**為什麼**直接比對字也有困難?因為我們詞的結構是非常 flexible 的' (**SpnL**: *global* projection) (2)

In (1) the prosodic highlight prompted (as shown by bolded underline) **PJR** '**最早的一篇**' in the *prosodic phrase* (PPh) '那也是最早的一篇文章' would have its respective **PJN** trajectory fall by the end of the following NP '文章', hence forming a location projection. As for global projection, the prosodic highlight indexed **PJR** '**為什麼**' in (2) can entail a projection with larger trajectory that could extend over the immediate PPh boundary in the following.

3. Methodology

The methodology incorporated in the current study involves mainly extraction of acoustic feature F0 across PJR-PJN information unit. First of all, F0 values (in semitone) across the unit are automatically extracted by using software program PRAAT (© [15]). In order to facilitate further comparison across speakers and eliminate in-between speaker discrepancies, all extracted values are subject to Z-score normalization [16], following (3):

$$Z = \frac{x - \mu}{\sigma} \quad (3)$$

where µ stands for average F0 and σ standard deviation of F0 values from each speaker.

The next step involves calculation of average F0 value. Here we take *prosodic word* **PW** as the base unit for calculation. Since **PJR-PJN** unit could be of different length depending on its projection trajectory size, we take the average F0 value from sampling points including: 1). the 1st PW at the starting point of **PJR**; 2). the ending PW right by the completion of *projection*; and 3). PWs at pre-/post- **PPh** boundaries, depending on the trajectory size. Finally, after deriving average F0 values from each sampling point, we further attempt the removal of intonation effect from higher-level discourse units: by each **PPh** unit we derive a linear regression line, whose position and slope is such so as to minimize the distance from the line to each data point. Finally, each slope from the linear regression line has undergone normalization so the slope of the line is zero.

4. Discussion

4.1. PJR-PJN unit by PPh

As the trajectory size of the projection varies by each *projector* **PJR**- *projection* **PJN** information unit, we wonder what the general distribution of the unit size could be across different speech genres. Hence we first calculate the range distribution of **PJR-PJN** unit. Here the calculation is done by the discourse-prosodic unit *prosodic phrase* (**PPh**) and results are summarized in Table 2.

From Table 2, it is demonstrated that over 50% of **PJR-PJN** unit have their trajectory size by the boundary of at least one PPh. This suggests that **PJR-PJN** units annotated across our speech data are not limited to merely local projections by adjacent syntactic units. Moreover, over 90% of the unit can be accounted for by up to 3 PPhs. Based on the findings, in following analyses we concentrate on **PJR-PJN** units expanding from 1 to 3 PPhs across speech genres.

Table 2. *Summary of total time and number of syllable of speech data from 4 genres.*

Genre PPh #	SpnL	SpnC	CNA	WB
1	66%	55%	63%	77%
2	18%	28%	25%	13%
3	7%	8%	6%	3%
Over 3	9%	9%	6%	7%

4.2. Acoustic feature: F0

We start out by calculating the average F0 value across **PJR-PJN** information unit, after normalization. Following the methodology from Section 3, we calculate F0 values by *prosodic word* **PW** and across sampling points include the first and last PWs from the **PJR-PJN** unit, as well as from PWs located right by pre-/post- **PPh** boundaries, depending on the projection trajectory size. The results are summarized in Fig. 1 (Note that top panels are results from SpnL/SpnC, and bottom panels CNA/WB; the vertical axis stands for normalized F0 and horizontal axis sampling point positions).

4.2.1. Discussion

From Fig. 1, a general tendency of *high-to-low* falling pitch contour is observed across PJR-PJN unit, regardless the projection trajectory size. Note also there are exceptions such as a slight final rising contour is observed from: SpnC when the unit extends up to 2 PPhs, and WB when the unit expands to 3 PPhs. The slight final rising contours in these cases, however, never reach higher than the F0 value taken from their respective **PJR** beginnings. Most of all, based on the standard error bars from Fig. 1 across all panels, obviously the **PJR** beginnings are distinct from the corresponding **PJN** endings regardless of trajectory size. Further statistical test actually reports that significant differences are found (h = 1, P<.05 presented across all panels).

Based on the findings, therefore, it is suggested that while planning for the **PJR-PJN** pair as an information unit, mostly speakers tend to start at a higher F0 from **PJR** beginning and continues with a falling contour across the projection trajectory. Although there are cases when slight rising contours do occur, the rising would never reach higher than the mean F0 from the starting points of **PJR**. Moreover, there seems to be a tendency that the larger the projection (i.e. when the trajectory expands over 2 PPhs), the greater the difference between the mean F0 from the starting and ending points of the units. This may suggest that in the fore-planning of the larger projection trajectory, speakers would also have to prepare for a greater F0 range across the unit in order to allow for further allocation of prosodic highlights within the unit. Given the observation, we wonder if further removal of higher level intonation effect might offer additional evidences. Hence we attempt the removal of intonation effect in the next section.

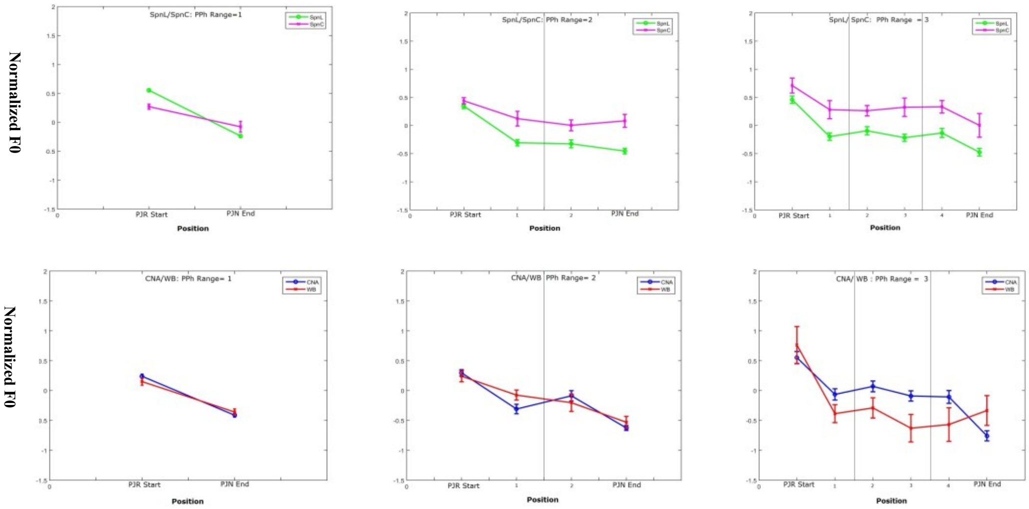

Figure 1: *F0 from **PJR-PJN** information unit (By positions: 1/2= PW prior/post to 1st PPh boundary; 3/4= PW prior/post to 2nd PPh boundary; the solid black lines stand for PPh boundaries).*

4.3. Acoustic feature: F0 without intonation effect

We try to remove the higher-level intonation effect from F0 values reported previously. Following the methodology from Section 3, we present results of F0 without intonation effect in Fig. 2. Note the results are arranged by numbers of **PPhs**, depending on the projection trajectory size.

4.3.1. Discussion

Fig. 2 demonstrates that, after removing intonation effect, the falling pitch contour across **PJR-PJN** unit can still be observed. This is more clearly shown in the two spontaneous speech genres (top panels). As for read speeches, it is the least obvious when the projection is only of local scale (i.e. within 1 **PPh**). Although we do find slight rising contours at projection endings in some panels, the risings in most cases do not reach higher than the corresponding **PJR** beginnings (*except* for read speech data when **PJN** = 1PPh). Additional T-test taken between the values from **PJR** beginning and **PJN** ending points report that they can be distinguished (all h = 1, P≦.05), except for read speech data in which the projection is only of local, i.e. within 1PPh. The result is thus evident in that although localized phenomena (such as local projections) vary across speech genres, a general tendency could be well preserved and captured when taking into account the global planning of information.

Another interesting finding is that, when the trajectory of projection extends over PPh boundaries, we further notice a *down-stepping* trend from the overall pitch contour. This trend is most noticeable by **PPh** boundaries. To further substantiate the observation, we carry out a calculation of ***down-stepping degree***, which is defined as the difference between values from the starting and ending points of **PJR-PJN** unit. The results are summarized in Table 3:

Table 2. *Down-stepping degree across **PJR-PJN** unit.*

Down-stepping degree **Genre**	**Within PPh**	**Cross 1 PPh**	**Cross 2 PPhs**
CNA	0.067	0.234	0.789
WB	0.049	0.452	0.614
SpnL	0.294	0.600	0.700
SpnC	0.173	0.316	0.553

From Table 3, it is clearly that a *positive correlation* can be derived between the down-stepping degree and *projection* trajectory size; in other words, as the trajectory size gets longer, the *degree* difference between the starting and ending points of the unit also increases. Most of all, we arrive at such result after the removal of the higher-level intonation effect. Thus this further implies that the intonation effect resulting from discourse-prosodic boundaries (cf. [17]) does not override the overall intonation planning across **PJR-PJN** unit. In turn, our findings here reinforces that in order to plan for a larger size of projection trajectory, speakers by default orient to a noticeable falling contour and larger down-stepping degree so as to allow for more prosodic variations in corresponding to prosodic highlight allocations and hence information planning within the projection trajectory, especially for higher-level, context-attributed prosody.

5. General discussion and summary

The current study focuses on acoustic features of advance information *projection* in continuous speeches and discourses of various genres. We examine particularly F0 feature of perceived *projector* **PJR** tagged with its range of intended trajectory of *projection* **PJN** as the systematic prosodic indexing of upcoming focal information. Though the projection trajectory size in each case differs, we are able to quantitatively derive a general tendency of falling contour across the *projection* trajectory, especially when extending over PPh boundaries. Moreover, by removing higher-level intonation effect and calculating ***down-stepping degree***, the finding is further substantiated. In the end a positive correlation between the projection size and the value differences from the beginning and ending of the named unit have also been validated. Eventually we offer evidences to the constitution of global *context prosody* showing that it is in fact both accountable and predictable.

In sum, current findings specifically foreground the acoustic feature of advance prosodic indexing of information projection. A solid acoustic down-stepping pattern has been derived only when we take into consideration the largest trajectory of projection and without intonation effect. Most of all, our analyses illustrate a generalization reached only when taking into consideration expected projection expanding beyond local syntactic associations and out of surface variations. In future studies, we plan to explore: 1). other acoustic features such as duration; 2). the actual information arrangement and allocation by a direct association with perceived prosodic highlights within PJR-PJN unit; and 3) further validations from across-language speech data.

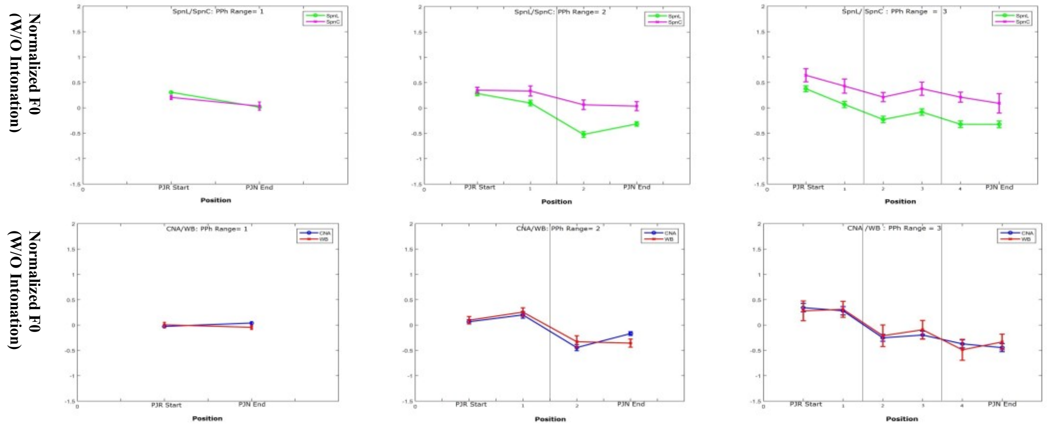

Figure 2: *F0 from **PJR-PJN** information unit without intonation effect (By positions: 1/2= PW prior/post to 1st PPh boundary; 3/4= PW prior/post to 2nd PPh boundary; the solid black lines stand for PPh boundaries).*

6. References

[1] M. A. K. Halliday, "Notes on transitivity and theme in English," *Journal of Linguistics* Vol. 3, 1967, pp. 199-244.

[2] J. B. Pierrehumbert, J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," In *Intentions in communication*, P. Cohen et al. (eds.), Cambridge: MIT Press, 1990, pp. 271-311.

[3] S. Falk, On the notion of salience in spoken discourse-prominence cues shaping discourse structure and comprehension. *Travaux interdisciplinaires sur la parole et le langage*, Vol. 30, 2014, pp. 1-30.

[4] P. Auer, "On-line syntax: thoughts on the temporality of spoken language," *Language Sciences*, Vol. 31, 2009, pp1-13.

[5] H. Chen, W. Fang, C. Tseng, "Prosodic prompts of information content in speech - A cross genre comparison of prominence as key, projector and projections," in *IACL 2016 – the 24th Annual Conference of International Association of Chinese Linguistics, July 17-19, Beijing, China*, 2016.

[6] H. Chen, W. Fang, and C. Tseng, "Prosodic prompts and information planning units in continuous speech— Relative allocation and compensation of prosodic highlight," in *the 12th Phonetic Conference of China, July 25-26, Tongliao, China*, 2016.

[7] L. Dilley, "Rhythm, context effects, and prediction," Proc. In *Speech Prosody 2016* – 8th Speech Prosody conference, May 31-Jun 1, Boston, USA, 2016.

[8] A. Clark, "Whatever next? Predictive brains, situated agents, and the future of cognitive science," *Behavioral and brain sciences*, vol. 36, no. 3, 2013, pp.181-204.

[9] N, Chomsky, *Knowledge of language: Its nature, origin, and use*, Greenwood Publishing Group, 1986.

[10] L. Haegeman, *Introduction to Government and Binding Theory*, Blackwell Publishing, 1994.

[11] C. Tseng, S. Pin, Y. Lee, H. Wang, Y. Chen, "Fluent speech prosody: Framework and modeling," *Speech communication*, vol. 46, no. 3-4, 2008, pp. 284-309.

[12] C. Tseng and Z. Su, "Discourse prosody and context–global F0 and tempo modulations," in *INTERSPEECH 2008 – 9th Annual Conference of the International Speech communication Association*, Sept. 22-26, Brisbane, Australia, Proceedings, 2008, pp.1200-1203.

[13] C. Tseng, "An F0 analysis of discourse construction and global information in realized narrative prosody," *Language and Linguistics*, Vol. 11, no. 2, 2010, pp. 183-218.

[14] J. De Ruiter, H. Mitterer, N. J. Enfield, "Projecting the end of a speaker's turn: a cognitive cornerstone of conversation," *Language*, Vol. 82, No. 3, 2006, pp. 515-535.

[15] P. Boersma, D. Weenink, *Praat: Doing phonetics by computer*, (www.praat.org) (retrieved on 20 Nov, 2012.)

[16] R. J. Larsen, L. Marx *An Introduction to Mathematical Statistics and Its Applications*, Prentice-Hall, Englewood Cliffs, NJ, 2000.

[17] C. Tseng and Z. Su, "Boundary and lengthening — On relative phonetic information," In the 8th Phonetics Conference of China and the International Symposium on Phonetic Frontiers. Beijing, China. 2008.