An Initial Investigation of L1 and L2 Discourse Speech Planning in English

Chiu-Yu TSENG[1], Zhao-Yu SU[2] and Chi-Feng HUANG[3]
Phonetics Lab, Institute of Linguistics, Academia Sinica
Taipei, Taiwan
cytling@sinica.edu.tw

Tanya VISCEGLIA[4]
Department of Applied English, Ming Chuan University
Taipei, Taiwan
orlandotaipei@hotmail.com

***Abstract*-A perceptually-based hierarchy of prosodic phrase group (HPG) framework was used in this study to investigate similarities and differences in the size and strategy of discourse-level speech planning across L1 and L2 English speaker groups. While both groups appear to produce similar configurations of acoustic contrasts to signal discourse boundaries, L1 speakers were found to produce these cues more robustly in English. Differences were also found between L1 English and L1 Taiwan Mandarin speaker groups with respect to the distribution of prosodic break levels and break locations. These differences in L1 and L2 organization of discourse speech prosody in English can be largely attributed to between-group differences in speech planning and chunking strategies whereby L2 speakers use more intermediate chunking units and fewer larger-scale planning units in their prosodic discourse organization. Through more understanding of prosody transfer, we believe that technology developed on the basis of L1 Mandarin spoken language processing may be applied to L2 English produced by the same speaker population, with little modification.**

Keywords*- *HPG, L1 and L2 English, discourse planning, information chunk

I. Introduction

Recent studies have shown that organization of discourse prosody differs between L1 and L2 speakers, and that some differences are likely to affect comprehensibility [1], defined as level of difficulty in following the speaker's intended meaning and/or sequencing of information [2]. Such studies have often measured differences between L1 and L2 speech by tracking F0 movement in a speaker's register, within and across intonational paragraphs [3][4]. The present study can more precisely investigate the layering of acoustic cues which comprises discourse prosody, as well as the individual contributions of F0, duration and amplitude at each prosodic level, using the perception-based hierarchical discourse prosody framework HPG (Hierarchy of Prosodic Phrase Group), which we have developed in previous research [5]. The data presented here represent the results of applying the HPG framework to compare L1 and L2 strategies for prosodic organization at the discourse level in English.

HPG's prosodic units, in ascending order of size, are defined as 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. The physio-linguistic unit BG, absent from many other frameworks, corresponds to an audible and complete change of breath; it has been included [6][7][8] to accommodate the physical necessity of breathing during continuous speech production. The five discourse boundary break strengths corresponding to each of the HPG units are: B1/SYL, B2/PW, B3/PPh, B4/BG and B5/PG. The relationship between these prosodic units and boundary breaks is revealed only in larger units of discourse; it can be expressed schematically as: SYL<PW<PPh<BG<PG and B1<B2<B3<B4<B5 [5]. Applying HPG to L1 and L2 productions of the same English text, we analyzed speech rate, break distribution and planning scale. Overlap of B4 (breath group) position between groups was also analyzed, and a multi-layered acoustic analysis was performed on prosodic boundaries, with the aim of comparing L1 and L2 use of acoustic cues to discriminate B3, B4 and B5.

II. Materials and Procedure

Speakers were instructed to read Aesop's fable "The North Wind and the Sun" aloud. This passage is recommended by the IPA for the purpose of eliciting all phonemic contrasts that occur in English. It contains 144 syllables, 113 words, 8 independent clauses, 5 dependent clauses, 5 sentences, and 3 paragraphs; when read aloud, it is approximately 40~50 seconds in duration. Data was collected from 10 L1 English speakers and 514 L1 Taiwan Mandarin speakers. Pre-processing of recorded data begins with automatic annotation of segmental labeling using the HTK toolkit included in the CMU dictionary. Segmental labeling is then spot-checked by experienced transcribers to ensure precise alignment of phone boundaries. Manual labels of perceived prosodic boundaries (B2, B3, B4 and B5) are also labeled by trained transcribers using HPG protocol [5].

III. Data Analysis

The multiple regression model used to analyze Mandarin discourse in our previous work was modified to reflect the English vowel inventory [5]. HPG was then applied to the English data in order to observe patterning of acoustic correlates at each prosodic layer, using the formula

$$x_i = \mu_i + \sum_{j=1}^{k} factors_j + \varepsilon_i$$

in which x_i denotes response variables and ε_i unpredictable noise. Predictors for x_i are intrinsic attribute (μ_i) and the effect of multiple prosodic layers ($factor_j$), in which j represents the index of each prosodic layer. Intrinsic attribute and the effects of multiple layering also consider vowel identity

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

and the syllable position corresponding to each respective prosodic layer. Since the phonotactic combinations in our data were insufficient to train a higher-level segmentation model, a quantization strategy was adopted for the purpose of modeling higher-level acoustic correlates. PW and PPh are quantized into three syllables and nine syllables, respectively. First, at the SYL layer, predictions of vowel features are determined in relation to the original signal while residues are treated as contributions from the next higher layer. The residues are then included in the next round of predictions, in this case the PW layer. The same predictions are repeated at each layer while prediction accuracy by layer was treated as contributions from respective layers. Cumulative predictions from all layers jointly constitute output prediction.

Whenever insufficient amount of training data were available for a particular prosodic position, a polynomial curve fitting formula (in which the order n is set 3) was used to generate more robust patterns of F0, duration and amplitude.

$$y(t) = a_1 t^n + a_2 t^{n-1} + \ldots + a_n t + a_{n+1} t^0$$

Before modeling, we quantize the size of each prosodic unit as follows: 3 syllables for a PW, 3 PW's (9 syllables) for a PPh, and 3PPh's (27 syllables) for a PG. Results obtained are presented in Section VI.

IV. COMPARISONS OF SPEECH RATE, BREAK DISTRIBUTION AND PLANNING SCALE

A. *Speech Rate*

Table I shows speech rate comparisons of L1 and L2 groups calculated in three ways: syllable number per minute [1], word number per minute [10][11] and number of stressed syllables per minute [12]. These measurement techniques have been used in previous studies to evaluate speaker fluency [13]. However, any method of calculation which employs means and averages cannot capture the internal dynamics present in the flow of continuous speech. This observation may account for our otherwise somewhat puzzling findings that the speech rate of L1 speakers is slower than that of L2 speakers, and that L2 speakers exhibit a much higher level of within-group variation. The HPG framework, in contrast, has been demonstrated in previous studies to reflect and account for dynamic changes in global speech rate [14][15].

TABLE I. SPEECH RATE BY UNIT OF MEASUREMENT AND SPEAKER GROUP

Measurement Speakers	Syl/min *(μ / σ)*	Words/min *(μ / σ)*	Stress/min *(μ / σ)*
L 1	234 / 19	183 / 15	84 / 7
L 2	199 / 40	156 / 32	72 / 15

B. *Distribution of Prosodic Breaks*

TABLE II shows distribution of prosodic boundaries for L1 and L2 speaker groups. The most pronounced difference with respect to distribution of prosodic breaks was found at the B3 level. L2 speech contains more than twice as many B3 breaks as L1 speech, but fewer B4 and B5 breaks overall. Thus, it appears that L2 speakers use more intermediate chunking units and fewer larger-scale planning units in their prosodic discourse organization.

TABLE II. BREAK NUMBER SPEAKER GROUP

Break Speaker	B2 *(μ / σ)*	B3 *(μ / σ)*	B4 *(μ / σ)*	B5 *(μ / σ)*
L1	39.2 / 5	11.3 / 4	4.5 / 1	3.9 / 0
L2	33 / 3	25 / 9	3 / 1	4 / 1

The relative size of discourse units across speaker groups has been calculated by number of syllables, words and stressed words. TABLE III. shows that the size of PW is larger for L2 than L1 speakers. The most pronounced difference between speaker groups is in the size of PPh and BG. L2 PPhs contain fewer syllables than L1 prosodic phrases. L2 BG seems to be larger than L1, but it also has a larger range of variation due to L2 speakers' inconsistent positioning of the B4 (BG) boundary. The size of PG is the same in L1 and L2, most likely influenced by the visible breaks in text presentation, as PG boundary locations were consistent with paragraph breaks in text for both groups.

C. *Chunking and Planning Unit Size*

Table II shows the size of each prosodic unit layer by number of syllables and words. Combining these results with those given in Table I, we found that L2 speech planning not only exhibits more B3s, but that those B3s also contain fewer syllables. L1 and L2 speakers' planning strategy appears to differ with respect to the use of intermediate-level chunking units, which suggests that L1 speakers are able to plan on a larger scale than L2 speakers at every prosodic layer (with the exception of BG). Section D will offer some possible explanations for why BG was found to be larger in L2 than L1.

TABLE III. SIZE OF CHUNKING AND PLANNING UNITS BY PROSODIC LAYERS, SPEAKER GROUPS AND UNIT OF MEASUREMENT.

Measurement &Group Prosodic units	*Syl Num*		*Word Num*	
	L1	*L2*	*L1*	*L2*
PW (μ /σ)	3.5 /1	3 / 1	2.7/ 0.9	2.5 / 0.8
PPh (μ /σ)	8.3 / 4	5 / 3	6.4 / 3.5	4.2 / 2
BG (μ /σ)	18 / 7	21 / 8	14.1 / 5	16.7 / 6
PG (μ /σ)	38 / 7	38 / 11	30 / 6	30 / 10

D. *Consistency of Discourse Planning in Text*

Overlap of B4 location was measured to investigate within- and between-group consistency of discourse planning in text (See Figure 1. Four B4 positions have a high level of consistency among L1 speakers; 9 to 10 L1 speakers show agreement on those B4 locations. L2 speakers' B4 locations demonstrate a much higher level of variation, and their patterns are different from those of L1 speakers. For example, at the first B4 position, which exhibits a high level of consistency for L1 (position index=50), only 4 out of 9 L2 speakers produced B4. It seems that L1 speakers have a high level of agreement on the planning structure for a fixed text, but L2 speakers do not. Possible explanations will be discussed in Section VII.

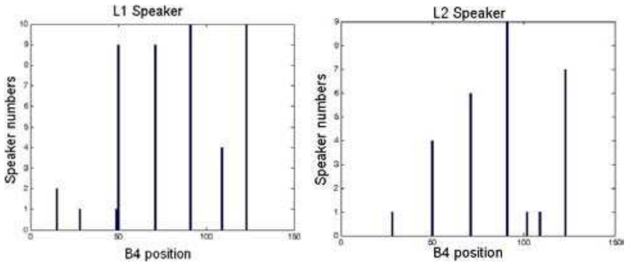


Figure 1. Distribution of B4 postion by speaker group.

V. ACOUSTIC ANALYSIS OF PROSODIC BOUNDARIES

A. Analysis of Pause Duration

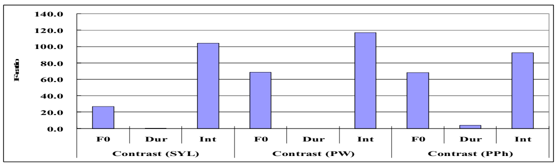
Table III shows the means/standard deviations of pause duration by speaker group and prosodic break strength. Consistent with our previous studies of Mandarin data [8], results show that pause duration is a feature also used to discriminate B3, B4 and B5 in English. In our previous studies, variation of pause duration at B3 in Mandarin was found to be greater than the variation found at the B4/B5 levels [8]. Even though pause duration at B3 was highly variable, transcribers could still perceive B3 consistently, which suggests that acoustic cues other than pause are more salient perceptually in differentiation of boundary strength [16]. Subsequent analysis showed that boundary neighborhood features make up contrast patterns which improved discrimination of boundary break levels [16]. These contrast patterns can compensate for variation in the duration of pauses, or even for the lack of a pause at every prosodic level. In Section V, we will examine the contrast features which contribute to the perception of a break in order to investigate the relative acoustic prominence of each feature which serves to distinguish B3, B4 and B5.

TABLE IV. PAUSE DURATION (MS) BY BREAK SIZE AND SPEAKER GROUP

Break ***Speaker***	***B3***	***B4***	***B5***
L1 (μ/σ)	91/135	533/189	762/173
L2 (μ/σ)	167/243	550/180	710/272

B. Boundary Discrimination among B3, B4 and B5

F-ratios discriminating B3, B4 and B5 by speaker group, prosodic unit and acoustic correlates are summarized in Figure 2. Overall patterns are similar across L1 and L2 speaker groups, which can explain why the same HPG units can be perceived by transcribers of both L1 and L2 data. Results indicate that (1) the degree of distinction (F-ratio) between break levels is higher for L1 speakers at all levels, (2) PW is the level at which the strongest distinctions among B3, B4 and B5 in English can be observed and (3) intensity is the acoustic feature used most extensively in English to distinguish B3, B4 and B5.



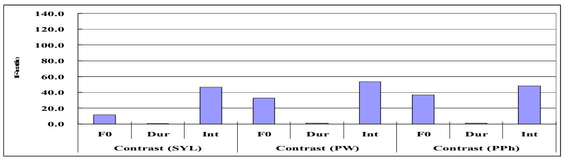


Figure 2. F-ratio distinctions of B3, B4 and B5 by acoustic features.

Contrastive feature means are summarized by speaker group, prosodic break, acoustic correlates and scale of feature extraction in Table VI. The means of F0 contrast and intensity contrast are ordered B5>B4>B3. In addition, the scale from B3, B4 to B5 in terms of F0 and intensity contrast is much larger than the scale of duration contrast. Thus, it seems reasonable to infer that F0 and intensity contrast between B3, B4 and B5 are much more salient cues to prosodic break level than duration contrast. However, it must be noted that duration patterns were calculated and extracted based on a Taiwan Mandarin syllable-timed template, so the effect of lexical stress and English stress timing were not incorporated into this analysis. Future studies will be designed to address this aspect of the data.

TABLE V. CONTRASTIVE FEATURE MEANS BY SPEAKER GROUP, BREAK SIZE, ACOUSTIC CORRELATES AND SCALE OF FEATURE EXTRACTION

Scale&Feature		***Contrast (SYL)***			***Contrast (PW)***			***Contrast (PPh)***		
Group &Break		***F0***	***Dur***	***Int***	***F0***	***Dur***	***Int***	***F0***	***Dur***	***Int***
L1	***B3***	0.31	-1.71	0.06	0.02	-0.76	0.02	-0.38	-0.11	-0.22
	B4	0.98	-1.52	0.50	0.61	-0.63	0.49	0.13	0.20	0.15
	B5	1.77	-1.76	1.48	1.86	-0.78	1.07	0.91	0.08	0.46
L2	***B3***	0.16	-1.18	-0.06	0.04	-0.34	-0.04	-0.16	-0.02	-0.09
	B4	0.82	-1.08	0.35	0.57	-0.35	0.34	0.05	0.25	0.18
	B5	1.24	-1.58	0.91	1.42	-0.70	0.62	0.99	-0.11	0.41

VI. ACOUSTIC FEATURE PATTERNS BY PROSODIC LAYER

A. F0 Domain

The F0 patterns derived after removing intrinsic vowel effect for each speaker group and prosodic layer are shown in Figure 3. Down-stepping can be observed at both the PPh and PG layers, and it is at these levels that we find the major differences between L1 and L2. The patterns in both prosodic layers have a larger range in L1 than L2, especially at the PW layer. Future work will investigate the relationship of this feature to differences in overall pitch range between L1 and L2 speakers.

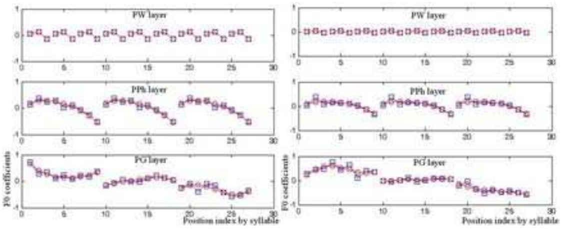


Figure 3. F0 patterns by speaker group and prosodic layer. The curve by blue squares denotes original prediction, while the curve by red diamonds denotes predictions after the polymonial fitting.

B. Temporal Domain

Figure 4 presents duration patterns derived after removing intrinsic vowel effects by speaker group and prosodic layer.

Differences between L1 and L2 speakers are observed only at the PPh layer: L1 speakers produce more pronounced final lengthening at the PPh layer than L2 speakers do.

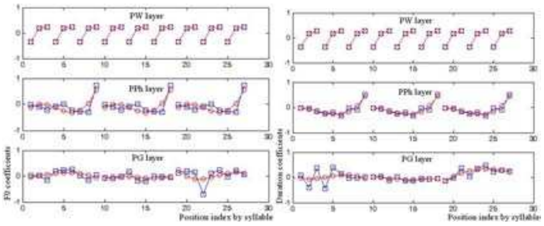


Figure 4. Duration patterns by speaker group and prosodic layer. The curve by blue squares denotes original prediction, while the curve by red diamonds denotes predictions after the polymonial fitting.

C. Intensity Domain

Figure 5 shows intensity patterns derived after removing intrinsic vowel effects by speaker group and prosodic layer. In this domain, little difference was found between L1 and L2 speaker groups.

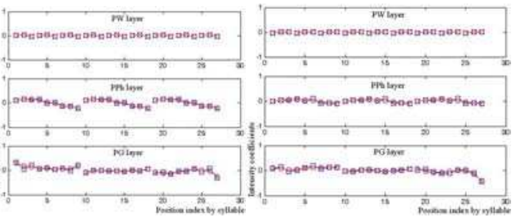


Figure 5. Intensity patterns by speaker group and prosodic layer. The curve by blue squares denotes original prediction, while the curve by red diamonds denotes predictions after the polymonial fitting.

VII. Discussion

A. Break distribution and planning scale

HPG analysis revealed that acoustic signals to perceived breaks are comprised of information from multiple prosodic layers. It was found that 1) L2 speakers produce more and shorter PPhs (B3) than L1 speakers do and 2) L2 speakers exhibit a larger range of variation with respect to the physio-linguistic unit BG (B4). L1 speakers also produced more consistent within-group B4 locations than L2 speakers did. The results indicate that L2 speakers set the look-ahead threshold by the sentence level, following punctuation marks rigidly. When the sentence structure is more complex, it was then divided into smaller phrase units. Large-scale discourse planning is not evident. On the contrary, L1 speakers are fully capable to set the look-ahead threshold above the sentence level, simultaneously planning both discourse and sentence units as they read, and as a result disregard punctuation marks in order to express discourse structure. In short, the speech planning scale is rather stiff to L2 speakers, while it is much more flexible to L1 counterparts.

B. Acoustic Analysis of Prosodic Boundaries

The results presented in Section V suggest that the F-ratio patterns used to distinguish B3, B4 and B5 prosodic break levels are similar across L1 and L2 speaker groups, although the acoustic cues produced by L1 speakers are more acoustically robust. It was precisely because the combination of acoustic contrasts used to mark prosodic breaks was similar for both groups that transcribers were able to use a consistent set of criteria to parse L1 and L2 prosodic breaks, which provides evidence for the cross-linguistic perceptual validity of the prosodic boundary categories represented in the HPG framework.

C. Acoustic Feature Patterns Across Prosodic Layers

Derived acoustic patterns of English discourse by prosodic layer were given in Section IV. While general configurations and patterns of acoustic cues appear to be similar across speaker groups, the extent to which those cues are realized has been shown to differ, particularly with respect to production of F0 range at the PW and PPh layers and production of final lengthening. L1 speakers exhibit a larger pitch range than L2 speakers in their production of PW and PPh down-stepping, and L1 speakers produce a greater degree of final lengthening. Between-group intensity differences, in contrast, were negligible at all tested levels.

VIII. Conclusion

We have uncovered both similarities and differences between L1 and L2 speech planning of English discourse using HPG to tease apart the multiple levels of prosodic contributions. The data presented here suggest that many of the perceived differences between L1 and L2 speech at the discourse level can be attributed to differences in prosodic break level distribution and location which correspond to how listeners perceive information chunks across speech flow, rather than differences in use of individual acoustic cues to signal prosodic breaks. In particular, the differences reflect L2 speakers' relatively smaller scope of speech chunking and planning. Future work will explore between-group difference of intermediate phrases to further reveal a more comprehensive picture of discourse planning.

IX. References

[1] Warren, P., Elgort, I., and Crabbe, D. 2009. Comprehensibility and prosody ratings for pronunciation software development. *Language Learning & Technology*, 13(3), 87-102.

[2] Derwing, T. M., & Munro, M. J. 1997. Accent, intelligibility, and comprehensibility: Evidence from four L1s. Studies in Second Language Acquisition, 19, 1-16.

[3] Pickering, L. 2004. The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. English for Specific Purposes , 23, 19-43.

[4] Wennerstrom, A. 1998. Intonation as cohesion in academic discourse. Studies in Second Language Acquisition, 20, pp 1-25.

[5] Tseng, Chiu-yu, Pin, Shao-huang, Lee, Yeh-lin, Wang, Hsin-min and Chen Yong-cheng. 2005. Fluent speech prosody: framework and modeling. *Speech Communication, Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation*, 46(3-4): 284-30.

[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. and Fu, B. 2005. Duration, intensity and pause predictions in relation to prosody organization. *Proceedings of the Interspeech 2005*, (September 4-8, 2005), Lisbon, Portugal, pp. 1405-1408.

[9] Riggenbach, H. 1991. Towards an understanding of fluency: A microanalysis of nonnative speaker conversation. *Discourse Processes* 14, 423–441.

[10] Dewaele, J.-M., 2000. Saisir l'insaisissable? Les mesures de longueur d'e´ nonce´ s en linguistique applique´ e. *International Review of Applied Linguistics* 38, 31–47.

[11] Dewaele, J.-M., Pavlenko, A., 2003. Productivity and lexical diversity in native and non-native speech: A study of cross-cultural effects. In: Cook, V. (Ed.), The effects of the second language on the first. *Multilingual Matters*, Clevedon, pp. 120–141.

[12] Vanderplank, R., 1993. Pacing and spacing as predictors of difficulty in speaking and understanding English. *English Language Teaching Journal* 47, 117–125.

[13] Judit Kormos, Mariann Denes, Exploring measures and perceptions of fluencyi n the speech of second language learners, *System, Volume 32, Issue 2*, (June 2004), Pages 145-164.

[14] 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*, (April 18-20, 2008), Beijing, China. 6 pages.

[15] Tseng, Chiu-yu and Su, Zhao-yu 2008. Discourse Prosody and Context – Global F0 and Tempo Modulations. *Interspeech 2008*, (Sep. 22-26, 2008), Brisbane, Australia. 1200-1203.

[16] Tseng, C. and Chang, C. 2008. Pause or no pause? –Prosodic phrase boundaries revisited. *Tsinghua Science and Technology*, 13.4: 500-509.