Discourse Prosody Planning in L1 and L2 English

Tanya Visceglia[1], Chiu-yu Tseng[2], Zhao-yu Su[2] and Chi-Feng Huang[2]
1. *Department of Applied English, Ming Chuan University, Taipei, Taiwan*
2. *Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei, Taiwan*
orlandotaipei@hotmail.com, cytling,morison,chifeng@sinica.edu.tw,

Abstract

L1 English and L1 Taiwan Mandarin discourse-length English speech data extracted from the TWNAESOP corpus was analyzed using a perceptually-based hierarchy of prosodic phrase group (HPG) framework in order to investigate similarities and differences in the organization of discourse-level speech planning in English across L1 (native) and L2 (non-native) speaker groups. While both groups appear to produce similar configurations of acoustic contrasts to signal discourse units and boundaries, L1 speakers were found to produce these cues more robustly. Between-group differences in discourse units were also found through the distribution of prosodic break levels and break locations. These findings can be attributed to the size and scope of speech planning and chunking, whereby L2 speakers, possibly due to on-line processing limitations in L2, use more intermediate chunking units and fewer larger-scale planning units in prosodic discourse organization. Future cross-L1 comparisons will investigate whether these differences represent L2-universal processing limitations and strategies.

1. Introduction

Generations of contact between English and other language groups represented in Asia has given rise to a rich variety of Asian Englishes. English is also being studied and spoken as a second language in more countries in Asia than ever before. Thus, understanding the range of variation in English spoken in the world today is a fundamental issue for the development of English language education as well as speech science and technology. The Asian English Speech cOrpus Project (AESOP), a multi-national research effort, was designed to collect and compare English speech corpora from as many Asian countries as possible, in order to derive a set of core properties common to all varieties of Asian English, as well as to discover features that are particular to individual dialects [1]. The data presented here represent part of the ongoing research conducted by the TWNAESOP research team, which was formed to develop a systematic understanding of the acoustic characteristics present in L2 Taiwan English speech.

It should be emphasized here that the major research goal of AESOP is not to normalize Asian Englishes to any particular ENL (English as a Native Language) standard, but instead to catalog and predict similarities and differences among the varieties of English found across Asia. Asian L2 speaker populations have grown to outnumber ENL speakers, and the majority of English speakers in the world today are either ESL (English as a Second Language) or EFL (English as a Foreign Language) speakers engaged in communication with other ESL or EFL speakers. This suggests the need for an international and flexible set of phonological standards, rather than a single rigidly defined ENL norm [2]. It is hoped that our collective findings will contribute to the further development of English speech tools and interfaces such that these applications can be better tailored to accommodate Asian users.

Moreover, current research has refuted the idea that L2 speech necessarily becomes less intelligible as a result of being different from native pronunciation. Many studies have demonstrated weak or no correlation between global accent ratings and level of overall intelligibility [3]. Thus, our analysis of Taiwan English pronunciation is not as much concerned with accentedness, defined as how different a speaker's pronunciation is perceived to be from that of the L1 community, as it is with intelligibility, defined as how well the speaker's intended message is understood, and comprehensibility, defined as perceived level of difficulty in following the speaker's intended meaning [4].

In addition to the segmental, lexical and utterance-level features demonstrated to correlate with comprehensibility in L2 speech, realization of the prosodic cues to discourse structure and information sequencing in continuous speech have also been found to correlate with intelligibility. A positive correlation was found between L2 speakers' use of intonation to

signal topic change (paratone) and their overall score on the SPEAK test of oral English proficiency [5]. And while L1 English speakers produce a hierarchical system of prosodic units to create semantic cohesion within spoken paragraphs and to signal aspects of information structure, there is evidence that L1 Mandarin International Teaching Assistants are much less adept at using prosodic cues in this way, even when they are necessary to clarify aspects of sequencing and emphasis [6].

The abovementioned studies measure prosodic differences between L1 and L2 speech by tracking F0 movement within a speaker's register in and across intonational paragraphs, which are delimited solely in terms of F0 downstepping and reset. Their data are extracted from unscripted speech of varying lengths, which better allows for observation of L2 spontaneous speech characteristics, but fails to control for discourse length and information structure. In contrast, the set of materials analyzed in this study consists of a text of fixed length and content, which allows for more precise observation of between- and within-group similarities and differences in speech chunking and planning strategies. This design also has the advantage of facilitating both the data collection process and automatization of data segmentation, which has allowed us collect data from a much larger L2 speaker group than those represented in previous studies. Moreover, our data analysis uses a perception-based hierarchical discourse prosody framework HPG (Hierarchy of Prosodic Phrase Group), which is designed to tease apart the layering of not only F0, but also duration and amplitude cues present in discourse prosody, traces of which can found in lower- and higher-level units of prosodic representation [7].

2. Materials and Procedure

These data are drawn from a subset of the core phonetic experimental tasks developed by AESOP. The AESOP materials, which include sets of both read and spontaneous speech tasks, as well as a recording platform and recording protocol manual specifically designed for this project [8] were developed in a collaborative effort by AESOP teams in Taiwan, Japan and Hong Kong. The current data set consists of only one read speech task; speakers were instructed to read Aesop's fable "The North Wind and the Sun" aloud at a natural speech rate and volume. This passage is recommended by the IPA for the purpose of eliciting all phonemic contrasts 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 North American English speakers and 514 L1 Taiwan Mandarin speakers. Pre-processing of recorded data included automatic annotation of segmental labeling using the HTK toolkit in the CMU dictionary [9], followed by spot-checking of segmental labeling by experienced transcribers to ensure precise alignment of phone boundaries. Manual labels of perceived prosodic boundaries (B2, B3, B4 and B5) were also labeled by trained transcribers using HPG protocol.

3. Data Analysis

Following phonetic segmentation and perceptual labeling of discourse boundaries, the HPG framework was applied to compare L1 and L2 strategies for prosodic organization at the discourse level. 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 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 (no identifiable pause), B2/PG (perceived slight tone of voice change follows), B3/PPh (clearly perceived pause), B4/BG (clearly perceived change of breath) and B5/PG (final lengthening, complete stop before new paragraph, change of breath) [10].

Speech rate, break distribution and planning scale were analyzed by applying HPG to L1 and L2 productions of the same English discourse. Overlap of B4 (breath group) position between groups was also measured, 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. The multiple regression model used for Mandarin discourse in our previous work was modified to reflect the English vowel inventory [7]. 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 and the syllable position corresponding to each respective prosodic layer. As our discourse contained an insufficient number of phonotactic combinations to adequately 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. The following polynomial curve fitting formula (order n = set 3) was also used to generate more robust patterns of F0, duration and amplitude.

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

4. Results

4.1 Speech Rate

Table I shows speech rate comparisons of L1 and L2 groups calculated in three ways: syllable number per minute, word number per minute, and number of stressed syllables per minute, measurement techniques which have all been used in previous studies to estimate speaker fluency [11]. However, any method of calculation that employs means and averages cannot capture the internal dynamics present in the flow of continuous speech, which may account for our otherwise somewhat puzzling finding 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 [12].

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

4.2 Prosodic Break Distribution

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 I. BREAK DISTRIBUTION BY SPEAKER GROUP

Break / Speaker	L1	L2
B2	66%	50%
B3	19%	39%
B4	8%	5%
B5	7%	6%

Table II shows that the size of PW is larger for L2 speakers than it is for L1 speakers, but the most substantial difference between speaker groups is in the size of PPh and BG. L2 PPhs contain fewer syllables than their L1 counterparts. L2 BG seems to be larger than L1 BG, but it also exhibits a larger range of variation, which can be attributed 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.

4.3 Chunking and Planning Unit Size

Table III 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 strategies appear 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.

Table III: Size of chunking and planning units by prosodic layer, speaker group 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

4.4 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, only 4 out of 9 L2 speakers produced B4. Thus, it seems that L1 speakers have a high level of agreement on the planning structure for a fixed text, but L2 speakers do not.

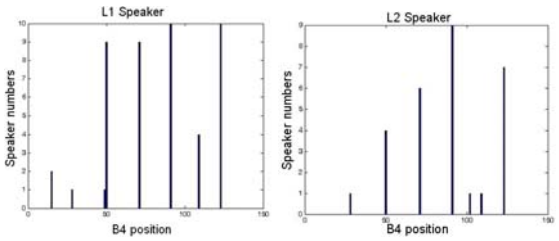

Figure 1. Distribution of B4 postion by speaker group.

4.5 Analysis of prosodic boundaries

4.5.1 Analysis of Pause Duration

Table IV shows the means/standard deviations of pause duration by speaker group and prosodic break strength. Consistent with our previous Mandarin data, results show that pause duration is a feature also used in English to discriminate B3, B4 and B5. 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. However, 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 perceptually salient in differentiation of boundary strength. Subsequent analysis showed boundary neighborhood features forming contrast patterns, which improves discrimination of boundary break levels. These contrast patterns can compensate for variation in the duration of pauses, or even for the lack of a pause at every prosodic level.

TABLE IV: PAUSE DURATION (MS) BY BREAK SIZE /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

4.5.1 Boundary Discrimination: 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 in both L1 and L2 speech. 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 neighborhood contrasts among B3, B4 and B5 can be observed and (3) amplitude is the acoustic feature used by both speaker groups most extensively to distinguish among break size categories B3, B4 and B5, although amplitude has received much less attention than duration and F0 in the literature.

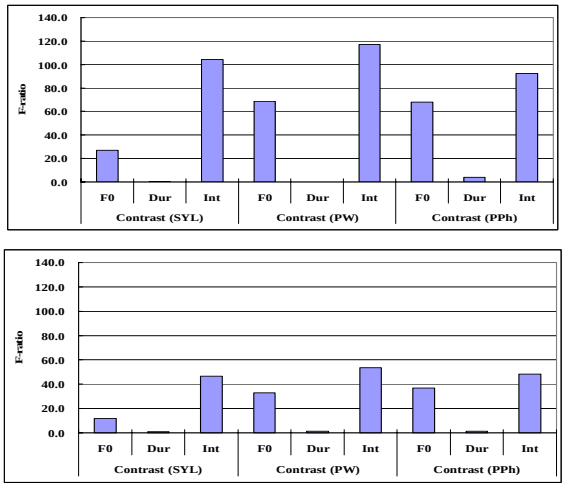

Figure 2. F-ratio distinctions of B3, B4 and B5 by acoustic features.

Table V summarizes contrastive feature means by speaker group, prosodic break, acoustic correlates and scale of feature extraction. Means of F0 contrast and intensity contrast are ordered B5>B4>B3. In addition, the scale from B3 to B5 in terms of F0 and intensity contrast is much larger than the scale of duration contrast. It seems that F0 and intensity contrast among B3, B4 and B5 are much more salient cues to prosodic break level than duration contrast. However, duration patterns were calculated and extracted based on a Taiwan Mandarin syllable-timed template, so effects of lexical stress and English stress timing were not incorporated into this analysis. Templates will be adjusted in future work to address this factor.

TABLE 5: CONTRASTIVE FEATURE MEANS BY SPEAKER GROUP, BREAK SIZE, ACOUSTIC CORRELATES AND SCALE OF FEATURE EXTRACTION

Scale&Feature / *Group &Break*		*Contrast (SYL)*			*Contrast (PW)*			*Contrast (PPh)*		
		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

4.6 Feature patterns by prosodic layer

4.6.1 F0 Domain

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. 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 between-group differences in overall pitch range.

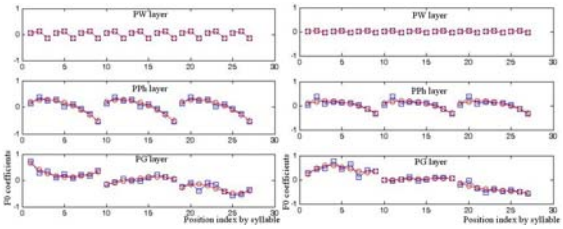

Figure 3. F0 patterns by speaker group and prosodic layer.

4.6.2 Temporal Domain

Figure 4 presents duration patterns derived after removing intrinsic vowel effects by speaker group and prosodic layer. Differences between L1 and L2 speaker groups are observed only at the PPh layer: L1 speakers produce more pronounced final lengthening at the PPh layer than L2 speakers do.

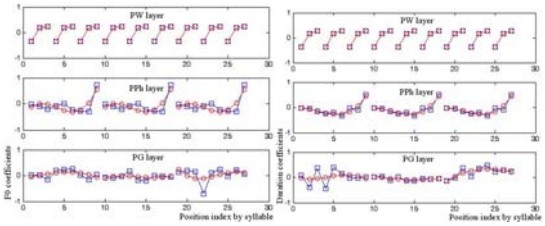

Figure 4. Duration patterns by speaker group and prosodic layer

4.6.3 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.

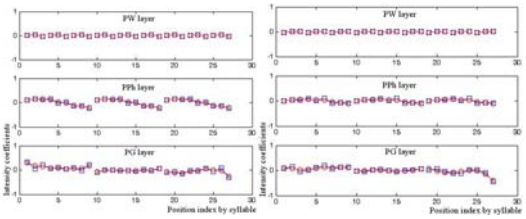

Figure 5. Intensity patterns by speaker group and prosodic layer

5. Discussion

5.1 Break size and distribution

The analyses given in Section 4 reveal that L2 speakers produce more and shorter prosodic phrases than L1 speakers do. More extensive use of intermediate-level boundaries by L2 speakers suggests that the difficulty presented by large-scale discourse planning causes them to divide a discourse into smaller, more manageable units. This is consistent with previous findings that L1 Greek, Spanish and Korean speakers tend to divide English utterances into shorter units than L1 English speakers do [13]. Thus, smaller units may be a universal L2 strategy for reducing the processing load of online speech production. It is interesting to note that speech divided into smaller chunks tends to sound slower, even when acoustic analysis disconfirms that perception. Smaller chunks in L2 may also explain L2 speakers' inconsistent placement of B4 boundaries; speakers are stealing smaller breaths with each B3.

5.2 Boundaries and Features across Prosodic Layers

F-ratio patterns used to distinguish B3, B4 and B5 prosodic break levels are similar across L1 and L2 speaker groups, although the cues produced by L1 speakers are more acoustically robust. These results are consistent with our findings on L2 production of the acoustic correlates of English lexical stress [14]: L1 and L2 speakers use the same cues, but L1 speakers are able to use those cues more effectively, and are able to maintain lexical contrasts even when those contrasts interact with higher-level prosodic contexts, whereas L2 speakers are not able to do so.

When acoustic cues are teased apart to determine the contribution made by each cue at each prosodic layer, a similar data pattern emerges. While general configurations of acoustic cues appear to be similar across speaker groups, the extent to which those cues are realized differs, 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.

Previous studies have suggested that the absence of clear pitch sequence structuring in L2 English discourse prosody is exacerbated by the L2 speakers' overall narrower pitch range [6], which is

also quite possibly a L2-universal phenomenon. L2 Mandarin studies have proposed that many L1 English speakers' lexical tone errors in Mandarin can be attributed to their reduced pitch range in L2. [15]. A bi-directional study of L2 Mandarin and L2 English, which found reduced pitch ranges in L2 for both speaker groups, also found that L2 speakers failed to perform discourse-level initial pitch setting in production of short dialogues [16]. Future work will compare the L2 English of L1 Taiwan Mandarin and L1 Putonghua. The pitch range of the former has been observed to be narrower than that of the latter, so it is possible that cross-linguistic differences in pitch range may contribute to our L1 Taiwan Mandarin findings. Lastly, our finding that intensity provides the strongest cue to boundary strength suggests that acoustic cues to prosody are produced within a system of trading relations. This underscores the importance of using a multi-layered data analysis approach, which can examine acoustic cues both separately and in combination.

6. Conclusion

Differences in the distribution and location of prosodic break levels, as well as differences in the acoustic robustness of the cues used to signal prosodic breaks, appear to represent the largest sources of variation between L1 English and L1 Mandarin realization of English discourse prosody. We attribute the differences between speaker groups found in this study to L2 speakers' relatively smaller scope of speech chunking and planning, to L2 speakers' smaller F0 range, and to the increased level of challenge L2 speakers face in embedding multiple levels of prosodic information in production of discourse-length units of speech.

Further analyses of the same data set will investigate other possible differences in prosodic realization of different levels of information sequencing and structure, such as use of prosodic highlighting to signal prominence and transition, and prosodic embedding of grammatical structure and dependency. Future work will also mine the multi-L1 AESOP database for cross-L1 comparisons in order to investigate the question of whether the between-group differences found in this study reflect universal L2 planning strategies and processing constraints.

7. References

[1] Meng, H., Tseng, C., Kondo, M., Harrison, A. and Visceglia, T., "Studying L2 Suprasegmental Features in Asian Englishes: A Position Paper" Interspeech 2009 1715-1718. Brighton.

[2] Jenkins, J. 2002. A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics*, 23(1), 83–103.

[3] Munro, M. J. 2008. Foreign accent and speech intelligibility. In Hansen Edwards, J. G. & Zampini, M. L. (Eds.). *Phonology and Second Language Acquisition* (pp. 193-218). Amsterdam: John Benjamins.

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

[5] Wennerstrom, A. 1998. "Intonation as cohesion in academic discourse: a study of Chinese speakers of English" *Studies in Second Language Acquisition* (20), 1-25.

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

[7] Tseng, C. Pin, S., Lee, Y., Wang, Hs. and Chen Y. 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.

[8] ftp://140.109.150.30

[9] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[10] Tseng, Chiu-yu, Cheng, Yun-Ching and Chang, Chun-Hsiang. 2005. Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech. Oriental COCOSDA 2005 23-28. Jakarta, Indonesia

[11] Judit Kormos, Mariann Denes, Exploring measures and perceptions of fluency in the speech of second language earners, System, Vol. 32, Issue 2, June 2004, 145-164.

[12] 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.

[13] Hewings, M. 1995. "Tone choice in the English intonation of non-native speakers." International Review of Applied Linguistics, 33(3), 251-265.

[14] Visceglia, T., Tseng, C, Su, Z. and Huang, C. "Interaction of Lexical and Sentence Prosody in Taiwan L2 English." To be presented at the SLaTE Workshop, Interspeech 2010. Tokyo.

[15] Shen, X. 1989. "Toward a register approach to teaching Mandarin tones." Journal of the Chinese Language Teachers Assoc. (24), 27-47.

[16] Visceglia, T. & Fodor, J.D. 2006. "Fundamental frequency in Mandarin and English: Comparing first- and second-language speakers". In Interfaces in Multilingualism, Lleó, Conxita (ed.), 27–59.