Global F0 Features of Mandarin L2 English - Reflection of Higher Level Planning Difficulties from Discourse Association and Information Structure

Chao-yu Su [1, 2, 3] *& Chiu-yu Tseng* [1]

[1] Institute of Linguistic, Academia Sinica, Taiwan
[2] Taiwan International Graduate Program, Academia Sinica, Taiwan
[3] Institute of Information Systems and Application, National Tsing Hua University, Taiwan
cytling@sinica.edu.tw

Abstract

It has been reported that one major feature of global prosody in continuous speech is to express cross-phrase association and cohesion through adjustment of individual phrase intonations. Another major feature of global prosody that also requires phrase intonation to adjust is to express information structure. Both features involve multi-level large scale speech planning; their interactions multifaceted. These higher level prosodic expressions are systematic and predictable in L1 speech, but these expressions are realized in L2 speech that can be attributed to foreign accent remains unknown. The current acoustic study thus attempts to address these two issues through corpus analysis with focus on normalization procedures that remove respective interactions in order to derive pattern that better reflect each feature involved. The global F0 constitution of L1 vs. Mandarin L2 English is compared to see the difference. As expected, respective positive correlations and patterns can be derived from L1 speech whereas divergent patterns found in L2 speech provide explanations of L2 accent. Overall cumulative effects of interaction further account for how both features collectively contribute to accent specific to Mandarin L2 English. These findings are readily applicable to CALL that targets global prosody training.

1. Introduction

In addition to second/foreign language education of English, the current world lingual franca, studies of foreign accent grow more in recent years due to their applications to language identification, speech recognition and CALL (Computer Aided language Learning). Topics related to foreign accent used to concentrate on linguistic specifications, known as canonical or bare forms that include segments, words (and individual phrase/sentence intonations at the phonetic/phonological, lexical and syntactic levels. However, more and more literature has demonstrated that the planning and production of continuous speech involves additional complex higher level planning that knowledge of bare forms alone could not account for; their direct reflection is prosody. The most notable issues are discourse prosody that expresses paragraph association and cohesion, and information structure that expresses information weighting [1, 2, 3, 4, 5, 6, 7]. Known as discourse intonation or discourse prosody, defined by a group patterned individual intonations, has been proposed for native English (L1) [1, 2, 3] specifying how successions of varied intonation patterns are schemed/organized comprehensively for overall communicative intention instead of arbitrarily or individually assigned.

Assuming that such discourse coherence of cohesion is not language specific, independent studies of L1 Mandarin speech examined corresponding prosodic patterns, taking care to separate lower-level contributions from segments [7, 8], and found similar patterns of how individual Mandarin phrase intonation adjust systematically to achieve coherence/cohesion. These cross-linguistic findings collectively demonstrate that individual phrase intonations are not independent isolated prosody units, but rather, constrained by paragraph association and subject to systematic modifications to yield global paragraph /discourse prosody. Needless to say, such prosodic features are also important in L2 speech learning and production.

Another important condition that contributes to intonation modification is information structure, planning and allocation. Evidence was reported in an F0 study on Cantonese using the command response model that separates the contribution of global unit intonation contour from local accentuation showed that in addition to phrase boundaries, extra insertions of phrase command which by the model's definition represents overall contour, are related to emphases of information focus [9]. In other words, information structure would further trigger intonation variations in addition to syntactic specifications. Similar findings of English intonation echo how variations can be better correlated with information structure (IS) in the context of the information structure and prosody interface [10].

It is no surprise that these higher-level specifications are much less understood in L2 speech. A previous study examined L2 English prosody variations by speech paragraph alone [11] and found how native speakers were able to match prosodic signals with discourse structural cues to mark the relationship within discourse

978-1-5090-3516-8/16/$31.00 ©2016 IEEE

units or between discourse units, while non-native speakers were unable to do the same. Another study compared L1 vs. Mandarin L2 English speech by discourse structure and information allocation [12] using the location of phrase commands by boundaries inside speech paragraph and perceived emphases. Results showed the following L2 features that marked their departure from The above L1 studies suggest clearly how these contributing factors of global prosody merits more attention L2 intonation and accent. L1: (1) less overlap of discourse boundaries but instead more insertions of phrase command at lower-level discourse unit resulting in smaller chunking units. (2) Less degree of F0 contrast of expressed emphases but more distinct F0 contours in non-emphases resulting in overall less degree of F0 contrast. These studies suggest clearly how these contributing factors of global prosody merits more attention L2 intonation and accent.

Following the same vein, the present study assumes that discourse associations and information structure involved higher-level planning, that is, in addition to lower-level linguistic specifications, and the required multi-level interaction, are two major reasons of L2 difficulty that would contribute significantly to L2 accent in the domain of global prosody. Our goals are thus two-fold. The first goal attempts to tease apart how L2's global F0 constitution differs from L1 by discourse structure and information planning while considering their interaction at the same time. The second goal is to further model L1's phrase commands by discourse structure and information planning jointly to test the predictability. Together, we hope the findings would provide better account of L2 global prosody. Methodologically, special attention is given to normalization procedures that would better reflect global F0 features and amplitude extraction of phrase command defined by command response model [13].

2. Speech Materials and Annotation

The speech materials are read speech of the story 'The North Wind and the Sun'. L1 speech is provided by 11 native American English speakers (5M/6F) whereas Mandarin L2 speech is provided by 30 Taiwan (TW) Mandarin speakers (15M/15F), coded N&S in AESOP-ILAS (Asian English Speech cOrpus Project— Institute of Linguistics Academia Sinica) corpus [14]. The passage contains 386 phoneme, 143 syllables, 114 words and 30 phrases.

2.1. Preprocessing

The three layers of preprocessing include segmental identification, lexical stress and focus status. Segmental identities are trained using force-alignment using the HTK Toolkit followed by manual spot-checking by trained transcribers. Word/lexical stress, i.e., primary, secondary and tertiary, is assigned to the syllable unit as specified in the CMU electronic dictionary. Focus status (narrow-, broad- and non-focus) is tagged in word unit by an American English native speaker and further aligned into corresponding phrase units. We assume focus status reflects information structure/planning which are used for contributing factors of prosody analysis and modeling.

2.1.1. Tagging discourse units by perceived boundaries and breaks across continuous speech

5 levels of prosodic units are used pin down the prosody of continuous speech, namely, the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath group (BG, a physio-linguistic unit constrained by change of breath while speaking continuously) and the multi-phrase speech paragraph PG. These units were manually tagged by 5 levels of perceived discourse boundaries B1 through B5 [15]; the default unit/boundary correlations can be expressed as SYL/B1, PW/B2, PPh/B3, BG/B4 and PG/B5.

3. Methods

3.1. Variables for modeling phrase command

Acoustic variables, namely intonations, are examined by higher-level linguistic specifications derived from discourse and information structure and compared between L1 and L2, the acoustic variables and higher-level linguistic specifications are defined in 3.1.1 and 3.1.2.

3.1.1. Acoustic variables

The major acoustic variable, namely intonations, are represented by phrase command defined by the command-response model [13]. The model, by definition, decomposes three contributing components as long-term/global tendency (phrase component), short-term/local humps (accent component) and a constant (base frequency). The 3 components are represented by (1). In other to separate the above 3 components, a previous method based on filter is adopted [16]. The present study adopts the phrase component to represent global F0 contour for analysis. The amplitude of phrase command are used for analysis and modeling in the present study.

$$F_0(\mathrm{t}) = \ln(F_b) + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}[G_a(t - T_{1j}) - G_a(t - T_{2j})] \quad (1)$$

where $G_p(t) = \alpha^2 t \exp(-\alpha t), for\ t \geq 0$

$$G_a(t) = min[1 - (1 + \beta t)\, exp(-\beta t), \gamma], for\ t \geq 0$$

$$\alpha = 3,\ \beta = 20$$

3.1.2. Higher-level linguistic specifications by discourse and information structure

Higher level specifications by discourse and information planning are directly derived by annotation in Tabel1. Based on an assumption that amount of information content in phrase may be related to phrase length, an informative feature, information density, is proposed for following analysis. The information density is defined by average amount of information content divided by PPh length and the aim is to quantify amount of information content while PPh length is normalized.

Table 1: *Higher-level linguistic specifications by discourse and information structure*

Information allocation	Focus number in phrase
	Narrow focus Number in phrase
	Narrow focus Position in phrase
	Information density in phrase
Discourse structure	Phrase length by word
	Pre break by discourse
	Post break by discourse
	Normalized position By PG
	Normalized position By BG
	Distance to pre Ap

3.1.3. Normalization

3.1.3.1 PG

By definition the largest discourse unit (or speech paragraph) PG is always a multi-phrase unit and hence would differ in the number of phrases within and its length. Thus PG is normalized by 6 positions. '1' and '6' represent the first and final phrases in a given PG; the rest of the phrases between '1' and '6' are normalized from 2 to 5.

3.1.3.2 BG

The majority of second largest discourse unit BG is also a multi-phrase unit that also differ in the number of phrases within and its length, Thus BG is also normalized, this time in 3 positions 'Initial', 'Medial' and 'Final'. 'Initial' and 'Final' represent the first and final phrases in a given BG; the rest of the phrases between 'Initial' and 'Final' are represented as 'Medial'. A small number of BG's only contain one phrase and is defined as 1-PPh BG and categorized accordingly.

3.1.3.3 Information density

To compare different amount of information content contained in a given PPh, information density is equally normalized into 5 levels '1' to '5'.

3.2. Modeling procedure for amplitude of phrase command

In the present study, the response variable is the amplitude of phrase command and the explanatory variables are higher level specifications by discourse, and information structure. In order to approximate the amplitude of phrase command by two separate higher (discourse and information) level specifications, three regression techniques are adopted, namely, multivariable linear regression, robust regression and neural network. Multivariable linear regression (MLR) approximates the relationship between a response variable and linear combination of explanatory variables [17]. Robust regression (RoFit) is an extension of multivariable linear regression; the derived model is less sensitive by outliers [18]. A feedforward neural network (FNN) is a modeling technique for approximating response variable by non-linear functions which contains sets of adaptive weights learned from explanatory variables [19]. The number of layers used is 30. The results from the three adopted regression methods will then be compared

4. Results

4.1. L1-L2 difference by discourse structure

4.1.1. L1-L2 difference of planning size by discourse unit

By definition the manually tagged perceived boundaries correspond to levels of discourse unit (*2.1.1.*). We derive L1's average of boundary levels larger than B3 and regard them as the set of L1 norm of discourse unit. Results of L1 and L2 are listed in Table 2. L1-L2 difference is found in higher-level boundaries of L1 norm (B4/B5) while L2 produce more intermediate prosodic boundaries B3. However, we note taht speaker variation occurs not only among L2 speakers, but also among L1.

Table 2: *L1-L2 difference by level of discourse boundaries*

L1	5	3	3	4	3	5	4	5	4	3
L2	5	3	3	3	3	5	3	4	4	3

4.1.1.1 Discussion

The results are consistent with previous studies that more intermediate boundaries in L2 speech than L1 speech [20], indicating that L2 speakers plan phrases instead of larger units. The results also imply that L2 speakers may pay more attention to individual phrase intonation than to global and association when compared with their L1 counterparts. In addition, less number of high-level boundaries also indicates that

boundaries marking the beginning and ending of speech paragraph PG may be produced by L2 speakers. In other words, overall L2 speakers produce less coherent discourse structure compared to L1 speakers.

4.1.2. L1-L2 difference by PG position

Figure 1 shows amplitude patterns of phrase command by PG position for L1 and L2. Note that the graphic pattern of L1 (in blue) shows an overall decline of phrase command from, indicating an overall tendency of global intonation (F0) declination across PG positions. This pattern is consistent with declining patterns of speech paragraph in reported findings of Mandarin [8]. However, the pattern of L2 (in red) speech is different from L1; the overall tendency hard to account for from discourse planning.

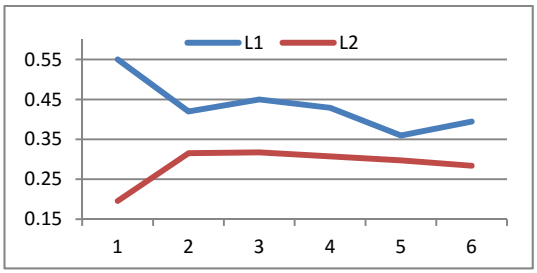

Figure 1: *Amplitude of phrase command by PG position for L1 and L2. Vertical axis - amplitude of phrase command, horizontal axis- normalized PG position*

4.1.2.1 Discussion

The results by PG position demonstrate how L1s' overall intonation pattern is signaled clearly by an overall global paragraph F0 declination from PG beginning toward end which can be taken as enhancing semantic association and cohesion. However, L2 speakers are not able execute similar pattern. Their form, distinct from the L1 norm, may impair overall association and cohesion on the one hand, and impede comprehension on the other hand.

4.1.3. L1-L2 difference by BG position

Figure 2 shows derived amplitude patterns of phrase command by BG position of L1 and L2. The patterns of L1 English (in blue and red) show significant F0 rising in BG-Initial position, followed by a sharp F0 dip to – Medial position, then to a slight F0 rise towards the – Final position. The 1-phrase PG (red only) shares the – Initial feature of higher F0. This pattern is consistent with reported findings from L1 Mandarin speech data [8], once again verifying cross-linguistic consistency at the higher level(s). However, the patterns of TW L2 speech (in green and purple) are not only reversed mirroring from the L1 norm, but are also lower in overall F0. The 1-phrase PG (purple only) is no exception.

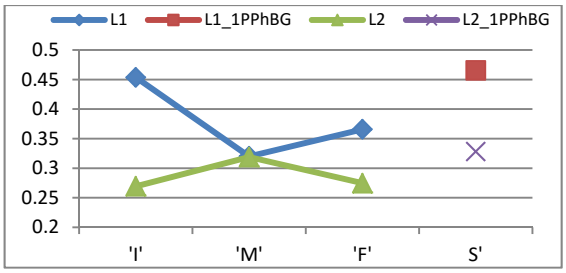

Figure 2: *Amplitude of phrase command by BG position for L1 and L2. Vertical axis - amplitude of phrase command, horizontal axis- normalized PG position*

4.1.3.1 Discussion

The results by three BG positions demonstrate how L1s' overall intonation pattern constrained by change of breath is also signaled clearly by an overall global F0 declination from beginning to end, confirming that similar patterns are employed by L1 at two consecutive layers of discourse planning. Once again, the patterns of L2 exhibit the opposite from the L1 norm. We therefore believe a second layer of deviation would further impair semantic association and cohesion.

4.2. L1-L2 difference by information allocation

4.2.1. L1-L2 information density by PPh

Figure 3 shows patterns of distribution of the phrase command by information density for L1 and L2 by intermediate phrase PPh. The graphic pattern of L1 (in blue) shows an overall ascending tendency, suggesting that information density increases across PPh. However, the pattern of L2 (in blue) shows irrelevant variation and hence no overall tendency of information density.

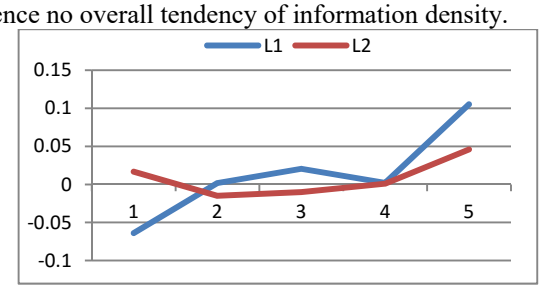

Figure 3: *Amplitude of phrase command by information density for L1 and L2. Vertical axis - amplitude of phrase command, horizontal axis- information density*

4.2.1.1 Discussion

The results by information density demonstrate that by L1's production of phrase command, the distribution and placement of information density follows a general tendency, increasing from the PPh beginning to end. This general tendency, consistent at the paragraph and its immediate lower-level levels, not only suggests that information placements follow a consistent pattern, but also provide reinforcing evidence essential to enhance

communication. Our L2 results, on the other hand, suggest that L2 speakers appear to be much less sensitive to information structure and content during speech planning; their speech output therefore lacks the contrasts that signal the information structure expressed through focus, less focus and no focus.

4.3. Modeling amplitude of phrase command for L1 by discourse and information structure

4.3.1. Prediction accuracy and contribution by discourse and information planning

The amplitude of phrase command is modeled by different regression methods and the error of root mean square by each method is listed in Table 3. It turns out that MLR performs the best. However, the difference among the 3 methods is not significant. Following the modeling, the contribution weight by two types of linear regression is further analyzed and listed in Table 4. The top 3 contributing weights 'Distance to pre Ap', 'Normalized position by BG' and 'Normalized position by PG' are identical across two types of regression MLR and RoFit. We note here also that 'information density', by default correlates most positively to information content is rank 5th.

Table 3: *Error of root mean square by regression type*

MLR	RoFit	FNN_30
0.181	0.182	0.201

Table 4: *Contribution weight by MLR and RoFit*

Higher level specification \ Regression		MLR	RoFit
Discourse structure	Phrase length by word	1.37	0.66
	Pre break by discourse	0.53	0.53
	Post break by discourse	0.10	0.09
	Normalized position by PG	2.43	3.63
	Normalized position by BG	6.17	7.11
	Distance to pre Ap	11.37	10.88
Information allocation	Focus number in phrase	-1.90	-1.20
	Narrow focus Number in phrase	0.00	0.05
	Narrow focus Position in phrase	0.88	0.45
	Information density in phrase	1.44	0.98

4.3.1.1 Discussion

The results substantiate how the global prosody of L1 English can be accounted for by intonation patterns correlating to discourse structure and information content. It also demonstrates by the F0 amplitude of how each current phrase command is related to its proceeding and following phrase command in similar ways. The results suggest that intonation planning is through a patterned global tendency by BG and PG instead of a linear succession isolated phrase intonation without higher level information. Specifically, higher level associations among speech paragraph and relative positions are significant contributing factors to overall intonation planning. Information density at rank 5th suggests information planning also contributes to intonation production. The overall cumulative effects of L1's positive correlation also provide account of the predictability of L1 global prosody.

5. Discussion

The above results reveal how the global F0 patterns of TW L2 English are different from L1 by both discourse structure and information planning. The discourse patterns show two distinct L2 features, i.e., (1) how L2 continuous speech lacks the cohesive pattern found in L1 and (2) why and in what way L2 prosody sounds flatter than L1. As for information structure, we found also through the production of phrase command in TW L2 speech that speakers are less sensitive to information content and hence less able to express weighted information content necessary to speech communication; their speech less expressive than the L1 norm. Global F0 is found systematic and closely related to discourse planning and intonation planning in native speech. The results suggest the L1's F0 planning is predictable by discourse structure and information planning.

6. Conclusions

Assuming discourse structure and information planning collaboratively contribute to L2 incomprehensibility, the present study examines how L2's global F0 constitution is different from L1 by discourse structure and information planning. By removing interaction between each level of specifications and modeling the amplitude of phrase command of L1, the global F0 patterns by discourse and information structure are extracted, and L1-L2 comparisons achieved. In so doing we have successfully teased apart the F0 constitution of multi-phrase discourse unit PG and BG and showed how and in what way TW L2 speech lacks the overall declining global prosody that signals cross-phrase association, resulting their less cohesive speech paragraphs when speaking continuously. We found further that L2 speech also lacks the overall tendency to increase information density across phrases, resulting in their prosodic expression of focal information less distinct when producing continuous speech. The above results combined help explain how discourse structure and information planning must be addressed in L2 speech. Future work will center on how global F0 constitution interacts with local F0 constitution and how to apply these findings to the design of CALL systems to improve L2 fluency and comprehensibility.

7. References

[1] Brazil, D., "Discourse intonation I", Birminghan:English Language Research Monograph, 1975.

[2] Brazil, D., "Discourse intonation II", Birminghan:English Language Research Monograph, 1978.

[3] Cormbie, W. "Process and relation in discourse and language learning", Oxford: Oxford University Press, 1985.

[4] Bailly, G., Holm, B., "SFC: a trainable prosodic model", Speech Communication 46: 348-364, 2005.

[5] Fujisaki, H., Wang, C., Ohno, S., Gu, W., "Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command–response model", Speech communication 47: 59-70, 2005.

[6] Xu, Y. "Speech melody as articulatorily implemented communicative functions", Speech Communication. 46, 220–251, 2005.

[7] Tseng, C. Y., Pin, S. H., Lee, Y. L., Wang, H. M. and Chen Y.C., "Fluent speech prosody: framework and modeling", Speech Communication, Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation 46(34): 284-309, 2005.

[8] Tseng, C. Y. and Su, C. Y., "Discourse Prosody and Context – Global F0 and Tempo Modulations", Interspeech 2008, 1200-1203. Brisbane, Australia, 2008.

[9] Gu, W.T, Hirose, K. and Fujisaki, H., "Modeling the Effects of Emphasis and Question on Fundamental Frequency Contours of Cantonese Utterances", IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, pp. 1155 -1170, 2006

[10] Dominguez, M., Farrús, M, and Wanner, L. "Combining acoustic and linguistic features in phrase-oriented prosody prediction", Speech Prosody 2016, 796-800, 2016.

[11] Levis, J. and Pickering, L., "Teaching intonation in discourse using speech visualization technology", System 32(4), 505–524, 2004.

[12] Su, C. Y., and Tseng, C. Y., "Melody of Mandarin L2 English—When L1 Transfer and L2 Planning Come Together", Oriental-COCCOSDA 2015 90-95. Shanghai, China, 2015.

[13] Hirose, K., Fujisaki, H. and Yamaguchi, M., "Synthesis by rule of voice fundamental frequency contours of spoken Japanese from linguistic information", IEEE, 1984.

[14] Visceglia, T., Tseng, C. Y., Kondo, M., Meng, H. and Sagisaki, Y. "Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project)", Oriental COCOSDA 2009. Beijing, China, 2009.

[15] Tseng, C. Y. "Beyond Sentence Prosody", Interspeech2010. Makuhari, Japan, 2010

[16] Mixdorff, H, "An Integrated Approach to Modeling German Prosody", Volume 25, Studientexte zur Sprachkommunikation, Dresden, 2002.

[17] Pedhazur, E J., "Multiple regression in behavioral research: Explanation and prediction", (2nd ed.). New York: Holt, Rinehart and Winston, 1982

[18] Andersen, R., "Modern Methods for Robust Regression", Sage University Paper Series on Quantitative Applications in the Social Sciences, 2008.

[19] Auer, P., Harald, B., Wolfgang, M., "A learning rule for very simple universal approximators consisting of a single layer of perceptrons", Neural Networks 21, 2008.

[20] Tseng, C. Y., Su, C. Y., Huang, C. F, and Visceglia, T., "An Initial Investigation of L1 and L2 Discourse Speech Planning in English", The 7th International Symposium on Chinese Spoken Language Processing (ISCSLP) 55-59. Tainan/Sun Moon Lake, Taiwan, 2010.