Why Is L2 Less Natural?—A Prosody Account

Chiu-yu Tseng

Institute of Linguistics, Academia Sinica, Taipei, Taiwan

cytling@sinica.edu.tw

Abstract

This paper is about what prosody is when speaking in units that are more than one sentence at a time, and in what way prosody makes speech sound natural and more intelligible. The discussion will focus on how in addition to linguistically defined prosody, it is mainly discourse associations, pragmatics and information structure that contributes to more expressive and natural sounding speech which is in fact more intelligible. Our approach to analyze and understand prosody phenomena is a relative one, hence degree of contrast and contrast patterns are used to test and illustrate while arguments are based on prosodic differentiation patterns that set L1 and L2 speech apart. Acoustic analysis of units and patterns of chunking and phrasing with boundary properties and pauses will be used to illuminate discourse associations; while perceived accentuation are analyzed by contrast degree as representation of keyword landmarks and speaker intention. Examples and evidence is derived using corpus linguistic approach and computational modeling. Issues presented are discourse associations and global prosody, discourse boundaries, stress patterns, focus and post-focus compression. Robust contrast is therefore necessary to make differentiation distinct. General L2 features, i.e., slower speaking rate, more intermediate units and pauses, highly varied intonation patterns, less distinct stress patterns and less post-focus compression, decrease the degree and of contrast robustness and makes L2 speech less expressive. Under-differentiated expressions are therefore a major reason of why L2 speech is less natural.

Keywords prosody, discourse association, focus, compression, stress, accentuation, pragmatics, speaker intention, information structure, contrast degree, prosodic expression, undifferentiation.

Introduction

The term prosody refers to the melody, rhythm and quantity of speech as reflected pitch, tempo and loudness patterns. Acoustic correlates to be analyzed are fundamental frequency (F0), duration and amplitude. However, acoustic analysis by lifting fragments from speech string and measuring them by phonological and/or syntactic units at face value is methodologically flawed because in realistic speech these purely linguistic units and structures representing abstract notions are laden with additional information from higher level discourse associations, pragmatics and speaker intention and largely reflected in prosody. In this paper, we will address prosody from its functions in speech and present analysis that reveals more facts of how we speak naturally.

In a nutshell, prosody is used to express three functions (1) purely linguistic information including lexical, phonological, semantic and syntactic; (2) mid-level pragmatic and information structure and (3) higher level discourse information and associations. In fact, in addition to the Anglo-American attentive syntax defined declarative/declining and interrogative/rising sentence intonation patterns [1] and the continuation-rise attentive complex utterance intonation [2, 3]; substantial variations occur to individual word stress and sentence intonation when they are produced in succession from when uttered in isolation. This goal of the this paper is to address why robust differentiation from discourse associations and speaker intended information is an intrinsic part of naturally occurring continuous speech, and how less robust realization of differentiation patterns makes L2 speech less expressive and henceforth less natural.

1. Framework and Methodology

1.1 Paragraph and Discourse Organization

Observations of fluent continuous speech revealed that speech paragraph is notably characterized by chunking, phrasing and associative patterns rather than individual sentence intonation. All of these features correspond to overall global planning of the speaker and are unmistakably perceived by the hearer. We have proposed a hierarchical prosody framework of discourse prosody from perceived chunking and phrasing units termed HPG (Hierarchy of Prosodic Phrase Group) [4, 5, 6] to account for discourse prosody. The HPG (see Figure 1 for a schematic representation) specifies both same-level adjacent sister relations and cross-over associations from larger-scale units, and at the same accommodates contributions from higher-level constraints. Thus by default phrases and sentences within a speech paragraph cannot be studied as unrelated units to the paragraph. The framework and units made identifications of layer-dependent prosodic contributions possible;

and at the same time accounted for contributing sources of overall prosody from different sized discourse unit. Not shown in Figure 1 are corresponding boundary breaks defined between each prosodic unit at the same level. The respective sizes of the HPG units and boundaries are SYL/B1>PW/B2>PPh/B3>BG/B4>PG/B.

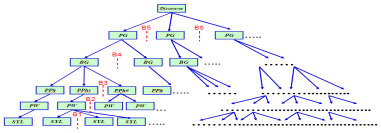

Figure1. A schematic representation of HPG (Hierarchy of Prosodic Phrase Group). The prosodic units from the lowest level are the syllable (SYL), the prosodic word (PW), the prosodic phrase (PPh), the breath group (BG) and the multiple phrase group (PG) or paragraph [4, 5, 6].

1.2 Methodology and Materials

Corpus and computational linguistics is adopted for investigations. Portions of preprocessed speech data from the two following corpora used (1) Sinica COSPRO (Mandarin Chinese Continuous Speech Prosody Corpus http://www.aclclp.org.tw/corp_c.php) 10.5GB 111-speakers Mandarin L1 read speech featuring both single- and multiple-speaker narrative pieces () and (2) AESOP-ILAS (Asian English Speech cOrpus Project, Institute of Linguistics, Academia Sinica) about 14GB/540-speaker L1 and L2 English speech featuring phnotactic, focus/prominence and discourse aspects [7]. Preprocessing includes automatic annotation (forced alignments of consonant and vowel identities with the HTK toolkit) followed by manual correction of spot-checked segmental alignments. At the prosody level, the 5 levels of HPG-specified perceived prosodic units and boundary breaks SYL, PW, PPh, BG and PG specified are manually tagged, using the Sinica COSPRO Toolkit [8]. Perception-based manual annotation accommodates the sometime non-overlap between perceived chunks and syntactic units [9], and multiple-phrase units made possible the departure of studying prosody by individual sentence intonation. Computational predictive modeling using a tailored linear regression model made possible a quantitative account of both the respective and cumulative prosodic contributions to the final overall output. In other words, in addition to prosodic contributions from specifications from lexicon, semantics, phonology and syntax, additional contributing factors from higher level discourse information and information structure cause lower level units to adjust and modify, as reflected in discourse associations, keyword prominence and speaker accentuation. In one study of the F0 composition of 4 speakers from COSPRO [10] we showed how at the syllables layer that correct prediction of Mandarin tone identities (lexical prosody) at the SYL level amounts to only 40~45%, indicating contribution of lexical prosody by the syllable to output prosody is less than half. However, by including additional contribution from the next higher level PW and linear neighborhood interaction as same-level contextual information of 15~20%, cumulative prediction accuracy is increased to 65%. Then by further including contributions from additional higher layers BG and PG of 7~35%, conservative cumulative predictions reached to 74~88%. It is therefore clear why additional F0 information from higher paragraph layers is crucial to the final output since patterns of global modulations are distinct [11].

Some major evidence illustrating global discourse associations and information structure are found to contribute to prosody. All of the evidences are phrase patterns with respect to HPG specified discourse positions PG PG-Initial, -Medial and –Final which denote paragraph topical initiation, continuation and termination of a speech paragraph.

2. Discourse Associations

2.1 F0 Down-stepping

Through the command-respond model [12] that decomposes the F0 of a speech section in semitones into the baseline, global contour Ap and local humps Aa, the model assumes that the F0 output is composed of a declining global trajectory indication intonation contour with modulations of local humps indicating accentuation. Hence the relative pitch height of both Ap and Aa can be derived. Averaged Ap values (phrase contours) of adjacent multi-phrase speech paragraphs by discourse positions reveal distinct but systematic high-to-low down-stepping across speaker and genre by discourse positions. The global pattern would not surface if we analyze sentences from a paragraph one at a time [13].

2.2 Tempo Adjustment

Pre-boundary tempo patterns are compared by the SYL, PW and PPh. Results found are similar to F0 findings. That is, systematic pre-boundary lengthening is only found at the PPh level and by the phrase, indicating how tempo adjustment is global [14]. The result echoes music composition where lengthening regarding overall phrasing structure by larger units is common.

However, analysis of two speech genres read speech vs. spontaneous university classroom lecture further reveals that tempo adjustment of phrases is genre conditioned. Phrase tempo of read speech by discourse position is a gradual fast-to-slow

modulation (PG-Initial<-Medial<-Final) whereas well-organized spontaneous university classroom lecture features a gradual normal-slow-fact pattern (PG-Initial<-Medial>-Final) [15].

3. Why Focus/Accentuation

3.1 Focus/Accentuation and Information Structure

We argue that the information structure is both semantics triggered and speaker intended. Weighting of information is reflected through prosodic variations of highlighted and compressed chunks in the speech string to denote the landmarks of key information. These accentuations are layered over discourse associations, making the outcome more expressive as shown in Figure 3.

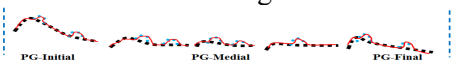

Figure3. A schematic representation of perceived pitch contours of a 5-phrase PG (Prosodic phrase group) with focus/accentuation. The black dashed lines-discourse association, the blue dotted lines-location of perceived prominence, the red solid lines-final output.

Perceived as prominence, focus and accentuation must be robust enough and clearly differentiable from the less important chunks.

3.2 Distribution of Focus by Genre

We define emphasis (E) in four degrees, i.e., reduced pitch, volume and segment contraction (E0), normal pitch, volume with no segmental contractions (E1), higher pitch or louder volume irrespective of speaker's tone of voice (E2) and higher pitch or louder volume marked by speaker's tone of voice (E4); and compared their respective distribution in three speech genres (1) read narratives CNA, (2) simulated weather broadcasting WB and (3) spontaneous university classroom lecture SpnL. Results show though similar distinctions of E/no-E are found regardless of genre, the ratio of E3/E4 distribution is SpnL (0.24), CAN (0.05) and WB (0.24), respectively. Furthermore, emphasis differs by genre, prosodic boundary type and discourse positions.

The results showed how allocation of focus/accentuation is systematic and borne by discourse associations to retain coherence, but distribution patterns diverge. Moreover, the E/no-E differentiation is robust and genre independent, but the E2/ E3 differentiation is genre dependent [16, 17].

3.3 Focus as an Additional Layer

We normalized contributions from emphases and found discourse structure remained distinct. Thus we prove how focus/accentuation can be accounted for as an additional layer over discourse organization [17, 18]. It became clear how both discourse coherence and information weighting contributes to the coherence and expressions of output prosody. The significance is how we could explain what causes prosodic modulations and why the seemingly highly varied modulations are in fact systematic.

4. Contrast and Differentiation

In the literature prosody related issues are more often studied in isolation and investigated separately, for example pitch height, pre-boundary lengthening, focus and post-focus compression. We argue, however, that they can be better accounted if studied from a relative perspective and understood by differentiation robustness. Findings from word stress and contrastive/narrow focus are presented below to illustrate the points.

4.1 Stress/Unstress Differentiation

Degree of contrast by F0 height, duration and intensity patterns of stressed/unstressed syllables are compared between L1 and L2 English (Taiwan speakers). Results of contrast degree for F0 and duration is L1>L2 while no significant difference is found for intensity. The results suggest that L1 English is more contrastive; its differentiation function is more robust. Distinct contrast is found in both F0/pitch and duration/rhythm patterns of L1 English whereas L2 English lacks the high/low pitch contrast but maintains the long/short rhythm. The result L2 is sounding flatter and less melodic than L1. The result suggests how learning stress assignment is simply about location and placement, but also about adequately realized contrast and differentiation (forthcoming).

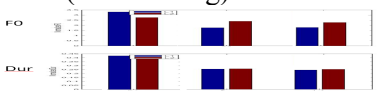

Figure4. Prosodic patterns by acoustic correlate, stress level (Primary, Secondary and Tertiary) and speaker group.

4.2 Narrow Focus and Contrast

Contrast differentiation was further investigated using elicited narrow focus produced by L1, Beijing (BJ) and Taiwan (TW) speakers [19, 20]. Results of F0 contrast is L1>TW>BJ though while post-focus F0 compression is only found in L1. Duration contrast showed different degrees of post-focus lengthening by L1, L2 TW speech and no contrast by BJ speakers. In other words, similar results of lexical stress contrast are also found at the sentence level.

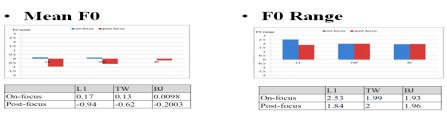

Figure5. Mean F0 and F0 range comparison between on-focus and post-focus constituents for L1,

TW and BJ speakers [20] .

5. Discussion and Conclusion

Speaking more fluently means having to maintain global prosodic patterns of discourse association. But speaking more naturally requires more expressions reflecting pragmatics and information structure while maintaining discourse association. More distinct high-low pitch contrast, slow-fast rhythm change and loudness adjustment is required to designate information weighting and speaker intention. Robust contrast is therefore necessary to make differentiation distinct. General L2 features, i.e., slower speaking rate, more intermediate units and pauses, highly varied intonation patterns, less distinct stress patterns and less post-focus compression, can all be characterized as underdifferentiating. Based on the above results, we argue that underdifferentiation is what makes prosodic output less natural.

6. References and appendices

1. Ladefoged, P (2006). A Course in Phonetics. *5th ed. Boston: Thomson.*
2. Halliday, M. (1967). Intonation and Grammar in British English. *Mouton, The Hague.*
3. Crystal, D. (1969). Prosodic Systems and Intonation in English. *Cambridge University Press, Cambridge*
4. Tseng, C-Y and Pin, S-H. (2004). Modeling prosody of Mandarin Chinese fluent speech via phrase grouping. *Speech and Language Systems for Human Communication (SPLASH-2004/Oriental-COCOSDA2004) 53-57. New Delhi, India*
5. Tseng, C-Y and Fu, B-L. (2005). Duration, Intensity and Pause Predictions in Relation to Prosody Organization. *Interspeech 2005 1405-1408. Lisbon ,Portugal.*
6. Tseng, C-Y. (2006). Recognizing Mandarin Chinese Fluent Speech Using Prosody Information—An Initial Investigation. *The 3rd Speech Prosody 2006. Dresden, Germany.*
7. Tseng C-Y. (2011). Phonotactic and Discourse Aspects of Content Design in AESOP (Asian English Speech cOrpus Project). *Oriental COCOSDA 2011. Hsinchu, Taiwan.*
8. Tseng, C-Y, Cheng, Y-C and Chang, C-H. (2005). Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech. *Oriental COCOSDA 2005 23-28. Jakarata, Indonesia*
9. Tseng, C-Y and Su, Z-Y. (2007). What Do Speakers Do and Why –The Story of Prosody-Syntax Non-Overlap and Higher Level Discourse Information. *Oriental COCOSDA 2007 27-32. Hanoi, Vietnam.*
10. Tseng, C-Y and Su, Z-Y. (2008). What's in the F0 of Mandarin Speech–Tone, Intonation and beyond. *ISCSLP 2008 45-48. Kunming, China.*
11. Tseng, C-Y (2012). Information Structure by Way of Discourse Prosody. *The 4th International Conference on Sinology 2012. Taipei, Taiwan.*
12. Hirose, K., Fujisaki, H. and Yamaguchi, M. (1982). Analysis and Synthesis of Voice Fundamental Frequency Contours of Spoken Sentences. *Speech, and Signal Processing, IEEE.*
13. Tseng, C-Y and Su, Z-Y. (2008). What's in the F0 of Mandarin Speech –Tone, Intonation and beyond. *ISCSLP 2008 45-48. Kunming, China.*
14. Tseng, C-Y and Su, Z-Y. (2008). Boundary and Lengthening—On Relative Phonetic Information. *The 8th PCC 2008.* Beijing, China.
15 Tseng, C-Y and Su, Z-Y. (2008). Discourse Prosody and Context – Global F0 and Tempo *Modulations. Interspeech 2008 1200-1203. Brisbane, Australia.*
16. Tseng, C-Y., Su, Z-Y. and Lee, L-S. (2010). Prosodic Patterns of Information Structure in Spoken Discourse—a Preliminary Study of Mandarin Spontaneous Lecture vs. Read Speech. *Speech Prosody 2010. Chicago, U.S.A*
17. Tseng, C-Y, Su, Z-Y, and Huang, C-F. (2011). Prosodic Highlights in Mandarin Continuous Speech ─Cross-Genre Attributes and Implications. *Interspeech 2011. Florence, Italy.*
18. Tseng, C-Y. and Su, Ch-Y. (2012). Information Allocation and Prosodic Expressiveness in Continuous Speech: A Mandarin Cross-genre Analysis. *ISCSLP 2012 243-246. Hong Kong.*
19. Visceglia, T., Tseng, C-Y, Su, Z-Y and Huang, C-F. (2011). Realization of English Narrow Focus by L1 English and L1 Taiwan Mandarin Speakers. *The 7th ICPhS 2011. Hong Kong, China*
20. Visceglia, T., Su, Ch-Y. and Tseng, C-Y. (2012). Comparison of English Narrow Focus Production by L1 English, Beijing and Taiwan Mandarin Speakers. *Oriental COCOSDA 2012 47-51. Macau, China.*