MELODY OF MANDARIN L2 ENGLISH—
WHEN L1 TRANSFER AND L2 PLANNING COME TOGETHER

Chao-yu Su¹,²,³ and Chiu-yu Tseng¹

¹ Phonetics Lab, Institute of Linguistics, 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 is always more difficult for L2 speakers to produce the melody and tempo of continuous speech because it requires simultaneous planning of L2 linguistic specifications, higher level discourse associations and information placements. We assume that higher level planning requires within-phrase chunking and cross-phrase paragraph phrasing while information arrangements through emphasis weighting assignment and allocation. The above involved planning is most notably delivered through distinct global melodic modulations and patterns. The present compares the onset features of extracted Phrase Commands and their consistency with tagged discourse units and perceived emphases using speech data of L1 English, Taiwan (TW) L2 English and TW L1 Mandarin. Explicitly, we study F0 contour features to compare L1/L2 chunking units and their global patterns to pinpoint L2 features. Results of distinct TW L2 features compared with L1 English are (1) less consistent discourse chunking, (2) fewer distinct contours by prosodic words, (3) less degree of emphasis contrast in prosodic phrase and (4) more distinct contours in non-emphases. While (1) and (2) may reflect general L2 planning difficulties, our results show that (3) and (4) namely, flatter overall contour, are Mandarin inherent transferred to L2. We believe our proposed methods of extracted Phrase Command more accurately and better represent global melodic features that could be applied to other L1/L1 comparison in general; the findings could also be directly applied to CALL development of L2 prosody enhancement to improve overall intelligibility.

Index Terms: discourse structure, information structure, F0 contour features, L1 and L2 speech planning, prosody, expression

1. INTRODUCTION
Studies of the features and variations present in L2 speech, often referred generically as foreign accent [1,2], grow more in recent years due to their applications to language identification, speech recognition and CALL (Computer Aided language Learning) in addition to English language education. Topics related to foreign accent used to concentrate on issues of bare form linguistic specifications (canonical forms) such as segments and individual phrase intonations. However, more recent literature reveals that additional features of L2 prosody variations also contribute significantly to L2 accent, producing as much of an effect on the intelligibility, comprehensibility and perceived accented-ness as well [3, 4, 5]. Studies of methodical training of prosody also showed positive effects towards L2 intelligibility [6, 7]. To this end and very much in need is more understanding that can be attributed to properties of higher level planning and their interaction with canonical linguistic information. The present study focuses on melodic features specific to the prosody of Taiwan (TW) L2 English on the one hand, but also aims at more generic method that could be applied across language as well.

Our present goal is to sort out from the speech signal interactive planning of canonical prosodic information, i.e., English phonology specified word stress and syntax specified sentence intonation, with systematic higher level constraints of discourse/paragraph structure and information weighting. Reported results of word stress produced by TW L2 speakers have revealed very interesting findings. When elicited target words are correctly pronounced, they are realized with less degree of F0 and amplitude contrasts than the L1 norm [8]. In addition, when the same polysyllabic words are embedded in sentences of different syntactic structures, their surface forms are more varied for both L1 and TW L2 speakers. However, the L1 variations turned out to be systematic and predictable while the L2 variations random. What happened was that L1 speakers would merge the secondary stress to either primary or tertiary, simplifying the three-way stress contrasts to a parsimonious binary (stress/no-stress) contrast, by merging the pre-primary secondary stress to primary and post-primary secondary stresses to tertiary, shown most markedly in the F0 domain (high/low contrast). As
a result, in output speech the primary/post-primary contrast is more robust while secondary stress appears in varied surface forms [9]. The same study also confirmed that this kind of predictable merge is difficult for L2 speakers, making it reasonable to assume that when more levels of planning is involved, it would trigger more deviations in L2’s global prosody and generate more inconsistencies in L2 speech. As expected, an initial L1/L2 comparison of the consistency of discourse boundaries by speaker and by language revealed more overlaps in L1 speech than L2 [11]. In addition to less consistent discourse structure in L2 speech, the present study assumes that simultaneous planning of L2 linguistic specifications, higher level discourse associations and information placements are the main difficulties that contribute to L2 melody. Along this vein, the following study aims to investigate global F0 difficulties that contribute to L2 melody. Along this vein, the following study aims to investigate global F0

2. SPEECH MATERIALS AND ANNOTATION

2.1. Speech Data

Five types of continuous speech by L1 English, TW L2 English and L1 Mandarin are used to examine discourse and information planning. The English materials are reading passage of “The North Wind and the Sun” by 13 American English L1 speakers (6M/7F) and 30 TW L2 speakers (15 males and females each), coded N&S in AESOP-ILAS (Asian English Speech Corpus Project—Institute of Linguistics Academia Sinica) corpus [10]. The TW Mandarin materials include 2 read and 1 spontaneous speech: (1) 1 male and 1 female reading of 26 discourse pieces (coded CNA in the Sinica COSPRO database) [11], (2) 1 male and 1 female simulation of broadcasting weather forecast (coded WB) and (3) 1 male spontaneous university classroom lecture. The speech details are summarized in Table 1.

Table 1. Summary of 5 types of speech data by language English (E), Mandarin (M) and genres story N&S, prose CAN, weather simulation WB and lecture Lec

<table>
<thead>
<tr>
<th>Information</th>
<th>Speaker</th>
<th>Length (Min)</th>
<th>Size (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language, L1/L2, genre</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E L1 N&amp;S</td>
<td>13 (6M/7F)</td>
<td>8.6</td>
<td>221</td>
</tr>
<tr>
<td>E L2 N&amp;S</td>
<td>30 (15M/15F)</td>
<td>28.2</td>
<td>597</td>
</tr>
<tr>
<td>M L1 CNA</td>
<td>2 (1M/1F)</td>
<td>107.4</td>
<td>165</td>
</tr>
<tr>
<td>M L1 WB</td>
<td>2 (1M/1F)</td>
<td>52</td>
<td>94</td>
</tr>
<tr>
<td>M L1 Lec</td>
<td>1 (1M)</td>
<td>145.2</td>
<td>265</td>
</tr>
</tbody>
</table>

2.2. Processing and annotation

The speech data of L1 English, TW L2 English and L1 Mandarin were tagged in layers for discourse as well as information structure. The preprocessing layer is force-aligned segments by the HTK Toolkit, followed by manual spot-checking by trained transcribers. Discourse units and information status were then manually tagged independently.

2.2.1. 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 [12]; the default unit/boundary correlations can be expressed as SYL/B1<PW/B2<PPh/B3<BG/B4<PG/B5. Though the smallest discourse unit by tagging protocol is the syllable (SYL), for the present study that focuses on global contour patterns from higher level information, we only adopt boundaries from B2 and above for comparison.

2.2.2. Tagging information structure by perceived degree of prominence

The same speech data are also manually tagged by trained transcribers by degree of perceived strength of emphasis/non-emphasis in to tokens (ETs) as a reference of information weighting. 4 degrees of perceived prominence are defined as follows:

- •E0—reduced pitch, lowered volume, contracted segments
- •E1—normal pitch, normal volume and clearly produced segments
- •E2—raised pitch, louder volume and with the speaker’s tone of voice
- •E3—higher raised pitch, louder volume and with the speaker’s change of tone of voice

2.3. The command-response model

The basic assumption of the command-respond model is that F0 contours of compound words and sentences can be generally characterized by the combined superposition of Baseline, Phrase Command (PCs) and Accent Commands (ACs) which correspond to gradual declination from the onset towards the end of a speech fragment, and local humps corresponding to accentuated positions. The model therefore allows contributions from global (PC) and local (AC) accountable and derivations of F0 output irrespective of the speakers register (Baseline) [13]. Within- and cross-linguistic comparison of patterns and magnitude by PCs and ACs allowed some intrinsic English/Mandarin difference to emerge [14]. That is, steeper high-to-low global contour and less local magnitude was found in English, whereas reversed patterns were found in Mandarin for both commands. These features combined provided
explanation of why TW L2 English sounds flatter and less distinct than the L1 norm. In the present study, the same model will be used to capture F0 features with respect to discourse and information structure.

### 3. METHOD

In order to better examine melodic features by F0 representations, the following two sets of comparisons are made. We first study how global F0 contour interacts with discourse planning by assuming that the onsets of PC (OPC) represent coarse-grained boundaries between retrievable intonation contours in the speech signal, and therefore may correlate positively indicate locations of weighting assignment from information planning. Since the data are already manually tagged for discourse boundaries, overlaps between manually-tagged [12] and OPC-derived boundaries are calculated. We then assume that correspondence correlations also exist between information weighting and F0 peaks, and further examines extractable peaks with manually tagged location and degree of prominence. By definition of the command-response model [13] should be characterized by the AC. However, evidence of Cantonese demonstrated how the emphasis effect is more pronounced by the global contour instead [15]. Therefore we propose to assume that the OPC also serves as a reference of emphasis placements before we investigate the AC and its correlation to emphasis in the future. More specifically, the present study will present distribution of overlap between the onsets of PCs (OPC) to (1) discourse boundaries (DB) and (2) emphasis boundaries (EB). We hope to explain the following by discourse/information planning: (1) in what additional and specific ways TW L2 English sounds different from L1 English, (2) what more intrinsic prosodic difference between Mandarin and English exists, and (3) whether indeed prosodic transfer from L1 Mandarin occurs.

#### 3.1. Extraction of onset positions of PC

In order to extract onset of PCs which represent boundaries of global F0, separating phrase component from F0 contour has been a major research issue. Low frequency contour (LFC) is a popular method that extracts phrase component form original F0 [16, 17]. LFC had been reported to be efficient for tone languages that we apply LFC on Mandarin speech in the present study. However, for non-tonal language such as English, LFC contour extraction derives negative values of accent component that is more tonal by nature, and further adjustment is needed. Hence in order to more efficiently extract phrase component we proposes a simple method for English. The method replaces LFC by a bottom line representing global F0 trajectory while positive characteristics of accent component in English maintains. The details are listed in follows.

1. Connecting local minimums of F0 to derive a bottom line represent global F0
2. Finding local minimums of the bottom line to segment F0
3. Fitting segmented F0 with function of phrase component (PC). Fitting criterion is to minimizing RMSE between segmented F0 and PC

Figure 1 shows extraction of PCs in the “north wind and the sun”.

After optimization of PC, extracted onset positions are compared with discourse boundaries in 2.2.1 and emphasis boundaries in 2.2.2 with 100ms tolerance.

### 4. RESULTS

Speech data of L1 English, L2 English and L1 Mandarin are compared by distribution of overlap between onset of PCs (OPC) with (1) discourse boundaries (DB) and (2) emphasis boundaries (EB) in order to examine Mandarin-English prosodic difference, source of L2 accent and possible prosodic transfer.

#### 4.1. Overlap between OPC and DB by language, genre and L1/L2

Figure 2 and Table 2 show comparison among L1 English, L2 English and L1 Mandarin by distribution of overlap between onset of phrase command (OPC) and perceived discourse boundaries (DB). Results show higher percentage of overlap in all types of speech data.
than non-overlap. Among them, E_L2_N&S shows lowest consistency between OPC and DB, about 54.03% and M_L1_WB shows highest consistency between OPC and DB, about 71%.

Figure 2: Diagrammatic comparison among L1 English, L2 English and L1 Mandarin by overlap distribution between OPC and DB.

Table 2: List of overlap between OPC and DB among L1 English, L2 English and L1 Mandarin

<table>
<thead>
<tr>
<th>Language, L1/L2, genre</th>
<th>Overlap %</th>
<th>Non-overlap %</th>
</tr>
</thead>
<tbody>
<tr>
<td>E_L1 N&amp;S</td>
<td>64.07</td>
<td>35.93</td>
</tr>
<tr>
<td>E_L2 N&amp;S</td>
<td>54.03</td>
<td>45.97</td>
</tr>
<tr>
<td>M_L1 CNA</td>
<td>60.50</td>
<td>39.50</td>
</tr>
<tr>
<td>M_L1 WB</td>
<td>71.18</td>
<td>28.82</td>
</tr>
<tr>
<td>M_L1 Lec</td>
<td>65.83</td>
<td>34.17</td>
</tr>
</tbody>
</table>

Figure 3 and Table 3 further demonstrate the distribution of overlapped OPC by break type and speech type. Among all types of speech, L2 speech shows most distinct patterns from others. Lowest B2/B3 ratio, i.e., L1 is found in L2 English, whereby the same B2/B3 ratio in the other types ranges from 2.4 to 3.8 and is much higher than L2 English. In addition, lowest B4/B5 percentage is found in M_L1_Lec at only about 1%.

4.1.1. Discussion

The overlap between PCs and discourse boundaries shows, among all speech types, L2 speech is lowest and account for L2 speakers’ discourse planning by global F0 is not as consistent as native speakers for both Mandarin and English. More detailed distribution of overlapped PCs by break type demonstrates L2 speakers plan discourse with more boundaries of prosodic phrase; however, native speakers tend to place more PC in prosodic word than prosodic phrase. Above results show distinct features of L2 English are (1) more inconsistent placement of PC with discourse planning and (2) fewer insertion of PC in prosodic words (compounds). Another interesting finding is that PC in spontaneous lecture speech is not consistent with higher-level discourse boundaries, B4/B5, that is against our assumption. We listen to some of B4/B5 examples in spontaneous lecture speech. Observation found when prosodic fillers occur between spoken discourses, perceived B4/B5s are forecasted and labeled before prosodic fillers with significant F0 reset that one-phrase shift between boundary positions and really F0 valley is often found in B4/B5. Similar finding has been reported in read speech of Mandarin in previous studies [18].

4.2. Overlap between OPC and EB

Figure 4 and Table 4 show comparison among L1 English, L2 English and L1 Mandarin by overlap distribution between onset of phrase command (OPC) and EB. Results show all types of speech data with higher non-overlap percentage than overlap percentage. Among them, M_L1_Lec shows lowest consistency between OPC and EB, about 18% and E_L1_N&S shows highest consistency between OPC and EB, about 43%.
Figure 5 and Table 5 further demonstrate distribution of overlapped OPC by emphasis type and speech type. Among all types of speech, L2 speech shows most distinct patterns from others. Highest E1/(E2&E3) ratio, 1.54 is found in L2 English whereby the same E1/(E2&E3) ratio in the other types ranges from 0.57 to 1.23 and is lower than L2 English.

Table 4: List of overlap between OPC and EB among L1 English, L2 English and L1 Mandarin

<table>
<thead>
<tr>
<th>Overlap/Non-overlap Language, L1/L2, genre</th>
<th>Overlap%</th>
<th>Non-overlap%</th>
</tr>
</thead>
<tbody>
<tr>
<td>E L1 N&amp;S</td>
<td>42.51</td>
<td>57.49</td>
</tr>
<tr>
<td>E L2 N&amp;S</td>
<td>27.73</td>
<td>72.27</td>
</tr>
<tr>
<td>M L1 CNA</td>
<td>37.74</td>
<td>62.26</td>
</tr>
<tr>
<td>M L1 WB</td>
<td>31.90</td>
<td>68.10</td>
</tr>
<tr>
<td>M L1 Lec</td>
<td>18.27</td>
<td>81.73</td>
</tr>
</tbody>
</table>

Table 5: List of overlapped OPC by emphasis types among L1 English, L2 English and L1 Mandarin

<table>
<thead>
<tr>
<th>Emphasis Type Language, L1/L2, genre</th>
<th>E1 %</th>
<th>E2&amp;E3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>E L1 N&amp;S</td>
<td>36.62</td>
<td>63.38</td>
</tr>
<tr>
<td>E L2 N&amp;S</td>
<td>60.65</td>
<td>39.35</td>
</tr>
<tr>
<td>M L1 CNA</td>
<td>52.50</td>
<td>47.49</td>
</tr>
<tr>
<td>M L1 WB</td>
<td>42.57</td>
<td>57.43</td>
</tr>
<tr>
<td>M L1 Lec</td>
<td>55.24</td>
<td>44.76</td>
</tr>
</tbody>
</table>

4.2.1. Discussion

The overlap distribution between PCs and emphasis placement shows that about 18%–43% of PCs are consistent with emphasis placement. Compared with observation of Cantonese emphasis effect realized in PC instead of AC [14], our results here are further supported with consistency by quantity. In short, among all speech types, L1 English is distinct from the other types by overlap between PCs and emphasis placement, showing possible language specific patterns. In addition, patterns of TW L2 English are similar to all types of L1 Mandarin, thus providing evidence of L1 transfer. More detailed distribution of overlapped PCs by emphasis type revealed specific L2 English feature different from both L1 English and Mandarin. That is, larger proportion of PCs is inserted in unadorned emphases. In summary, distinct TW features of L2 English are (1) less emphasis effect in the phrase domain which might transfer from their mother tongue. (2) Among emphasized PC, more PCs inserted in unadorned emphases.

5. GENERAL DISCUSSION AND CONCLUSION

The above analysis reveals L2’s distinct melody (F0) features by discourse and information planning. Discourse related features are (1) less systematic discourse planning suggesting more variation by speaker and (2) fewer insertion of PC in prosodic words suggesting less pronounced compounding effect. We believe both features provide Mandarin evidence to general L2 features in the literature [11, 15]. Since no significant difference is found between L1 English and Mandarin suggesting higher level discourse effect is not language specific. In addition, information related features are (3) less degree of emphasis contrast in the prosodic phrase and (4) more distinct contours in unadorned emphases. Analysis of information structure reveals that intrinsic prosodic difference may be one of the main reasons of L1 Mandarin melody to Mandarin L2 English. Moreover, more PCs inserted in unadorned emphases may further contribute to distinct L2 melody of Mandarin speakers.

In summary, we believe our results have shed lights on the significance of melodic features of Mandarin L2 English can be attributed to difficulties of simultaneous planning of higher level discourse units as well as assignment and arrangement of information weighting and location. We believe our proposed methods of extracting Phrase Command more accurately and better represent global melodic features that could be applied to other L1/L1 comparison in general; the findings could also be directly applied to CALL development of L2 prosody enhancement to improve overall intelligibility. Immediate future work will concentrate on patterns and features of AC in relation to the same framework of discourse association and information placements.
6. REFERENCE