A PHONETICS BASED COMPUTER AIDED PROSODY TRAINING SYSTEM FOR L2 ENGLISH LEARNING

Chao-yu Su [1,2] *& Chiu-yu Tseng*[1]
[1] Institute of Linguistics, Academia Sinica, Taipei, Taiwan
[2] Institute of Information System & Application, National Tsing Hua University, Taiwan
E-mail: cytling@sinica.edu.tw

ABSTRACT

The usual practice of learning L2 English prosody is a bottom up process, from initializing a word stress model followed by layering over effects from semantic and syntactic specifications, and finally paragraph association. The present study derives normalized 3-way stress models of F0 and duration from speech data with the above specifications of both L1 and L2 English to see where the differences are and how deviations from the norm that may inhibit intelligibility could be improved. The results are a word stress model achieved by teasing apart the contributions form sentential/paragraph effects. The normalized results reveal more distinct L1-L2 differences both in word phonology (categorical contrasts) and in larger speech units (sentence and paragraph); and how a model with these results could help L2 learning of English prosody.
Keywords: L2 English prosody, word stress, sentence, paragraph, phonetics, L2 phonology.

1. INTRODUCTION

Segmental variations between L1 and L2 have been a main research focus of L2 accent. However, recent studies reveal that similar to segmentals, prosodic variations also cause as much effect on the intelligibility, comprehensibility and perceived accent of L2 speech [1, 2, 3]. As a result, there has been a growing interest of computer-assisted prosody training systems showing how indeed prosody training could help improve the overall intelligibility of L2 speech [4, 5]. We assume the processes involved in learning English prosody by L2 speakers is bottom up and additive: starting from building up phonetic and phonological specifications by words, followed by superimposing higher level syntactic/semantic specifications from forming compounds, phrases and sentences, then upward to add patterned prosodic projections of paragraph associations to produce fluent continuous speech. The task is difficult from the beginning since English word stress specifies how in a multi-syllabic word, a syllable assigned with primary stress is positively correlated to longer vowel duration and higher F0 while the remaining secondary and tertiary syllables are negatively correlated with shorter duration and lower F0. The layering-over of higher-level information causes more alternations of pitch (high/low), tempo (fast/slow), loudness (strong/weak) and silent pause by unit of various sizes to simultaneously deliver linguistic information and communicative expressions. So far it is still challenging for a computer aided training systems to address the contributions made by individual these layers and examine both their separate and cumulative effects on speech output. We believe that technology development with phonetic knowledge that accounts for the majority of these contributions would help enhance overall prosody of L2 English speech. The present study is an attempt to show how a learner of L2 English may be aided to better learn word stress and on to improve overall prosody as the speech unit grows by size.

Reported previous studies of Mandarin Chinese successfully examined the sentential and paragraph factors in output speech by separating the contributions made by these factors while computed their cumulative contributions to output prosody at the same time [6]. We believe such higher-level contributions are not language specific and can be adopted to Taiwan L2 English and used to extracting underlying stress patterns from acoustic signals containing layered-over information.

A series of previous studies on the stress patterns of Taiwan (TW) L2 English with minimal sentential/boundary effects have shown that the multiple demands of layering linguistic and paragraph information over to lower level segmental and phonological specifications from word and stress have already shown their effects in how Taiwan L2 stress differs from L1 and why they inhibit intelligibility. Namely, TW L2 word stress is featured by lack of sufficient degree of prosodic difference in contrast degrees in F0 (high/low pitch contrast) and duration (Fast/slow contrast tempo contrast); the result is sounding overall flatter and less differentiable than L1 speech [7]. The results also revealed why stress related F0 and duration alternations is particularly difficult for TW L2 speakers and in what way they may in need of additional help to learn [8]. We therefore believe

these features could be built into a training system to facilitate improvement.

The goal of the present study could also be two-fold, one is to better account for how abstraction of English stress from more realistic speech of larger speech units is different for L1 and L2 speakers; another to build a prosodic training model from bottom upward for L2 speakers.

2. METHOD

2.1 Speech Materials

The AESOP-ILAS speech database is specially designed to investigate comparative L1 and L2 English prosody [9]. AESOP (Asian English Speech cOrpus Project) is a multinational collaboration whose aim is to build up English speech corpora across Asia that would represent the varieties of English spoken in that region. AESOP-ILAS (Institute of Linguistics Academia Sinica, Taiwan) is part of the AESOP consortium that specifically collects L2 English of Mandarin L1 speakers in Taiwan. A subset of the AESOP-ILAS corpus is used in the present study.

The materials used here include Task1 to Task 5 with varied elicitation and context setup: (1) Task 1-- 2/3/4-syllable target words of all possible stress patterns embedded in carrier sentences i.e., *"I said JAPANESE five times."* for the purpose of baseline comparison (2) Task 2--the same target words at phrase boundaries in yes-no questions, wh-questions and declarative sentences (3) Task 3-- target words in narrow-focus positions. For example, *Context: Do you like Japanese and Korean food?* Reply: *"I like JAPANESE food, but Korean food is too spicy for me"*. The 2/3/4-syllable target words are 20 frequently used words from 2-, 3- and 4-syllables categorized according to 10 syllabicity and stress type: (a) 2-syllable initial stress 2, (b) 3-syllable initial stress, (c) 3-syllable medial stress, (d) 3-syllable final stress, (e) 4-syllable initial stress, (f)4-syllable medial 1 stress, (g) 4-syllable medial 2 stress, (h) 4-syllable final stress, (i) left-headed compounds (e.g orange juice), (j) right-headed compounds (e.g. afternoon). The chosen words are money, morning, white wine, hospital, apartment, department, tomorrow, video, overnight, January, supermarket, elevator, available, Japanese, afternoon, misunderstand, information, experience, California and Vietnamese. Two additional sets of experimental sentences are also included for sample amount including (4) Task 4, function words in stressed and unstressed positions and (5) Task 5, prosodic disambiguation of syntactic structures.

Speech data were recorded by trained proctors in quiet rooms directly into a laptop computer, using a recording platform developed specifically for AESOP. Experimental sentences and context were preloaded and appeared individually on a computer screen. Participants wore head-mounted Sennheiser PC155 microphones positioned 2 cm away from their mouths; they were instructed to speak naturally at a normal rate and volume. The speech data of a total of 17 speakers were analyzed: 8 L1 North American English speakers (3 male and 5 female) and 9 TW L2 speakers (4 male and 5 female)

2.2 Data Analysis

Normalized stress patterns are presented by F0, duration to represent underlying stress patterns using Z-score normalization by each sentence first to remove speaker and sentence variation. In order to extract the F0 of lexical stress without intonation effect for subsequent analyses, a straight line with minimal distance (RMSE) to original F0 contour is derived to represent intonation and subtracted, the residual is regarded as F0 without intonation effect. In turn, duration extraction is also refined to remove the effect of intrinsic segmental duration and boundary lengthening using a multi-layered normalization method shown below [6], in which *factor1* represents information at the segmental level, *factor2* represents respective syllable position within the word (to remove word-final boundary lengthening effects), and ε_i represents all other unpredictable values. Extracted values μ_i thus represent duration values which have been normalized for intrinsic segmental duration and boundary effect:

$$x_i = \mu_i + factor_1 + factor_2 + ... + \varepsilon_i$$

3. RESULTS AND DISCUSSION

L1 vs. L2 differences in normalized F0 and duration are compared by three-way stress categories, i.e., primary, secondary and tertiary to examine how L2 can be distinguished from L1. We assume that these normalization derived stress patterns are closer representations of abstract phonological categories. This section presents normalized 3-way primary/ secondary/tertiary patterns of F0/duration which we take to represent stress abstraction of L1/L2 English. In addition, stress patterns with minimal higher-level sentential/boundary effect in carrier sentence are also compared with the derived stress patterns to characterize L1/L2 phonological differences.

3.1. F0

Stress patterns of F0 by L1/L2 are shown in Figure 1, Figure 2 and Figure 3 which respectively represent raw F0 with patterns of sentential/paragraph effect, normalized/underlying F0 patterns without higher-level sentential/paragraph effect and raw F0 patterns after excluding higher-level sentential/boundary effect by corpus design. When sentential/boundary effect involves, raw F0 in Figure 1 shows L2's P (0.515) is slightly higher than L1's P (0.462) and P-S discrimination is very similar between L1 and L2. After normalizing higher-level sentential/paragraph effect in figure 2, the P-T discrimination increases 33.8% for L1 speakers but only increases 13.2% for L2 speakers. Sentential/paragraph-excluded F0 patterns in Figure 3 shows similar L1-L2 discrimination with Figure 2, namely, normalized results.

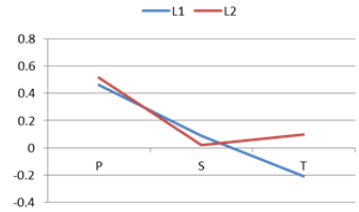

Figure 1: Sentential/paragraph-included F0 patterns by stress type in Task l to Task 5.

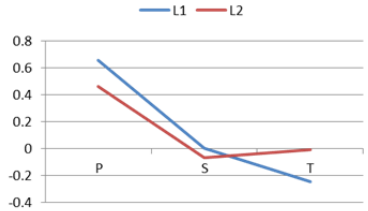

Figure 2: Normalized/underlying F0 patterns by stress type after removing higher-level sentential/paragraph effect in Task l to Task 5.

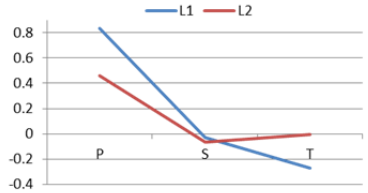

Figure 3: Sentential/paragraph-excluded F0 patterns by stress type in carrier sentence.

3.1.1. Discussion

In condition with patterns sentential/paragraph effect, raw F0 shows two features: (1) L1's P (primary stress) is lower than L2, and (2) L2's S-T (secondary/tertiary stress) differentiation is not clear and S (secondary stress) is even lower than T. In particularly, the first feature is counter-intuitive and contradicts the definition of stress. After normalization of sentential/paragraph effect, the results show larger 3-way contrast discrimination in L1 than L2 especially in P-S contrast. L1's P is found higher than L2's P after the removal of higher-level effect. Compared to raw F0 patterns, L2's normalized patterns still show S-T under-differentiation and the discrimination between L1 and L2 is now more distinct. Comparison of stress categories of F0 by normalized/underlying and sentential/paragraph-excluded show similar trends and patterns; the patterns are different from sentential/paragraph-included stress patterns. The above results show that L2 stress without higher-level information is quiet distinct from L1, suggesting different L2 phonology at work. Additional higher level information not only makes it all the more difficult for L2, but also distract the L2 learner from extracting phonological distinctions from larger and more varied speech units. Taiwan L2 learners need more attention calling to target F0 raising of the primary stress when learning single English words, but to overall F0 lowering when producing larger speech units.

3.2. Duration

Stress patterns of duration by L1/L2 are shown in Figure 4, Figure 5 and Figure 6 which respectively represent raw duration patterns with patterns of sentential/paragraph effect, normalized/underlying duration patterns without patterns of sentential/paragraph effect and raw duration patterns after excluding higher-level sentential/boundary effect by corpus design. When sentential/boundary effect is involved, the raw duration in Figure 1 shows L1's P (0.574) almost overlaps with the P of L2 (0.584) and the S-T discrimination is very similar between L1 and L2. After normalizing higher-level sentential/paragraph effect in Figure 2, L1's P (0.291) is slightly longer than the P of L2 (0.208). Duration patterns with minimal higher-level sentential/boundary effect in Figure 3 shows longer P duration in L1 (0.719) than P duration in L2 (0.303). The normalized patterns are closer to stress specified duration patterns by definition.

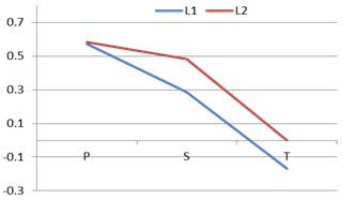

Figure 4: Sentential/paragraph-included duration patterns by stress type in Task l to Task5.

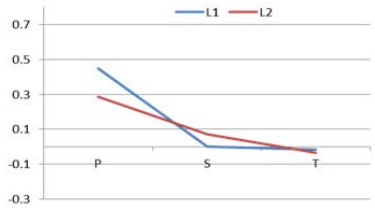

Figure 5: Normalized/underlying duration patterns by stress type after removing higher-level sentential/paragraph in Task l to Task5.

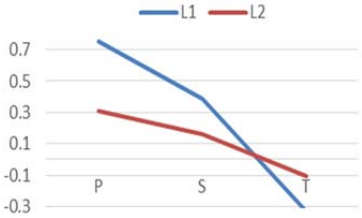

Figure 6: Sentential/paragraph-excluded duration patterns by stress type in carrier sentence.

3.2.1. Discussion

In condition with patterns sentential/paragraph effect, raw duration shows two features: (1) L1's P almost overlaps with the P of L2, and (2) L2's P-S differentiation is not clear. However, when the sentential/paragraph effect is removed, the patterns are reversed and the results show larger discrimination in L1 than L2 by P-S contrast. L1's P is found longer than the P of L2 when the higher-level effect is removed. Comparison of stress patterns of duration by normalized/underlying and sentential/paragraph-excluded shows L1's longer primary duration appears in both of them but not in sentential/paragraph-included patterns. Once again, these results suggest how the phonology of L1 is different from L2. The same results also suggest that special attention needs to be given to L2 learning of stress related duration adjustments of words produced in isolation and in larger speech units.

4. GENERAL DISCUSSION

From the above analyses of L1/l2 English stress related F0 and duration, we found that sentential/paragraph-included stress patterns are different from underlying/phonological stress patterns by normalization or sentential/paragraph-excluded patterns from corpus design. The results suggest that L2 learners such as TW Mandarin speakers may need special help to learn word stress, both at the beginning and later at intermediate stages, for different reasons and with different measures. When learning lexical stress by single words, they need to learn to raise the F0 of the primary stress to a higher caliber and to lower the F0 of the secondary and secondary stresses to different calibers by a normalized scale in order to produce the unfamiliar sing-song effect. Later when learning to speak in larger units and longer passages, the same speakers will benefit if their attention can be called to overall melodic patterns. Similar attention to tempo adjustment in both individual word and continuous speech should also help improve the awareness and learning of duration adjustment. We therefore propose that such phonetic understanding could be utilized to construct computer aided training systems for L2 speakers.

From a modeling point of view, these results also demonstrate construction of word stress systems may not benefit from of sentential/-paragraph-included speech materials, that is, even when the sentential/paragraph effect is minimal, as shown in the data design of AESOP. A spoken dictionary of words is still necessary. In turn, a spoken dictionary is by no means sufficient to model continuous speech, even when the sentences are short and simple. Our results also show that native (L1) speakers may choose to realize word stress through binary stress/no-stress contrast anchored by the position of primary stress. Post-primary secondary syllables are reduced to near-tertiary stress while pre-primary secondary syllables are elevated to near-primary magnitude in F0. The primary/secondary/tertiary contrast is merged into a binary stress/no-stress contrast with robust prosodic contrast between the primary stress and its following syllable(s). As expected, the position-related merge of the secondary stress is difficult for TW L2 speakers.

5. CONCLUSION

The present study examines more detail of normalized/underlying 3-way stress patterns of F0 and duration in viewpoint of bottom-up model and hope to derive patterns which represent 3-way stress abstraction of L1/L2 English. Furthermore, we are also interested in how phonologically L2's stress model is different from L1. More distinct L1-L2 difference is found by underlying/phonological stress patterns than sentential/paragraph-included stress patterns. In short, abstraction of lexical phonology could be achieved and derived by normalization to better represent stress in condition with sentential/paragraph effect while data-driven prosody training system could benefit from phonetic knowledge. Our prosody training system under construction should prove helpful to L2 learners.

6. REFERENCES

[1] Anderson-Hisieh, J., Johnson, R. and Koehler, K. 1992. The relationship between native speakers judgments of nonnative pronunciation and deviance in segmentals, prosody and syllable structure *Language Learning 42: 4* 529-555.

[2] Munro, M. J., & Derwing, T. M. 1994. Evaluations of foreign accent in extemporaneous and read material *Language Testing*, 11, 253–266

[3] Baker, W., & Trofimovich, P. 2006. Perceptual paths to accurate production of L2 vowels: The role of individual differences *International Review of Applied Linguistics in Language Teaching (IRAL)*, 44, 231-250.

[4] Hardison, D. M. 2004. Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, 8, 34-52.

[5] Hirata, Y. 2004. Computer-assisted pronunciation training for native English speakers learning Japanese pitch and duration contrasts. *Computer Assisted Language Learning*, 17, 357-376.

[6] Tseng, C. Y., Pin, S. H., Lee, Y. L., Wang, H. M. and Chen Y. C. 2005. Fluent speech prosody: framework and modeling *Speech Communication, Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation* 46(3-4): 284-309.

[7] Tseng, C. Y., Su, C. Y. and Visceglia, T. 2013. Underdifferentiation of English Lexical Stress Contrasts by L2 Taiwan Speakers. *Slate 2013* 164-167. Grenoble, France.

[8] Tseng, C. Y., and Su, C. Y. 2014. Prosodic Differences between Taiwanese L2 and North American L1 speakers — Under-differentiation of Lexical Stress. *The 7th Speech Prosody Conference.* Dublin, Ireland.

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