The Long Road from Phonological Knowledge to Phonetic Realization – An Acoustic Account of the Temporal Composition of Mandarin L2 English

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

[1] Institute of Lingustic, 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

Producing continuous speech in L2 is a challenging task. We accept that the composition of speech tempo involves multiple linguistic levels of contributions. We further hypothesize that respective contributions in the speech signal could be better accounted for through normalization of acoustic contributions, and examined the English phonetic inventory, the way stress type (primary, secondary and tertiary), boundary type (non-phrase final, continuation rise, final rise and final fall), as well as focus status (non-focus, function words, broad focus and narrow focus). Analyses of speech data of L1 vs. Mandarin L2 English not only verified the contribution of each factor examined, but also demonstrate in what explicit ways the temporal composition of Taiwan Mandarin L2 English differs from the L1 norm. In short, a discrepancy between linguistic awareness and phonetic execution leads to difficulties by lower level units such as segments and stress patterns; whereas higher level planning difficulties leads to deviations exhibited in boundary adjustments and realization of broad and narrow focus contrasts. We believe the results shed new light on temporal composition both L1 and L2 English, facilitate better understanding of tempo structure that can be directly applied to L2 teaching and computer aided training.

Index Terms: Mandarin L2 English, speech communication, temporal structure, communication function, L2 accent, L2 incomprehensibility

1. Introduction

Speech production is generally hypothesized a complex procedure accommodating multiple linguistic specifications and communicative functions while the communicative functions are encoded in parallel into melodic primitive, namely, prosody [1, 2, 3]. These linguistic specifications and communicative functions are reflected in various prosodic layers including at least lexical, syntactic, etc. that jointly attribute to output prosody. As a result, correlating acoustic parameters are systematically governed by multiple prosodic modulations and collectively attribute to intended communication goal [2, 3, 4]. Following this vein of rationale, it is reasonable to assume that producing L2 prosody is cognitively demanding and challenging since it involves simulating the outcome that is derived from complex interactions of multiple contributions. While each level of specification involved contributes to L2 prosodic variation, their effects to output speech is cumulative.

However, we note that the majority of reported L2 studies are focused on individual linguistic level at one time; their respective and collective effects not yet clear. For example, phonetic studies of L2 English consonants and vowels produced by Javanese and Swedish speakers showed temporal patterns that differ from L1 English due to respective L1 influences [5, 6]. Lexical stress studies at the word level reveal that both Japanese and Mandarin L2 English learners do not provide sufficient acoustic contrast between stressed and unstressed syllables as native speakers do [7, 8]. Similarly, studies of focus and emphasis in English sentences produced by L2 Taiwan and Beijing speakers revealed a general pattern of under-differentiation due to inadequate on-focus/post-focus contrasts due to respective L1 features [9]. Comparable results are also found later in Hong Kong L2 English due to insufficient post-focus compression [10]. While all of the above studies illustrate how and why L2 prosody varies from the L1 norm in one particular linguistic level, between- and among-level interactions are often not addressed since separating acoustic signal of surface prosody into particular specifications requires methodological refinements.

It was reported that a Mandarin based hierarchical prosody framework characterized by including larger prosody units of continuous speech like phrases and speech paragraphs made possible quantitative account of contributions from lower to higher linguistic specifications, global level between- and cross-phrase associations as well as their cumulative effects [4]. Its corresponding modeling procedures specify how from the lowest level upward the prosodic hierarchy, individual contributions from the syllable, words, phrases, sentences and paragraphs receive layers of superimposed constraints from higher level specifications to finally form coherent continuous speech. Following the same vein of rationale, in the study reported below is aimed at better understanding how these interacting factors behave individually in the acoustic signal for L1 and L2. We will use identical elicited speech data of L1 American and Mandarin L2 English and examine each contributing factor independently by removing other interacting factors through statistical normalization. The contributing factors considered are (1) intrinsic duration from physical composition of segments, (2) phonology specified word stress, (3) phrase- and sentence-specified boundary type and (4) focus status constrained by sentence structure (broad vs. narrow focus). We believe the results will facilitate account of the temporal composition of Mandarin L2 English better and help understand accent explicit to Mandarin L2 English.

10.21437/SpeechProsody.2016-4

2. Speech Materials and Annotation

2.1. Speech Data

Subsets of the AESOP-ILAS [11] speech database are used for the present study. AESOP (Asian English Speech cOrpus Project) is a multinational collaboration of data collection project whose aim is to build up English speech corpora across Asia that would represent the varieties of English spoken in that region, with special focus on prosodic properties. 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. The materials used include Task1 to Task 3 which were designed to elicit lexical stress, boundary effect and contrastive focus, respectively. A total of 20 frequency- controlled and stress-balanced (2-4 syllables) target words were embedded in carrier sentences (Appendix A). 15 of the target words reappeared in sentences controlled for boundary type, board and narrow focus (Appendix B). Speech data were recorded by trained proctors in quiet rooms directly into a laptop computer, using a recording platform developed specifically for AESOP by Hong Kong Chinese University. Participants were instructed to speak naturally at a normal rate and volume. The speech data of a total of 41 speakers were analyzed: 11 L1 North American English speakers (5 male and 6 female) and 30 TW L2 speakers (15 male and 15 female)

2.2. Processing and annotation

The speech data of L1 English, TW L2 English were tagged by multiple layers of linguistic specifications. The preprocessing layer is force-aligned segments by the HTK Toolkit followed by manual spot-checking by trained transcribers. Following the tagging of segment, lexical stress (primary, secondary and tertiary) is labelled manually in syllable unit by dictionary transcription. For higher-level of communicative function in addition to phoneme and lexical stress, phrase boundaries (non-phrase boundary, continuation rise, final rise and final fall) and focused status (function word, non-focus, broad focus and narrow focus) is further tagged in word unit by an English native speaker.

3. Method

In addition to corpus designed to elicit to layering-over effect by lexical stress, boundary effect and contrastive focus, computational normalization is further conducted to tease apart linguistic specifications combined in speech and derive individual model by each level.

3.1. Normalization and Computational modeling

Z-score normalization by each sentence is conducted first to remove speaker and speech rate. Speech rate-normalized segment duration is first clustered by phoneme type. Each cluster mean by phoneme type is calculated to represent phoneme models. We assume phone model and effect from immediate upon-phone level, i.e., stress, jointly compose duration output. In order to derive stress effect embedded in duration output, the values of duration output subtracting phoneme models are regarded as stress effect. Following the same rationale, each cluster mean by stress type is calculated to represent stress models and immediate upon-stress model; boundary type could be derived as well. The procedure is recursively conducted layer by layer till modeling duration in level of phoneme, lexical stress, boundary effect and contrastive focus achieved. A formulation representing temporal structure with multiple functional layers is shown below, in which x_i denotes surface tempo and PM, SM, BM, FM, ε represent phoneme model, stress model, model of boundary type and model of focus status and residual error respectively.

$$x_i = PM + SM + BM + FM + \varepsilon$$

4. Results

4.1. L1-L2 temporal difference by intrinsic duration from segmental composition

4.1.1. L1-L2 temporal difference of duration by vowel and consonant

Table 1 shows average L1-L2 difference by vowel and consonant duration. The mean values of temporal difference are 0.252 for vowel and 0.247 for consonant respectively.

Table 1. *Average L1-L2 difference by vowel and consonant duration*

Vow/Con / Stat	Vowel	Consonant
Mean	**0.252**	**0.247**
STD	**0.161**	**0.174**

4.1.2. L1-L2 difference by vowel duration

Table 2 lists duration models by specific vowel type. By L1-L2 difference, the top-5 vowels are ʌ, ʊ, o, ɑ, ə which range from 0.352 to 0.571.

Table 2. *Duration models by vowel type*

L1/L2 / Vowel Type	L1	L2	L1-L2 Diff
ʌ, ə + l	**0.832**	**1.403**	**0.571**
ʊ	**-0.596**	**-0.133**	**0.463**
o	**2.535**	**2.112**	**0.423**
ɑ	**-0.374**	**0.023**	**0.397**
ʌ, ə	**-0.403**	**-0.051**	**0.352**
ɪ	**-0.32**	**-0.014**	**0.306**
aɪ	**0.872**	**1.113**	**0.241**
ɛ	**0.322**	**0.543**	**0.221**
u	**0.015**	**0.23**	**0.216**
eɪ	**0.554**	**0.704**	**0.15**
ɒ	**0.027**	**0.172**	**0.145**
æ	**0.368**	**0.467**	**0.098**
i	**0.371**	**0.299**	**0.073**
au	**2.672**	**2.603**	**0.069**
ə, ɚ	**0.422**	**0.47**	**0.048**

4.1.3. L1-L2 temporal difference by consonant duration

Table 3 shows duration models by specific consonant type. By L1-L2 difference, the top-5 consonants are θ, ʒ, dʒ, h, ŋ which range from 0.419 to 0.0.789.

Table 3. Duration models by consonant type

L1/L2 Consonant Type	L1	L2	L1-L2 Diff
θ	0.021	0.809	0.789
ʒ	0.019	-0.478	0.497
dʒ	0.083	-0.412	0.495
h	-0.381	0.074	0.456
n	0.365	-0.055	0.419
ʃ	0.573	0.233	0.341
p	-0.091	-0.391	0.301
s	0.65	0.406	0.245
ɹ	-0.157	-0.397	0.239
j	-0.438	-0.674	0.237
g	-0.573	-0.776	0.203
f	0.37	0.174	0.197
l	-0.293	-0.476	0.182
n	-0.178	-0.333	0.155
m	0.047	-0.104	0.151
tʃ	0.311	0.164	0.147
d	-0.691	-0.831	0.14
z	0.306	0.438	0.132
k	-0.03	-0.158	0.128
b	-0.627	-0.755	0.128
v	-0.391	-0.484	0.092
ð	-0.741	-0.657	0.084
w, ʍ	-0.317	-0.401	0.084
t	-0.572	-0.491	0.081

4.1.4. Discussion

The above duration patterns by L1 and L2 show how segmental intrinsic temporal features do exist for Mandarin L2 speakers at the phoneme level. By the duration of vowel type, the majority of most L1-L2 difference is found in the central and back vowels (ʌ, ə, o, ʊ, ɑ). By duration of consonant type, L2 produced fricatives (θ, ʒ and h) are distinct from L1, illustrating how fricatives are more challenging to TW L2 speakers than the other consonant types. In short, the segmental inventory and the phonotactics patterns are both tasking to Mandarin L2 speakers.

4.2. L1-L2 temporal difference by word stress specified by phonology

Figure 1 shows duration patterns by stress type and speaker group (L1/L2) while segmental effects are removed. Mandarin L2 English shows less degree of contrast among primary, secondary and tertiary stresses than L1.

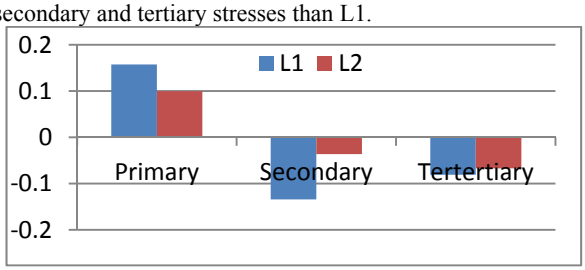

Figure 1: *Duration patterns by stress type and speaker group (L1/L2) without segmental effect.*

4.2.1. Discussion

The above results show the more realistic temporal status of lexical stress superimposed on segments (without interaction). Overall, TW L2 speech exhibits a general pattern of less contrast degrees. It is therefore reasonable to state that at the lexical level, TW L2 word stress is under-differentiated, and less differentiable than L1.

4.3. L1-L2 temporal difference by boundary type correlating to sentential structure

Figure 2 shows duration patterns by boundary type and speaker group (L1/L2) while lower-level effect (segmental and stress effect) are removed. Non-phrase final boundary, mid-phrase Continuation Rise, Final Rise in yes-no question and Final Fall (sentence final) are coded NB, CR, FR, FF, respectively.

L1 patterns show considerable pre-boundary lengthening except NB; the degree of lengthening among CR, FR and FF, namely, however, shows little distinction (0.173, 0.172 and 0.171). As expected, different patterns are found in TW L2 speech which shows less significant degree of lengthening across type, especially in type CR (-0.024) which is much longer in L1 speech (0.173).

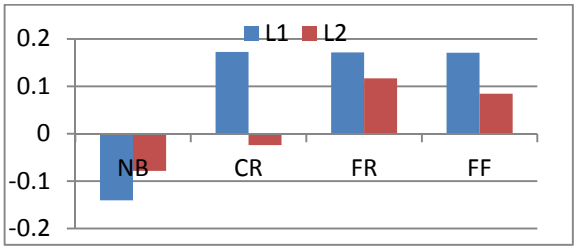

Figure 2: *Duration patterns by boundary type and speaker group (L1/L2) without lower-level effect (segmental and stress effect). Non-phrase final boundary, continuation rise, final rise and final fall are coded NB, CR, FR, FF respectively*

4.3.1. Discussion

The above results demonstrate duration adjustments by prosodic boundary in addition to segmental and stress effect. Per-boundary lengthening (final lengthening) is found across continuation rise, final rise and final fall while non-phrase final boundary is shortened instead. Note how mid-sentence phrase-final continuation rise is accompanied with similar degree of phrase final lengthening. In other words, boundary lengthening as a boundary final effect is only L1 evident. In the case of TW L2, while similar shortening is found in non-phrase final and pre-boundary lengthening in final rise and final fall, it is again marked by lesser degree, as exhibited in the stress related patterns (4.2.). Interestingly, duration lengthening associated with continuation rise appears to be overlooked by TW L2 since 2 pattern shows slight shortening instead of lengthening which is opposite of L1. It is therefore reasonable to state that at the phrase level, final lengthening in TW L2 speech is not only under-differentiated than L1, its mid-sentence continuation rise is the most distinct feature from L1 since it is treated as mid-phrase non-final boundary.

4.4. L1-L2 temporal difference by focus status correlating to information structure

Figure 3 and Figure 4 are duration patterns by focus status for L1 and L2, correspondingly. Function word, non-focus, broad focus and narrow focus are coded FW, NonF, BF and NF,

respectively. The lower-level effects (segmental, stress and boundary effect) are removed to derive the duration model of focus status. In particular, we assume that stress induced duration adjustment of segments may vary by their respective positions in the syllable and therefore further classify focus status by stress type. The results of L1 and L2 are shown in Figures 3 and 4, respectively. The patterns of L1 primary and secondary stress show a rising trend of lengthening by focus status, i.e., NonF <BF <NF, whereas tertiary stress is shortened, a possible effect of vowel reduction. The L2 patterns show similar trend with L1 only in tertiary stress shortening; while trends of primary and secondary stress are distinctly different from L1 patterns, and show no correlation by focus status.

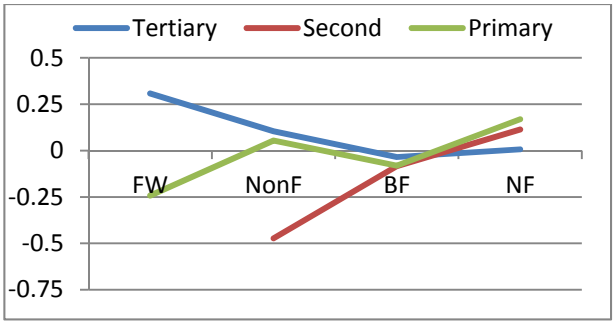

Figure 3: *Duration patterns of L1 speech by focused degree without lower-level effect (segmental, stress and boundary effect).Function word, non-focus, broad focus and narrow focus are coded FW, NonF, BF, NF respectively*

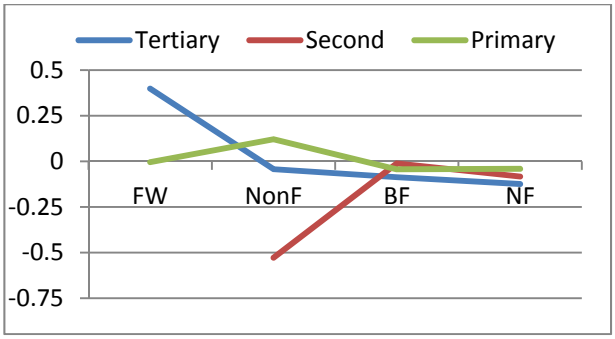

Figure 4: *Duration patterns of L2 speech by focused degree without lower-level effect (segmental, stress and boundary effect). Function word, non-focus, broad focus and narrow focus are coded FW, NonF, BF, NF respectively*

4.4.1. Discussion

The above results demonstrate that focus status would superimpose yet another level of temporal adjustments by focus status in addition to duration effects from segmental, stress, and boundary related adjustments. The L1 patterns show how focus degree is accompanied by increasing degree of lengthening from function word, non-focus, broad focus to narrow focus; primary and secondary stress requires duration lengthening while tertiary stress the opposite. It is important to note here that the derived patterns also illustrate how focus related duration adjustments, a higher level effect, are in sync with lower level word level stress specifications, thus demonstrating how intricate interactions must occur between linguistic levels. The same results also explain why an exaggeration method by placing equal degree of acoustic adjustments by words to enhance focus would not yield satisfactory results reported in [12]. On the other hand, the diverse L2 patterns by focus status which is more complex than under-differentiated stress patterns further imply how high level planning must require more proficiency of the target language.

5. Discussion

By sorting out the respective contributions in multiple levels of linguistic specification, it is now evident that each level involved does contribute independently to duration adjustments of various degrees. The physical constitution of segments at the lowest level of the prosodic hierarchy is the building block. Word level stress specifications are then superimposed and trigger systematic adjustments; primary stress requires duration lengthening while secondary and tertiary stress shortening. Note that their respective degree of contrasts must be robust enough to signal differentiation. Pre-boundary final lengthening is a phrasal effect. However, mid-sentence continuation rise requires similar degree of boundary lengthening as both phrase final rise and phrase final fall, suggesting that their respective differences must be signaled through other acoustic parameters such as the F0. Focus, a higher level sentential phenomenon, is related to duration lengthening that must observe lower level stress related specifications at the same time.

6. Conclusions

From the above results, we have reached the conclusion that the temporal composition of continuous speech could indeed be better understood using a simple but more refined methodology. We were able to tease apart the temporal constitution by separating contributions from phoneme type, word stress, boundary type and focus status; and compared their patterns in L1 and TW L2 speech to illustrate how L2 deviations are formed. It is clear now that L2 temporal variations are largely due to two reasons: (1) the discrepancy between linguistic awareness and phonetic execution at the lower levels, as shown in the case of segments both vowel quality (central and back vowels) and consonants (fricatives) as well as word stress, i.e. under-differentiation of stress category. (2) Difficulty to manipulate duration adjustments from higher level specifications, as shown in boundary lengthening and focus implementation. We believe our results shed new lights on the temporal composition of English in general, help sort out the differences in more detail of TW L2 English in particular. Future work includes similar investigations of F0 properties and data of narratives of longer passages. We believe our results could also be used in teaching English prosody as well as forming a bottom-up computer-assisted training system to improve overall proficiency of L2 English prosody.

7. References

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

[2] 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.

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

[4] 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(3-4): 284-309, 2005.

[5] Perwitasari, A., Klamer, M., Witteman, J. and Schiller, N.O., "Vowel duration in English as a second language among Javanese learners", International Conference on Phonetic Sciences 2015, Glasgow UK, August 2015.

[6] Thorén, B. "Swedish accent - duration of post-vocalic consonants in native swedes speaking English and German", International Conference on Phonetic Sciences 2007, Saarbrücken Germany, August 2007.

[7] Nakamura, S. "Analysis of Relationship between Duration Characteristics and Subjective Evaluation of English Speech by Japanese learners with regard to Contrast of the Stressed to the Unstressed", Journal of Pan-Pacific Association of Applied Linguistics, 14(1), 1-14, 2010.

[8] 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, 2013.

[9] Visceglia, T., Su, C. Y. and Tseng, C. Y. "Comparison of English Narrow Focus Production by L1 English, Beijing and Taiwan Mandarin Speakers", Oriental COCOSDA 2012 47-51. Macau, China, 2012.

[10] Gananathan, R.Y., Yin, Y., Ki, K., and Mok, P. "Interlanguage Influence in Cues of Narrow Focus: a study of Hong Kong English", International Conference on Phonetic Sciences 2015, Glasgow UK, August 2015.

[11] 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

[12] Lu, Jingli., Wang, Ruili., Silva, L C. "Automatic stress exaggeration by prosody modification to assist language learners perceive sentence stress", International Journal of Speech Technology, 2012, Volume 15, Number 2, Page 87, 2012

8. Appendix

Appendix A

Task 1:

Carrier sentence: " I said TARGET WORD five/ten times."

20 Target words by syllabicity (2-4) and stress type (syllable number/primary stress position): money, morning, wonderful, video, apartment, tomorrow, overnight, Japanese, elevator, January, available, experience, information, California, misunderstand, Vietnamese, supermarket, department store, white wine, afternoon.

Appendix B

Examples of Task 2:

Target words at prosodic boundaries:

Continuation rise (IP rise)

1. Do you know that in December and January,
2. Although Fred didn't have any experience,

Final fall (IP fall)

1 the sun rises at seven in the morning.

2.He had no trouble learning how to make a video

Final rise (IP rise)

1. Do you need any money?
2. Did he go to the hospital?

Examples of Task 3:

Target words in narrow focus:

1. Context: Are we allowed to make audio and video recordings?
   Answer: No. VIDEO recordings are not allowed.
2. Context: Can we open a branch of our office in this building?
   Answer: No. This is an APARTMENT building, not a commercial building.