Native and Non-native English Speakers' Assessment of Nuclear Stress Produced by Chinese Learners of English

Congchao Hua
Hubei University, City University of Hong Kong

Bin Li
City University of Hong Kong

Ratree Wayland
University of Florida

Abstract
This study compared naïve native and non-native English speakers' assessment of nuclear stress produced by Chinese learners of English and explored the effects of prosodic cues on their assessment. Adopting rapid prosody transcription (RPT), naïve raters comprising 36 highly proficient non-native English speakers and 30 native English speakers rated 176 sentence recordings produced by six Chinese learners of English. Results revealed that the native and non-native raters made generally comparable judgements and their ratings were reliable compared with expert rating. However, ratings by the two groups differed significantly on 20 sentences. Acoustic analysis showed that while native speakers relied on duration when identifying nuclear stress in learners' English, non-native speakers relied on both duration and intensity.

Anne Mette Nyvad, Michaela Hejná, Anders Højen, Anna Bothe Jespersen & Mette Hjortshøj Sørensen (Eds.), ***A Sound Approach to Language Matters – In Honor of Ocke-Schwen Bohn***
(pp. 187-209). Dept. of English, School of Communication & Culture, Aarhus University.

1. Introduction

Nuclear stress in English is particularly important for marking information focus (Dickerson, 1989; Lu et al., 2012). It marks the speaker's selection of priority in thought groups, and thus facilitates information processing by the hearer (Fouz-González, 2015). Misplaced nuclear stress often affects comprehensibility of both native and non-native English speech (Jenkins, 2002; Hahn, 2004; Luchini, 2005; Ingels, 2011; Frost, 2011).

In speech production, nuclear stress is realized through prosodic correlates such as F0, duration and intensity, yet roles that these cues play in stress marking vary across languages (Mennen, 2015). These cross-language variations may lead to difference in perception of nuclear stress by native speakers and second language (L2) learner.

Research has revealed that naïve first language (L1) listeners can reliably transcribe sentence stress in both L1 speech and L2 speech, yet it is not clear whether L2 listeners can also recognize sentence stress in a comparable manner. Thus, this study set out to compare L1 and L2 English speakers' perceptual judgement of nuclear stress produced by L2 English learners. In addition, an acoustic analysis was performed to identify effects of phonetic cues on L1 and L2 perception of nuclear stress.

2. Literature Review

2.1 Nuclear Stress in English and Mandarin Chinese

Nuclear stress in English refers to the stress associated with the nuclear tone in an intonation unit. Native speakers of English often follow a specific pattern for nuclear stress assignment. When a whole utterance is under focus (broad focus), nuclear stress is by default on the last content word (Crystal, 1969; Roach, 1991; Cruttenden, 1997). This is proved to be true by Alternberg (1987), who reports that 88% of the utterances in the London-Lund corpus have their nuclear stress on the last content word.

Example 1 --What happened?
--The baby is crying.
Example 2 --What's wrong?
--He cheated us.

Here the underlined syllable in each example carries nuclear stress and is the focus of the whole information structure. In Example 1, nuclear stress falls on the last word as it is the last content word in the utterance, whereas in Example 2, nuclear stress falls on the penultimate word, which is also the last content word in the utterance. However, there are some exceptions to the last-content-word rule.

Example 3 --Have you been to the lake?
--We walked around it.
Example 4 --What's the news?
--I met the president this morning.

In Examples 3 and 4, nuclear stress does not fall on the last content word. In example 3, it falls on 'around', which is a function word, and in Example 4, it falls on 'president', which is the penultimate content word. We will turn to these exceptions in detail later.

The above are all examples of broad focus. There is another type of focus: narrow focus. Narrow focus signifies contrast to known information or emphasis of new information. Often the word under narrow focus carries nuclear stress, despite its grammatical category or semantic weight.

Example 5 --Did you see Jane?
--No, I talked with Jane.
Example 6 --Who won the game?
--We won the game.

In Example 5, 'talked' carries nuclear stress because it contrasts with 'see', and in Example 6, 'We' carries nuclear stress because it directly answers the question 'Who' and is therefore emphasized. In both examples, nuclear stress does not fall on the last content word. To express contrast or emphasis in English, nuclear stress can fall on any word under focus.

Apart from the default pattern on the last content word, nuclear stress assignment involves over a dozen exceptional patterns that mainly include sentences ending with a function word, an early-stressed compound, a reflexive or reciprocal pronoun, a reporting phrase, a parenthetical, an empty word, a time or place adverbial, a phrasal verb ending with a preposition, a noun modifier, repeated information, and contrastive information, and event sentences and sentences containing a wh-object (Cruttenden, 1997; Wells, 2006).

The placement of nuclear stress is language specific. As a non-stress language (Selkirk & Shen, 1990), Chinese has less salient stress than English (Yu & Andruski, 2011) and tends to stress the final syllable of a word or phrase (Chao, 1979). Unlike English, Chinese relies more on syntax for focus marking, and prosody is only a supplementary means of focus marking (Xu, 2004). In Chinese, broad focus tends to be marked after the main verb or towards the end of a sentence (Chen, 1995) with no phonological manifestation (Xu, 2004), whereas narrow focus can be achieved either syntactically or phonologically (Xu, 2004). In other words, not all focuses in Chinese are realized through nuclear stress and narrow focus in Chinese is more likely to be realized through nuclear stress than broad focus. The following are two examples about how narrow focus is achieved in Chinese.

Example 7 -- 是谁 赢了 比赛？
Shishui yingle bisai?
Who won the game? (Who won the game?)
--是 我们 赢了 比赛。
Shi women yingle bisai.
It's we won the game. (It's we that won the game.)

Example 8 --谁 赢了 比赛？
Shui yingle bisai?
Who won the game? (Who won the game?)
--我们 赢了 比赛。
Women yingle bisai.
We won the game. (We won the game.)

Example 7 shows the syntactic marking of narrow focus, where the focus '我们 women' does not necessarily carry nuclear stress because there is the focus marker '是 shi'. In example 8, however, the focus '我们 women' carries nuclear stress, as the absence of the focus marker '是 shi' necessitates the prosodic marking of the focus.

The difference between the use of nuclear stress in English and Chinese often contributes to Chinese speakers' misplacement or misuse of nuclear stress in English. For example, they tend to assign nuclear stress to the final word or syllable in an utterance (Yu & Andruski, 2011), to every word in an utterance (Juffs, 1990) or even to pronouns (Deterding, 2006). Such deviations in their English may lead to communicative problems as native English speakers as well as other non-native English speakers may misinterpret their intended message.

2.2 Acoustic Realizations of Stress in English and Mandarin Chinese

The phonetic realization of nuclear stress in English has been widely investigated (Xu & Xu, 2005), including acoustic parameters such as F0 (pitch), duration, and intensity (Bolinger, 1986; Roach, 1991; Cruttenden, 1997; Pennington & Ellis, 2000; Chun, 2002; Ingels, 2011; Frost, 2011; Lu, Wang & de Silva, 2012). Acoustically, nuclear stress in English is indicated by a change in pitch height or pitch contour, a lengthening of the vocalic part in the stressed syllable, and an increase in intensity. It is 'generally accomplished by means of a co-occurrence of relatively extreme values of all three parameters' (Pennington & Ellis, 2000). A number of research suggests that pitch is the most indicative of nuclear stress in English, followed by duration and intensity (Lieberman, 1960; Roach, 1991; Cruttenden, 1997; Frost, 2011). However, there is also evidence suggesting a robust role of intensity in stress perception (Sluijter, van Heuven & Pacilly, 1997; Tamburini & Caini, 2005), and the co-dependent nature of duration to pitch increment (Bolinger, 1958; Ciszewski, 2012). In short, despite the disputes over the roles of phonetic cues to stress, a consensus is that pitch, duration and intensity are relevant cues and all contribute to English stress, but with decreasing importance (Roach, 1991).

Similarly, stress in Chinese is also realized through changes in pitch, duration and intensity. As a non-stress language, Chinese seldom marks focus with nuclear stress. When focus in Chinese is marked with stress (often contrastive stress for narrow focus), the pitch range of the element under focus is drastically expanded and that of the elements following the focus is greatly compressed (Shih, 1988; Xu, 1999; Yuan, 2004; Kabagema-Bilan, Lopez-Jimenez & Truckenbrodt, 2011), just as in English (Jin, 1996; Xu, 1999; Chen, 2003; Liu & Xu, 2005).

Unlike pitch, which has attracted wide attention, duration and intensity in Chinese stress have been relatively under-researched. Both Jin (1996) and Yuan (2004) report a lengthening of the syllable under stress and an increase in its intensity. Jin (1996) further claims that a stressed Chinese syllable is always longer but louder only in the sentence-final position. Likewise, Yuan (2004) found that syllable lengthening is especially salient for sentence-final stress, and the intensity of the stressed syllable is the highest and drops drastically thereafter. These findings are confirmed by Swerts and Krahmer (2004), who report that a stressed syllable in Chinese is the longest in sentence-final positions and that intensity rises and drops drastically after the stressed syllable.

Chen (2003) and Chen and Gussenhoven (2008), however, emphasize that the role of pitch at the sentence level is greatly weakened in Chinese. They found that the duration of word under contrastive focus is directly related with the degree of emphasis, yet pitch only varies in the focus and non-focus conditions but does not indicate the degree of emphasis.

In sum, previous research reveals that duration, intensity and pitch all contribute to sentence stress in Chinese, but with different importance.

2.3 Rapid Prosody Transcription (RPT)

Speakers of different languages perceive and interpret the acoustic parameters of pitch, duration, intensity and vowel quality differently in oral communication (Beckman, 1986; Low & Grabe, 1999; Pennington & Ellis, 2000). As a result of L1 influence, L2 learners tend to use cues to English stress in a different manner from its native speakers. Consequently, native speakers often find it hard to rely on prosody to interpret L2 learner speech (Gray, 2015; Ingels, 2011).

Studies on speech prosody have proposed and tested various methods to the evaluation of L2 learner's English. Among the most recent development, Rapid prosody transcription (RPT) (Cole et al., 2010, 2016) emerges as an effective method. It refers to assessing prosody by a group of naïve listeners (listeners with no phonetic or phonological knowledge) and the percentage of listeners who have assigned a prosodic feature (e.g., prominence or intonation boundary) to a certain word or position in an utterance will be the rating score for that feature.

RPT has been proven an effective method for marking prominence and intonation boundary in different languages with different transcribers. For example, Cole et al. (2010) and Cole et al. (2016) found RPT effective for marking prominence and intonation boundary in American English by American English speakers; Smith (2011, 2013) and Roux et al. (2016) found RPT effective for marking prominence and intonation boundary in French by native French speakers; Pintér1 et al. (2014) report that RPT ratings of L1 English by L1 and L2 English speakers were comparable; Smith and Edmunds (2013) report that L1 English speakers' RPT for L1 English and L2 English are both reliable. In addition, Smith (2009) compared native French speakers' RPT with expert transcription and found that their results are significantly correlated.

Previous findings have confirmed that naïve native speakers are able to make reliable judgements about both L1 and L2 prosody, so are naïve

L2 speakers about the target language prosody. However, it remains untested whether RPT can be applied with L2 speakers to assess L2 prosody, and whether there is a high degree of correspondence between L1 and L2 speakers' judgements. Variations in prosody across languages and L2 acquisition both suggest that naïve L1 and L2 speakers may differ in their assessment of L2 prosody. Therefore, this study aimed to answer the following research questions:

1. Do L1 and L2 English speakers yield comparable results when assessing nuclear stress produced by Chinese learners of English?
2. If there are discrepancies between the ratings by L1 and L2 English speakers, what acoustic cues contribute to these discrepancies?

3. Research Method

3.1 Participants

Speakers

Recordings of learner speech were from six English majors (1 male and 5 female) at a provincial university in central mainland China: three were first-year students (intermediate level English learners), and three were third-year students (advanced level English learners). They were between 18 to 22 years old. All speakers came from the same province and reported using Mandarin Chinese as their primary language in everyday communication.

Raters

The raters were 36 L2 English speakers and 30 L1 English speakers. The L2 English speaking raters (henceforth L2 raters) all spoke either Mandarin or Cantonese as their first language, had received postgraduate education related to English language (either in linguistics or literature), and had been studying or/and using English for over 15 years. These raters were between 22 to 45 years old. Ten were male and 26 were female. They were all highly proficient in English and reported using English frequently in their daily communication.

The L1 English-speaking raters (henceforth L1 raters) were all from the U.K. and spoke standard British English. They were between 23 to 50 years old. Twenty of them were male and 10 were female. None of them were fluent in Mandarin Chinese, though some had learned basic Chinese and could speak a little.

None of the raters reported having received systematic training in English prosody. The L2 raters participated on a voluntary basis, and the L1 raters each were paid 30 RMB yuan for their participation.

Expert

The first researcher served as an expert for nuclear stress rating. As a non-native English speaker, she had majored in English phonetics and phonology and had been systematically trained in English prosody. She had taught intermediate to advanced English learners at a Chinese university for over ten years and her own English proficiency was the highest at C2 Mastery for foreign language learners[1].

3.2 Stimuli

The stimuli included 30 sentences selected from the recording of a reading task done by each of the six participants chosen from two university classes, totaling 176 sentences (four sentences were of bad quality and thus excluded). The reading task was to assess the learners' mastery of nuclear stress and contained two parts: sentences in isolation and a dialogue. The dialogue and sentences were adapted from Wells (2006). The stimuli produced by each learner included 15 sentences in isolation and 15 in context (i.e., the dialogue) (See the appendix).

The 15 sentences in isolation represent all typical types of nuclear stress placement summed up in Wells (2006), including the default pattern (nuclear stress on the last content word) and 14 exceptions to the default pattern where nuclear stress does not fall on the last word in an utterance. These 14 exceptions (13 types) include: one event sentence, one wh-object sentence, two contrastive sentences (one long contrastive sentence broken into two parts), and 10 other sentences respectively ending with different components: a function word, an early stressed compound, a time adverbial, a reporting phrase, a parenthetical, an empty word, repeated information, a noun modifier, reflexive pronoun and a phrasal verb ending with a preposition.

[1] C2 Mastery is the highest among the six reference levels (A1-2, B1-2, C1-2) of language proficiency, according to *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment* by the Council of Europe (https://rm.coe.int/1680459f97) and *The Core Inventory for General English* by the British Council in 2017 (https://www.eaquals.org/resources/the-core-inventory-for-general-english/).

The 15 sentences taken from the dialogue represent some of the types, including five default pattern sentences, two contrastive sentences, two ending with a time adverbial, one ending with a function word, one ending with an empty word, one ending the an early-stressed compound, one ending with an early-stressed compound and a parenthetical, one ending with repeated information, and one ending with a post-modifier.

The task was designed in this way to assess if the participants had awareness of nuclear stress in English and if they could apply such awareness in context. However, this is not the focus of this study and the findings concerning these learners' awareness and application of nuclear stress in English is not reported here.

Three sentences produced by two native British English speakers (1 male, 1 female) were also included in the stimuli. The three sentences were all taken from the dialogue mentioned above. The recordings of these native speaker sentences were adopted from Wells (2006).

3.3 Procedure

The recordings of the learners' reading of the isolated sentences and the dialogue were firstly split into individual sentences. This yielded 180 sentence recordings (6 participants x 30 sentences), of which four were of bad quality and excluded. The 176 stimulus sentences were incorporated into 3 questionnaires designed on Qualtrics (www.qualtrics.com), each containing 60 sentences produced by two L2 English learners and the same three sentences produced by the two native speakers. In addition, questions about the rater's age, first language, and confidence level in rating were also included. For the L2 raters, information about their years of English learning and experience in English pronunciation learning was also elicited.

The questionnaires were distributed online to the target raters, who listened to the sentences individually and clicked on the word that they heard as the most prominent in each sentence. Eleven to 13 L2 raters and 10 L1 raters responded to each questionnaire. The expert rater rated all the 176 learner sentences. The recordings were randomized and rated twice by the expert rater with a two-week interval.

When all ratings were completed, the expert ratings were first compared and converted. Then the RPT results were converted and compared with the expert rating. Lastly, the acoustic cues contributing to their discrepancies were explored.

3.4 Data Analysis

First, we compared the ratings of the L1 and L2 raters. More specifically, we calculated the percentage of raters selecting a certain word as the most prominent for each word in each sentence recording. Then the percentage for each target word (the word supposed to carry nuclear stress according to theory, as underlined in the appendix) that was judged by the raters as carrying nuclear stress was converted to a numerical grade (0, 1, or 2) using the following coding scheme: Ratings higher than 60% were converted to 2, standing for good mastery of the nuclear stress production. Ratings lower than 60% but the highest in the sentence were converted to 1, representing partial mastery; ratings as the highest in the sentence but shared with other word(s) in the same sentence were also converted to 1. Other ratings lower than 60% were converted to 0, standing for non-mastery of the nuclear stress production.

Likewise, the expert ratings were also converted to 0, 1, and 2. A target word marked as carrying a nuclear stress in both expert ratings was given 2; that marked in one rating was given 1; and that not marked in either rating was given 0.

This conversion was necessary for direct comparison between the expert rating and the naïve raters' ratings. The expert rated each target word as 0 (not carrying nuclear stress) or 1 (carrying nuclear stress) in each round of rating, whereas the two groups of naïve raters' ratings for each target word were in percentage (the percent of naïve raters choosing the target word as the most prominent). Thus, it would be difficult to compare the numbers (0 or 1) with the percentages. The conversion of the ratings mentioned above was a solution to this problem and makes the comparison possible.

All three sets of scores, that is, scores from the expert, the L1 raters and the L2 raters, were compared using Kendall's tau correlation coefficients in SPSS 20.0 to assess the intra-rater and inter-rater reliabilities. This non-parametric statistic was chosen because not all of the three sets of scores were in normal distribution.

Secondly, we calculated the discrepancies (in percentage, non-converted) between the L1 and L2 raters. Then the 176 sentences were ranked ordered according to the degree of discrepancies (indexed by percentage scores). Sentences containing target words with L1-L2 discrepancies equal to or above 33.3% (meaning one third of the raters in each group were in disagreement) were identified for acoustic analysis. Likewise, sentences containing target words with high L1-L2 agreement (above 80% in both L1 and L2 ratings) were also selected. In total 20 sentences with great L1-L2 discrepancies and 20 sentences with high L1-L2 agreement were selected for acoustic analysis to explore further the relationship between L1 and L2 ratings.

The acoustic data collected included the following:

1) Duration: duration of the target words in the high-agreement sentences, that of words with ratings above 20% in the high-disagreement sentences, and also duration of the entire sentences. Duration ratio was then calculated by dividing word duration by sentence duration.
2) Fundamental Frequency (F0/pitch): F0 range for each word. Values of F0 peak, F0 valley, and mean F0 of the target words in the high-agreement sentences and of words with ratings above 20% in the high-disagreement sentences. F0 slope, calculated by dividing F0 range by word duration, and F0 ratio, calculated by dividing the mean F0 of each word by that of each sentence.
3) Intensity: intensity range for each word. Values of peak, valley, and mean of target words in the high-agreement sentences and of the words with ratings above 20% in the high-disagreement sentences. Intensity ratio, calculated by dividing the mean pitch of each word by that of each sentence.

All the calculations were done with raw values, and then z-normalized for cross-sentence and inter-speaker comparison. A series of Pearson product-moment coefficients were computed to explore the correlations between ratings and these cues. In addition, independent-samples *t*-tests were run to compare the acoustic characteristics of the exemplar sentences with high agreement with those of the sentences with high discrepancies.

4. Findings

4.1 Comparison between Ratings by Expert, L1 Raters and L2 Raters

To assess the reliability of RPT with naïve raters, recordings of three sentences read by native British English speakers were included in the rating task. Results showed that ratings for the three sentences were highly consistent, with 75%-95% of the L1 and L2 raters choosing the target words as the most prominent in each of these sentences. This high level of consistency among the raters on native production can serve as a bench mark against which different performances by the L2 English learners can be measured.

For the learners' recordings, ratings were less consistent, as expected. For some sentences, there was no agreement between L1 and L2 raters on the most prominent word, while for others their agreement could reach 100%.

Kendall's tau correlation coefficients were run to compare the two ratings by the expert for the 176 learner's sentences as well as the three sets of scores (expert rating, L1 speaker rating and L2 speaker rating) for these sentences. Results indicated that there was a strong positive correlation between the first expert rating and the second one (τ_b=.842, p<.001), representing high intra-rater reliability. For inter-rater reliability, there was a strong positive correlation between L1 and L2 ratings (τ_b=.610, p<.001), but a moderate positive correlation between the expert rating and the L2 rating (τ_b=.484, p<.001), and between the expert rating and the L1 rating (τ_b=.353, p<.001). The inter-rater reliability averaged at .482, which was moderate. Thus, the intra-rater reliability was higher than the inter-rater reliability.

4.2 Effects of Learner Proficiency

Intra- and inter-rater reliability for all three groups of raters is shown in Table 1. On average, the intra-rater reliability was high and the inter-rater reliability was moderate. However, both types of reliabilities varied as a function of talker, i.e., with the L2 learners' proficiency level. Specifically, higher degrees of reliabilities were obtained for higher proficiency learners (Talkers 4, 5, 6), and lower reliabilities for lower proficiency learners (Talkers 1, 2, 3). As shown in this table. The intra-rater reliabilities for Learners 1, 2, 3 varied from .70 to .942, which was in a lower range in comparison to those for Learners 4, 5, 6, varying from .801 to 1. The average inter-rater reliabilities followed the similar pattern: those for the lower proficiency learners were moderate (between .30 and .50) to high (above .50), and those for the higher proficiency learners varied at a higher range from .520 to .621. Besides, for all the six L2 English learners, inter-rater reliability was lower than intra-rater reliability. Among the six correlations between expert rating and L2 speaker rating, two were high at .843 (Learner 2) and .650 (Learner 6), one was low at .190 (Learner 1), and the rest three were moderate at .329 (Learner 3), .470 (Learner 4), and .464 (Learner 5). Among the six correlations between expert rating and L1 speaker rating, one was high at .613 (Learner 4), four were moderate at .450 (Learner 2), .349 (Learner 3), .386 (Learner 5), .440 (Learner 6), and one was low at .144 (Learner

1). Learner 1 was exceptional among all the learners. Her f0 contours extracted in Praat (Boersma & Weenink, 2018) were rather flat with little variation, which in part explains the greater disagreement between the ratings of the expert and of the two groups of naïve raters.

Learner	Intra-rater reliability		Inter-rater reliability Expert-L2		Expert-L1		L1-L2		Average
	τ_b	*p*	τ_b	*p*	τ_b	*p*	τ_b	*p*	
1	.700	.000	.190	.243	.144	.374	.690	.000	.341
2	.942	.000	.843	.000	.450	.007	.442	.006	.578
3	.810	.000	.329	.044	.349	.033	.547	.001	.408
4	.866	.000	.470	.008	.613	.001	.781	.000	.621
5	.801	.000	.464	.006	.386	.023	.710	.000	.520
6	1	.000	.650	.000	.440	.010	.632	.000	.574

L1: native English speaker raters; L2: non-native English speaker raters;
Average: the average of expert-L2, expert-L1 and L1-L2 correlations

Table 1 Intra-rater reliability and inter-rater reliability by learner

In addition, ratings by L1 and L2 raters agreed better than ratings by the expert and either group of naïve raters for all learners, except for Learner 2. Correlations between L1 and L2 speaker ratings for learners 1, 3, 4, 5, 6 were all high, at above .50. For Learner 2, the L1-L2 correlation was moderate at .442, which was the lowest among all L1-L2 correlations and much lower than the expert-L2 correlation of .843.

A Pearson product-moment coefficient was also run to test if sentence length affected judgement, as one may infer that longer sentences meant more challenges for raters as they would be faced with more choices. Results disputed such an inference (*r*=.134, *p*=.076). Therefore, it is safe to conclude that RPT ratings were not affected by sentence length.

4.3 Effects of Acoustic Cues

Although there were high correlations between L1 and L2 speaker ratings, the two groups of naïve raters disagreed greatly on 20 of the 176 sentences rated. On the other hand, the two groups agreed almost perfectly on another 20 sentences. These 40 sentences were chosen for acoustic analysis to uncover what may have led to the (mis)matching in perceptual judgement.

A series of Pearson product-moment coefficients were computed to explore the relations between acoustic cues (duration, f0, intensity) and L1 and L2 speaker ratings (in the original percentage). The results showed high correlations between duration and ratings, suggesting that the raters relied on temporal parameters in locating nuclear stress. Specifically, the correlation between word duration (z-normalised) and L2 speaker rating was moderate, *r*=.478, *p*<.001, between word duration and L1 speaker rating was strong, *r*=.505, *p*<.001, between duration ratio and L2 speaker rating was moderate, *r*=.471, *p*<.001, and between duration ratio and L1 speaker rating was also moderate, *r*=.498, *p*<.001.

Regarding intensity, no significant correlation was found for L1 ratings, suggesting that L1 raters did not rely on intensity to identify nuclear stress. L2 ratings, however, was slightly correlated with intensity, as indicated by a slight positive correlation between L2 speaker's rating and maximum intensity (*r*=.298, *p*=.018), intensity range (*r*=.272, *p*=.046), mean word intensity (*r*=.287, *p*=.023), and mean sentence intensity (*r*=.297, *p*=.018).

Surprisingly, we did not find any correlation between the naïve raters' ratings and f0 correlates including f0 peak, f0valley and mean, f0 slope and f0 range. This suggests that both groups of naïve raters did not rely on pitch variations for judging the placement of nuclear stress.

Next, a series of independent-samples *t*-tests were run to compare the acoustic cues (duration, f0, intensity) of the most rated words in the 40 sentences. For the 20 sentences with great L1-L2 rater discrepancy, all words with a rating of above 20% by at least one group were identified, yielding 48 words. The values of acoustic parameters of these 48 words were compared with those of the 20 target words in the 20 high-agreement sentences.

Results revealed that the two groups of words differed significantly in duration. The target words in the sentences with high agreement (*M*=0.70, *SD*=0.84) were significantly longer than the target words in the sentences with great discrepancy (*M*=-0.27, *SD*=0.93), *t*(66)=4.0, *p*<.001, *d*=1.09. Duration ratios confirmed that target words in sentences with high agreement (*M*=0.35, *SD*=0.08) were comparatively longer in their hosting sentences than those in the group of sentences with low agreement (*M*=0.24, *SD*=0.11), *t*(66)=3.66, *p*<.001, *d*=1.14.

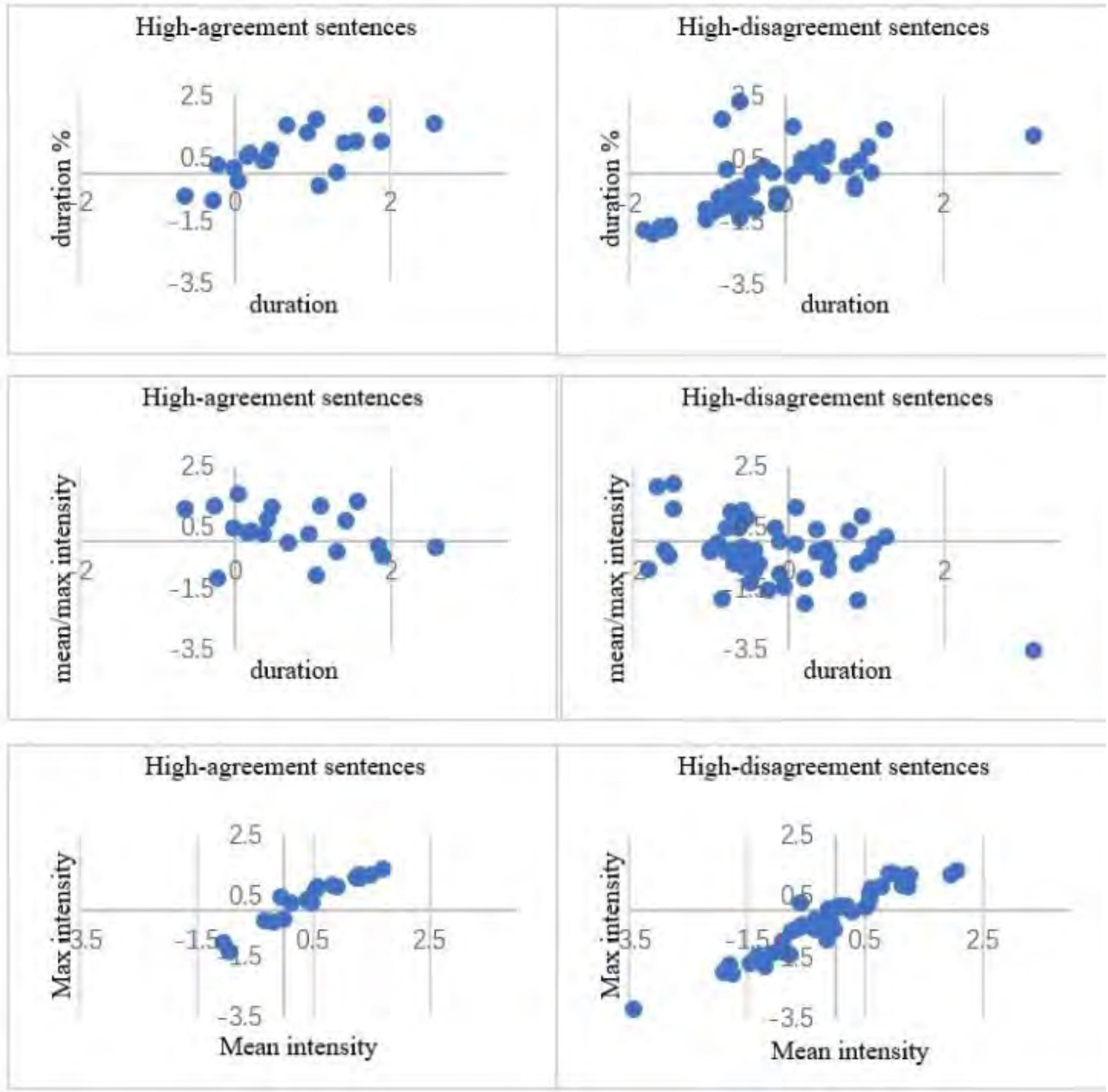

Figure 1. Scatterplots of z-normalised duration, duration ratio, maximum intensity, and mean intensity of the words in the two groups of sentences.

The two groups of words also differed significantly in intensity. More specifically, the target words in the high-agreement sentences had higher maximum intensity (*M*=0.51, *SD*=0.76) and higher mean intensity (*M*=0.45, *SD*=0.78) than the target words in the high-disagreement sentences (*M*=-0.21, *SD*=1.03), (*M*=-0.19, *SD*=1.03), *t*(66)=2.84, *p*=.005, *d*=0.80, *t*(66)=2.15, *p*=.055, *d*=0.70, respectively.

However, the two groups of words did not differ in any f0 dimensions. This echoes with the patterns in perceptual judgement where neither the L1 raters nor the L2 raters seemed to use pitch in their judgements.

These differences in acoustic parameters between the two groups of sentences are illustrated with the scatterplots in Figure 1. As shown in the scatterplots, most of the target words in the high-agreement sentences have a z-normalized duration and a z-normalized duration ratio of above 0, which means that they are longer than the mean word duration in the two groups and occupy a larger portion of the total sentence duration. In comparison, over half of the words with ratings above 20% in the high-disagreement sentences have a z-normalized duration and a z-normalized duration ratio of under 0, meaning they are shorter than the mean duration and take up a smaller portion of the entire sentence.

A similar pattern is found with mean intensity and maximum intensity of words. While the z value of the mean intensity and maximum intensity of most target words in the high-agreement sentences are above 0, that is, higher than the means for the two groups, those of most words in the high-disagreement group are below 0, lower than the means.

This means that when there were robust acoustic cues (duration and intensity, in this case), L1 and L2 raters found it easy to locate nuclear stress and therefore their ratings matched; when these acoustic cues were obscure, variations occurred in their perception and judgement.

5. Discussion

The moderate to high correlations between the expert score, the L1 rater score and the L2 rater score confirm the reliability of RPT among both native and non-native raters for assessing nuclear stress in L2 learner English. Naïve English speakers, native and non-native alike, can assess the nuclear stress in learner speech in a comparable manner.

However, RPT consistency across groups is affected by the learner's English proficiency level. Ratings by experts and by naïve raters were more reliable for high proficiency learners, but less so for low proficiency learners. For example, Learner 1 had the lowest intra-rater and inter-rater reliabilities. Acoustic analysis revealed few intonational fluctuations in her reading of the sentences and dialogue. Thus, the disagreement between the expert rating and naïve English speaker ratings can be attributed to the expert's awareness of and reliance on the acoustic cues associated with stress. While the presence of such cues made it easy for the expert rater to decide on nuclear stress, lack of robust cues may have posed a problem. The naïve raters, by contrast, were less explicitly aware of such roles of acoustic cues and their rating may have been more psychoacoustically-based. Therefore, they were less affected by the presence or absence of acoustic cues when making judgements about nuclear stress in a sentence. Based on this finding, the low agreement within RPT can be an indicator of an L2 English learner's poor mastery of nuclear stress.

The agreement between expert rating and the naïve raters' ratings, though not strong for all the six learners, lends support to Smith's (2009) finding that expert rating and L1 speaker rating are comparable. The agreement between L1 and L2 raters' performances echoes with Pintér1 et al.'s (2014) finding that RPT results for L1 English prosody by L1 and L2 raters are comparable, yet we have taken a step further by proving that RPT results for L2 English prosody by L1 and L2 raters are also comparable, at least to a certain extent.

Another major finding of our study is that both L1 raters and L2 raters relied on duration for assessing nuclear stress in L2 English learners' speech. This dependence on temporal cues for nuclear stress supports previous findings on the phonetic realization of stress in both English and Chinese (cf., Roach, 1991; Jin, 1996; Cruttenden, 1997; Yuan, 2004; Swerts & Krahmer, 2004).

However, apart from duration, L2 raters also relied heavily on intensity for the task, but L1 raters did not. Since the production and perception of stress are correlated yet independent, this difference can be justified from two perspectives. One possibility is that the learners produced nuclear stress with the same acoustic realizations as native English speakers, but L1 raters were not strongly sensitive to intensity because intensity is the least robust cue for stress in English, whereas L2 raters were more sensitive to it due to the important role of intensity for stress in Chinese. However, if this was the case, a question arises for the role of pitch variations in L1 ratings since pitch is the most important cue for stress in English.

The absence of the role of pitch in both groups' judgements suggests that the nuclear stress produced by the learners was acoustically different from that by native English speakers. The Chinese-speaking English learners may have relied on duration and intensity but not pitch to realize stress in their English speech, as many pronunciation teaching materials describe stressed English words as being longer and louder (e.g., Baker, 2009). Duration may have a far greater contribution to stress than intensity in these learners' English speech. Consequently, L1 raters relied only duration as a cue for nuclear stress location in these learners' speech.

In either case, there is evidence for L1 influence, either in the learners' or the raters' performance. The reliance on duration as a cue for stress by L2 learners and L2 raters and the absence of the role of pitch support Chen's (2003) and Chen and Gussenhoven's (2008) findings that unlike in English, duration is more important than pitch as a cue for stress in Chinese. The reliance on intensity echoes with previous research findings that intensity, together with duration, contributes greatly to stress in Chinese (Yuan 2004; Swerts & Krahmer 2004).

Given the absence of pitch as a cue in both L1 raters' and L2 raters' judgements of nuclear stress, it is highly likely that the L2 English learners did not make use of pitch to signal nuclear stress. This is worth L2 English teachers' attention. They need to raise their students' awareness of pitch as a cue for stress production to improve their production (and quite likely, their perception as well) of English stress.

6. Conclusion

This study adopted the Rapid Prosody Transcription (RPT) and acoustic analysis to examine production of nuclear stress in L2 English. Naive and expert raters who were L1 or L2 speakers of English provided perceptual assessments of stress placement, which was then correlated with acoustic findings to evaluate the robustness of phoetic cues to nuclear stress. Comparable ratings from L1 and L2 naive rater groups confirmed the reliability and effectiveness of RPT in assessing L2 speech prosody. Besides, correlation patterns between perceptual results and phonetic features revealed that L1 and L2 raters may rely on different acoustic cues in making perceptual judgements. The former group seemed to use duration only, while the latter deployed both duration and intensity in locating nuclear stress. The variation in perceptual reliance could be attributed to L1 raters' lack of sensitivity or L2 raters' sensitivity to certain cues in L2 English. Future research may further examine the perceptual reliance by increasing learner diversity such as recruiting L2 learners from various proficiency levels and language backgrounds. More diverse L2 production could also contribute to maximizing the potentials of RPT as an effective and reliable method to assess L2 prosody.

Acknowledgement: This study received support from the Research and Development Project EDB(LE)/P&R/EL/175/12, under the Language Fund by SCOLAR, Hong Kong S.A.R.

References

Altenberg, B. (1987). *Prosodic patterns in spoken English: Studies in the correlation between prosody and grammar for text-to-speech conversion*. Lund: Lund University Press.

Baker, A. (2009). *Ship or sheep? An intermediate pronunciation course* (3rd edition). Beijing: Beijing Language and Culture University Press.

Beckman, M. E. (1986). *Stress and non-stress accents*. Dordrecht, the Netherlands: Foris Publications.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program]. Version 6.0.39, retrieved 10 April 2018 from http://www.praat.org/

Bolinger, D. L. (1958). Intonation and grammar. *Language Learning, 8*(1), 31-37.

Bolinger, D. L. (1986). *Intonation and its parts: melody in spoken English*. London: Edward Arnold.

Chao, Y. R., (1979). *A grammar of spoken Chinese*. Bei Jing: China. Commercial press.

Chen, R. (1995). Communicative dynamism and word order in Mandarin Chinese. *Language Sciences*, *17*, 201-222.

Chen, Y. (2003). *The phonetics and phonology of contrastive focus in standard Chinese* (Unpublished PhD dissertation). Stony Brook: State University of New York.

Chen, Y., & Gussenhoven, C. (2008). Emphasis and tonal implementation in standard Chinese. *Journal of Phonetics*, *36*, 724-746.

Chun, D. M. (2002). *Discourse intonation in L2: From theory and research to practice.* Amsterdam: John Benjamins Publishing Company.

Ciszewski, T. (2012). Stressed vowel duration and phonemic length contrast. *Research in Language*, *10*(2), 201-214.

Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, *1*(2), 425-452.

Cole, J., & Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology*, *7*(1), 1-29.

Cruttenden, A. (1997). *Intonation* (2nd edition). Cambridge: Cambridge University Press.

Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.

Deterding, D. (2006). The pronunciation of English by speakers from China. *English World-Wide*, *27*(2), 175-198.

Dickerson, W. (1989). *Stress in the speech stream: The rhythm of spoken English*. Urbana, IL: University of Illinois Press.

Fouz-González, J. (2015). Trends and directions in computer-assisted pronunciation training. In J.A. Mompean & J. Fouz-Gonzalez (Eds.), *Investigating English pronunciation: Trends and directions* (pp. 174-195). Basingstoke, UK: Palgrave Macmillan.

Frost, D. (2011). Stress and cues to relative prominence in English and French: A perceptual study. *Journal of the International Phonetic Association*, *41*(1), 67-84.

Gray, M. (2015). Training L1 French learners to perceive prosodically marked focus in English. In J.A. Mompean & J. Fouz-Gonzalez (Eds.), *Investigating English pronunciation: Trends and directions* (pp. 174-195). Basingstoke, U.K.: Palgrave Macmillan.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201-223.

Ingels, S. A. (2011). *The effects of self-monitoring strategy use on the pronunciation of learners of English* (Unpublished PhD dissertation). Urbana-Champaign: University of Illinois.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics*, *23*(1), 83-103.

Jin, S. (1996). *An acoustic study of sentence stress in Mandarin Chinese* (Unpublished Ph.D. dissertation). Ohio: The Ohio State University.

Jones, D. (1956). *The pronunciation of English*. Cambridge: Cambridge University Press.

Juffs, A. (1990). Tone, syllable structure and interlanguage phonology: Chinese learners' stress errors. *International Review of Applied Linguistics*, 28(2), 99-117.

Kabagema-Bilan, E., Lopez-Jimenez, B., & Truckenbrodt, H. (2011). Multiple focus in Mandarin Chinese. *Lingua*, *121*, 1890-1905.

Ladd, D. R. (1980). *The structure of intonational meaning*. Bloomington, IN: Indiana University Press.

Levis, J. M. (1999). Intonation in theory and practice, revisited. *TESOL Quarterly*, *33*(1), 37-63.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustic Society of America*, *32,* 451-454.

Liu, F., & Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica*, *62,* 70-87.

Low, E. L., & Grabe, E. (1999). A contrastive study of prosody and lexical stress placement in Singapore English and British English. *Language and Speech*, *42*(1), 39-56.

Lu, J., Wang, R., & de Silva, L. C. (2012). Automatic stress exaggeration by prosody modification to assist language learners perceive sentence stress. *International Journal of Speech Technology*, *15*, 87-98.

Luchini, P. L. (2005). A new approach to teaching pronunciation: An exploratory case study. *The Journal of Asia TEFL*, *2*(2), 35-62.

Mennen, I. (2006). Phonetic and phonological influences in non-native intonation: An overview for language teachers. *Working paper WP-9*. Edinburgh, UK: QMUC Speech Science Research Centre.

Mennen, I. (2015). Beyond segments: Towards a L2 intonation learning theory. In E. Delais-Roussarie, M. Avanzi & S. Herment (Eds.), *Prosody and language in contact: L2 Acquisition, Attrition and Languages in Contact* (pp. 171-188). Heidelberg: Springer Verlag Berlin.

Pennington, M. C., & Ellis, N. C. (2000). Cantonese speakers' memory for English sentences with prosodic cues. *Modern Language Journal, 84*(3), 372-389.

Pintérl, G., Mizuguchi, S., & Tateishi, S. (2014). Perception of prosodic prominence and boundaries by L1 and L2 speakers of English. In *Proceedings of INTERSPEECH 2014* (pp. 544-547).

Roach, P. (1991). *English phonetics and phonology: A practical course* (2nd edition). Cambridge: Cambridge University Press.

Roux, G., Bertrand, R., Ghio, A., & Astésano, C. (2016). *Naïve listeners' perception of prominence and boundary in French spontaneous speech.* Paper presented at Speech Prosody 2016, Boston, MA, USA.

Selkirk, E., & Shen, T. (1990). Prosodic domains in Shanghai Chinese. In S. Inkelas & Z. Draga (Eds.), *The phonology-syntax connection* (pp. 313-338). Stanford: Stanford University.

Shih, C. (1988). Tone and intonation in Mandarin. *Working Papers of the Cornell Phonetics Laboratory, Number 3: Stress, Tone and Intonation* (pp. 83-109), Cornell University.

Sluijter, A. M. C., van Heuven, V. J., & Pacilly, J. J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of Acoustic Society of America*, *101*, 503-513.

Smith, C. (2009). Naïve listeners' perception of French prosody compared to the predictions of theoretical models. *Proceedings of IDP 09* (pp. 335-349).

Smith, C. (2011). *Perception of prominence and boundaries by naive French listeners.* Paper presented at ICPhS XVII, Hong Kong, 2011.

Smith, C. (2013). French listeners' perceptions of prominence and phrasing are differentially affected by instruction set. *Proceedings of Meetings on Acoustics* (pp. 1-7).

Smith, C., & Edmunds P. (2013). Native English listeners' perceptions of prosody in L1 and L2 reading. *Proceedings of INTERSPEECH 2013* (pp. 235-238).

Swerts, M., & Krahmer, E. (2004). Congruent and incongruent audiovisual cues to prominence. *Proceedings of Speech Prosody 2004* (pp. 69-72).

Tamburini, F., & Caini, C. (2005). An automatic system for detecting prosodic prominence in American English continuous speech. *International Journal of Speech Technology*, *8*(1), 33-44.

Wells, J. (2006). *English intonation: an introduction*. Cambridge: Cambridge University Press.

Xu, L. (2004). Manifestation of information focus. *Lingua*, *114* (3), 277-299.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, *27,* 55-105.

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, *33,* 159-197.

Yu, V. Y., & Andruski, J. E. (2011). The effect of language experience on perception of stress typicality in English nouns and verbs. *The Metrical Lexicon*, *6*(2), 275-301.

Yuan J. (2004). *Intonation in Mandarin Chinese: Acoustics, perception, and computational modeling* (Unpublished Ph.D. dissertation). Cornell University.

Appendix: Stimuli Sentences (Adapted from Wells (2006)

The underlined are the syllables that tend to carry nuclear stress in the sentences. The types of sentence are indicated in parentheses.

Sentences in isolation

1 I've just received a letter from her. (Ending with a function word)
2 You've told me what Emma wants, (Contrastive sentence Part 1)
3 what do you want? (Contrastive sentence Part 2)
4 I'm going to buy a new mobile phone. (Ending with an early-stressed compound)
5 Shall we walk to the restaurant? (Default pattern)
6 I'd prefer to go on foot. (Ending with repeated information)
7 You're looking rather pleased with yourself. (Ending with a reflexive pronoun)
8 How are you doing, he asked. (Ending with a reporting phrase)
9 I'll see you on Tuesday, then. (Ending with a parenthetical)
10 Let's go back to my place. (Ending with an empty word)
11 There's a mosquito on your finger. (Ending with a place adverbial)
12 What are you looking at? (Ending with a phrasal verb with a preposition)
13 Look at the tie he's wearing. (Ending with a noun modifier)
14 There's a train coming. (Event sentence)
15 Which route did you take? (Wh-object sentence)

Sentences in context

16 Are you planning to go away this year? (Ending with a time adverbial)
17 We've just been away. (Contrastive sentence)
18 We had a week in Cornwall. (Default pattern)
19 How was it? (Ending with a function word)
20 We had a marvelous time. (Ending with an empty word)
21 The only problem was the weather. (Default pattern)
22 It rained most of the time. (Ending with a time adverbial)
23 What did you do during all this rain? (Ending with repeated information)
24 The best thing we did was to go to the Eden Project. (Ending with an early-stressed compound)
25 What's that? (Default pattern)
26 It's a museum of ecology. (Default pattern)
27 I found it utterly fascinating. (Default pattern)
28 It's more like a theme park really. (Ending with an early-stressed compound and a parenthetical)
29 There's lots to do. (Ending with a noun modifier)
30 The children loved it (too). (Contrastive sentence)