Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Methodological influences on word stress identification: Implications for research and teaching

Mahdi Duris[1], *John M. Levis*[1], *Reza Neiriz*[1], *Alif O. Silpachai*[2]

[1]Iowa State University, Department of English
[2]Radboud University, Centre for Language Studies

mduris@iastate.edu, jlevis@iastate.edu, rneiriz@iastate.edu, alif.silpachai@ru.nl

Abstract

Accurate word stress influences intelligibility for L2 and L1 English speakers, making it important in language assessment and perception training. However, reliably rating word stress in English is difficult even for experts because stress is signaled by four possible acoustic correlates (pitch, duration, intensity, and vowel quality), which are not present in all spoken words. Multiple cues mean that judgments of word stress may involve embedded decisions, leading to varied levels of agreement in published studies.

To investigate the influence of methodological decisions on word stress judgments, we employed two approaches to stress identification. Three phonetically-trained expert listeners rated stress placement by 10 Chinese L1 speakers of English who read 100 multisyllabic English words (2-6 syllables), 1,000 words in total. In the first approach, raters identified whether each word was correctly stressed with embedded decisions justifying their answers. In the second, listeners made a series of binary decisions about stress placement, syllable count, and vowel quality. Rater scores resulted in agreement levels among all three listeners of as little as 41% (for Approach 1) to 91.6% (for primary stress placement in Approach 2), showing that ratings of word stress are sensitive to construct definitions.

Index Terms: methodological innovation, word stress, improvement of rating, assessment

1. Introduction

English word stress provides "islands of reliability" [1, 2] for listeners in identifying words in the stream of speech [3]. This reliability is important not only for L1 English listeners but also for listeners for whom English is an additional language [2, 4]. Accurate word stress also influences the intelligibility of L2 and L1 English speakers in both academic contexts and for pronunciation in general [4, 5]. Word stress has also been increasingly a focus of research in assessment [6, 7], English as a Lingua Franca [8], and perception training [9], with evidence that word stress deviations can affect understanding for both L1 and L2 listeners [2, 10]. Despite the importance of word stress for intelligibility, it is not clear that native English speakers are consistent in evaluating stress. In one troubling example, naive students were asked by the second author to evaluate the accuracy of word stress in transcriptions. They were relatively poor in identifying stress errors, especially in two- and three-syllable words without obvious vowel deviations. Instead of noticing stress errors, they seemed to assume the stress was fine if they could understand the word [11].

In research, reliably rating word stress in English is difficult because it is signaled by four possible acoustic correlates (pitch, duration, intensity, and vowel quality), which are not always available to listeners in all spoken words [12]. In addition, although English L1 listeners can make use of the suprasegmental cues of pitch, duration, and loudness if they have to [13], these are not their favored cues. Instead, they primarily rely on the segmental cue of vowel quality to make decisions about accurate stress placement [14].

In addition, the listener's language background also affects word stress judgments. Although L1 English listeners prefer to evaluate word stress based on vowel quality, L2 Dutch listeners are better at using suprasegmental features to judge stress placement than L1 English listeners [15]. In other cases, it has been reported that certain groups of L2 listeners may be unable to reliably recognize stress placement, the so-called stress deafness phenomenon [16]. The availability of multiple cues also means that direct judgments of word stress accuracy involve embedded decisions. Embedded decisions here refer to the reasons for correctness decisions and, therefore, cognitive demands and strategies in judging word stress. Asking listeners (even experts) to identify stress may result in low reliability across judges. To ensure sufficient agreement between raters, previous studies have most often judged stress based on interrater reliability, but such studies report varied levels of agreement (e.g., .70 in [17] and .88 in [18]).

The importance of vowel quality in decisions about stress placement led us to investigate whether a mispronunciation detection (MPD) system built for segmentals could be extended to be used for word stress identification. In other words, if the system could reliably distinguish between full and reduced vowels, it could also provide information about whether stress was accurate. In this first step to employing our MPD system for stress, we explored whether expert raters were accurate in identifying stress placement. We compared two approaches to word stress identification. The first approach (the embedded decision approach) asked raters to classify each word's stress placement by classifying it for accuracy first and then providing a reason for their classification. The second approach asked raters to evaluate stress with multiple separate binary decisions (e.g., correct primary stress, correct vowel quality, correct number of syllables).

2. Methodology

In two experiments, three phonetically-trained expert listeners rated the word stress placement and predicted intelligibility of 100 words read by each of 10 Chinese L1 speakers of English. Raters were asked to make embedded decisions or binary decisions about whether the stress was correct or incorrect. Embedded and binary decisions asked about vowel quality, potential intelligibility issues, and syllable count, which were chosen due to their potential importance for the MPD system.

10.21437/SpeechProsody.2024-148

2.1. Participants

Ten L1 Chinese speakers from a US Midwestern university participated in the study, with the majority identifying their L1 as Mandarin and no specific dialect pronunciation. Self-reported dialects spoken by the others included Cantonese, Minnan, and Wu. The average age for the participants was 27.5 (*SD* = 4.22), ranging from 22 to 34 years old (4 female and 6 male). All participants were paid for their participation.

2.2. Materials

2.2.1. Stimuli

A total of 100 multisyllabic words were read by each participant, producing 1,000 tokens. To test for a complete range of English stressed words, tokens included words of two to five+ syllables. The stress position in words was varied to account for all possible syllable stress positions, although as words included more syllables, it was not always possible to identify words with final stress. Figure 1 provides pattern examples combining the number of syllables, the stressed syllable position, and the total number of tokens.

	Stressed Syllable					
# of syllables	1st	2nd	3rd	4th	5th	Total Token
2	COMfort	conSENT	X	X	X	17
3	BEAUtiful	aMAzing	picturESQUE	X	X	34
4	HOnorable	conDItional	unmisTAken	nevertheLESS	X	42
5+	IMplementation	conCEPtualizing	diffeRENtiation	adminiSTRAtively	N/A	7

Figure 1: *Examples of stress patterns when considering the number of syllables*

Words included at least one full vowel in stressed position, and some included multiple full vowels; some pairs involved shifted stress items based on the same root word (e.g., igNORE vs. IGnorance). Control for word frequency was not established, but word familiarity was rated by the participants but not reported in this paper. All participants were asked to rate their familiarity with the word they produced through two 5-point Likert scale prompts (familiarity results are not reported in this paper), with the first asking how confident participants were about the meaning of the word (1 = not confident, 5 = very confident) and the second if they had heard the word before the experiment (1 = very uncertain, 5 = very certain).

2.2.2. Data Collection Interface

To capture the word recordings from participants, a web interface was designed in NodeJS to present all stimuli in one session. Each word was presented and recorded one at a time. Participants could start and stop recording of the word and listen to their recording as many times as they wished. Once they were satisfied, they would click "Next" to rate their familiarity with the word using two 5-point Likert scale options. Once the ratings were completed, the participant would again click "Next" to move on to the next word. During this procedure, a research assistant ensured that the participants were moving through the recordings and the ratings as expected from a drafted protocol. Any problems with the recordings or ratings were remediated immediately.

2.3. Raters

Three phonetically trained listeners involved in the research study conducted stress judgments for each word. Rater one (R1) is a native speaker of American English, while R2 and R3 are both L2 speakers of English (R2 = Thai L1, R3 = French L1) with native-like production of spoken English. Before carrying out individual stress judgments, all listeners rated a subset of 20 words to establish norms for categorizing word stress errors. Then, all raters discussed and adjusted the judgment criteria before rating all the items. Rater 1 rated 100% of the stimuli, while R2 and R3 covered 50% of the tokens each. All raters used the same interface, which presented each spoken token as a clickable WAV file and judgment options to complete the stress judgments. To avoid rater fatigue, all raters were instructed to conduct their judgments for a maximum of 60 minutes at a time.

3. Experiment 1: Embedded decisions

The rating categories for Experiment 1 were divided into two main types (correctly stressed vs. incorrectly stressed). From these two stress conditions, further judgments characterized the conditions in which the words were judged for stress placement. The correctly stressed category involved two sub-judgments, *Correct 1* and *Correct 2*. For *Correct 1*, the word was heard as correctly stressed, and all segments were considered intelligible. For *Correct 2*, the stress was correct, but the word could be considered by raters as potentially unintelligible because of non-stress pronunciation deviations. Based on discussion about reasons for stress errors, the raters considered four possible incorrect categories in applying stress to an English word. For *Wrong 1*, the stress was incorrect, but all vowels were correctly pronounced, for example, *CONcentrate* was said as *concenTRATE*. *Wrong 2* involved incorrect stress and incorrectly pronounced vowels. *Wrong 3* errors involved equal stress on multiple syllables. *Wrong 4* words had an incorrect syllable number, even if the expected syllable was stressed.

Figure 2 shows an example of a stress rating for participant 100 for the word "substance." The rater listened to the participant's production and saw information about the participant's familiarity with the word. In this example, the rater selected *Wrong 4* because the word sounded like "subastance."

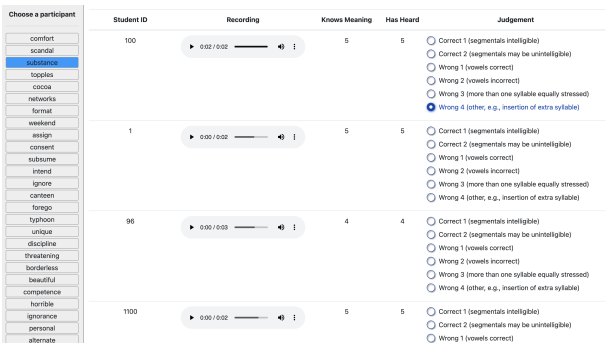

Figure 2: *Rating interface for embedded decisions. The rater is presented with the audio token for each participant with their self-reported familiarity rating of the word. Wrong 4 has been selected for this token.*

3.1. Results

All ratings for Experiment 1 were compiled by order of participants (n=10) with their respective ratings from R1, R2, and R3. Each rating category was assigned a numerical value from 1 to 6. For example, ratings for *Correct 1* were marked as 1, while ratings for *Wrong 4* were marked as 6. An agreement analysis covered six rating levels using simple frequency statistics. The first rating level indicates the agreement between raters at the categorical level, whether the item's stress is correct or incorrect. Out of 2,000 combined ratings by R1, R2, and R3 (R1=1,000; R2=500; R3=500), the categorical agreement level (Correct vs. Incorrect) reached 70%. At the rater level, R1 and R2 had a slightly higher agreement level at 73% compared to 67% for ratings between R1 and R3. Considering a Quadratic Weighted Kappa (QWK) analysis, the agnostic rater and item level reaches only 0.37, a fair agreement according to [19].

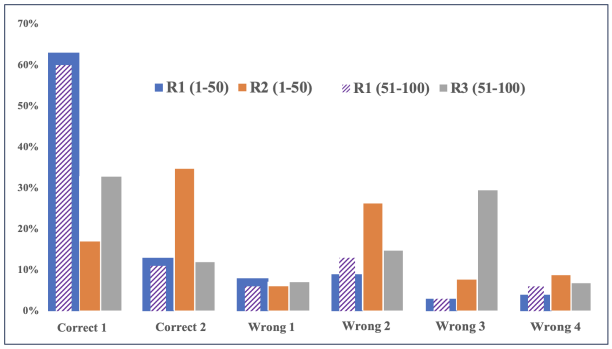

Figure 3: *Experiment 1 agreement between R1R2 and R1R3*

Further analysis of the agreement between raters at the sub-judgment level shows lower levels of agreement. Agreement for ratings based on all six categories (*Correct 1, 2*; *Wrong 1, 2, 3, 4*) was 41% at a rater-agnostic level, indicating that reliability for embedded decisions was low and that the raters were not able to agree on reasons the stress patterns were correct or incorrect. Figure 3 provides details of the stress ratings between R1, R2, and R3. Specifically, R1 (the English native speaker) rated Items 1-50 for the 10 L2 English speakers at a higher proportion for the *Correct 1* category (63%), indicating that most items were judged to have the correct stress, with all segmentals being intelligible, despite minor variations in vowel quality realization. In contrast, R2 rated *Correct 1* for the same items at 17%, indicating that evaluating both stress and segmental quality involved different criteria from those employed by R1. For R2, most items fell in the *Correct 2* (35%) and *Wrong 2* (25%) categories.

R1 and R3's ratings on items 51-100 show similar patterns for *Correct 1* but not for the other categories, suggesting that the L1 of the raters may also have played a role in the ratings of some categories. Most items (60%) were judged as *Correct 1* by Rater 1, while R3 rated *Correct 1* for just 33%. Additionally, R3 assigned *Wrong 3* to another third (27%) of the spoken items, while R1 only assigned 3% for the same sub-judgment category (Figure 3). Evaluating both stress accuracy and reasons for accuracy judgments resulted in weak agreement levels. The Quadratic Weighted Kappa shows a slightly higher agreement between R1 and R2 (0.38) compared to R1 and R3 (0.36). Overall, R1 showed consistency in rating all items, while R2 and R3 displayed broader variance in assigning stress judgment.

3.2. Experiment 1 Discussion

Asking raters, even expert raters who are trying to follow the same criteria, to rate the accuracy of word stress and to evaluate other phonological features for accuracy or intelligibility at the same time seems destined for poor agreement between raters. Reliability in even a basic decision about correctness seemed to be lowered because the raters were making multiple decisions at the same time. This indicates that rating word stress accuracy, even for words in isolation, can be made more difficult by also attending to other phonological or perceptual features.

It may be that there were too many decisions for raters to make at the same time. Even though the three raters were involved in creating the criteria for rating, and they normed their ratings and discussed together why they made their decisions, the task proved to be overly complex, leading to unexpectedly low agreement for all six categories and higher but still marginal agreement (∼70%, similar to that of de Jong et al. [17]) when taking into account only a binary decision of correct or incorrect. It may be that even this agreement level was suppressed because of the cognitive load of making additional decisions to sub-classify the primary decision on accuracy.

4. Experiment 2: Binary decision ratings

To control for decisions being made in a single rating session, a second experiment focused on making stress judgments using binary judgments for decisions about stress. First, the rating categories were divided into five questions for which raters needed to decide between "No" and "Yes" for every spoken stress token. Figure 4 shows all five categories, with the first selected, "Primary Stress on Correct Syllable?" indicating that the rater needed to complete the other four categories in order. This sequential process allowed raters to focus on one type of rating at a time for each of the tokens. (Note: Not all binary decisions were specifically about word stress accuracy. The criterion, "All segmentals are likely intelligible" was included to test an unrelated research question.)

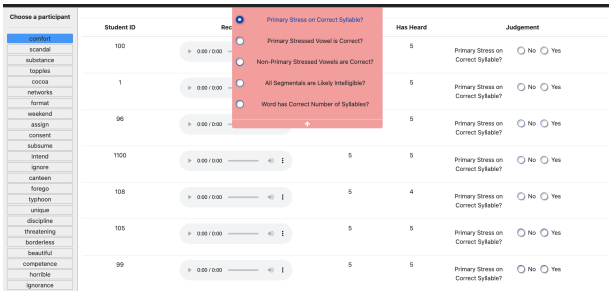

Figure 4: *Rating interface involving binary decisions*

Like Experiment 1, all raters used an interface that presented key information about the tokens being judged for stress and an input option for their judgments. As seen in Figure 4, the raters would first select the word to be judged in the far left column. In this example, the word "comfort" is selected, and all 10 recorded instances of the word were presented on the page. The "student ID" is displayed in Column 2, while the recordings of each participant are available in Column 3.

Columns 4 and 5 displayed the participants' knowledge of the word, as seen previously in Figure 2, about their knowledge of the meaning of the word and if they had "heard" the word before. The last column, "Judgment," displayed the rating

question being judged along with a binary option to be selected (i.e., yes or no). The same information was displayed for each of the five rating categories. Here, the first category is selected, where raters evaluated whether the primary stress was on the correct syllable. The same rating expectations were followed as in experiment 1, with raters 2 and 3 (i.e., R2 and R3) completing half of all tokens and R1 judging 100% of the items. Raters stopped for a short break after 60 consecutive judgments.

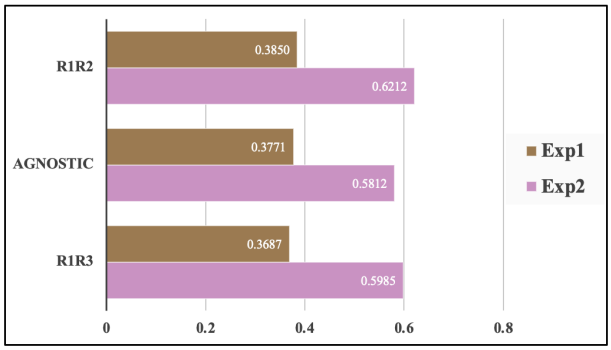

Figure 5: *Interrater reliability (QWK) results for R1R2, rater agnostic, and R1R3 across both experiments*

4.1. Results

For Experiment 2, each rater evaluated five categories using a dichotomous choice (i.e., Yes or No). Each category (henceforth Q1-5) gave R2 and R3 500 tokens to rate (1,000 for R1). These 500 tokens were the same as the 10 Chinese L2 English speakers in Experiment 1 (10 participants $\times$ 50 items), resulting in 10,000 ratings, 2,000 for each of the five rating categories (distributed overall as R1 = 5,000, R2 = 2,500, R3 = 2,500). Some ratings were lost during multiple data server transfers needed to handle the WAV files (mostly from R1's judgments), accounting for 1.89% of the total data from 10,000 ratings.

Overall, at a rater-agnostic level (R1-R2 and R1-R3), when asked if the "Primary Stress [is] on [the] Correct Syllable?" (Q1), raters agreed at 91.59% over 963 tokens. When rating for "Word has [the] Correct Number of Syllables?" (Q5), raters agreed at 95.38% (985 ratings). For ratings considering stress and intelligibility, raters agreed to "All Segmentals are Likely Intelligible?" (Q4) at 87.03%. The last two ratings, Q2 and Q3, received the lowest ratings agreement, which pertained to stress and vowel quality. For "[the] Primary Stressed Vowel is Correct?" (Q2), raters agreed at 83.03%, while "Non-Primary Stressed Vowels are Correct?" reached an agreement at 75%. The rater and item agnostic analysis using Quadratic Weighted Kappa (Figure 5) shows a moderate agreement level of 0.58, with R1 and R2 having higher agreement (0.62) than R1 and R3 (0.59). When comparing across raters, the Q5 category received the highest agreement across all items (1 through 100), followed by Q1. Table 1 details all agreements across raters for items 1-100.

4.2. Experiment 2 Discussion

The much-improved agreement levels in Experiment 2, in which each binary decision was made separately, indicate that simplifying the task resulted in much better agreement levels. Two of the criteria (Q1 & Q5) had agreements above .9, one (Q4) above .8, and two (Q2 & Q3) slightly lower than that.

Table 1: *Experiment 2 agreement levels between raters where the percentage represents the categorical agreement between raters on similar items*

	R1 & R2 (Items 1-50)	**R1 & R3 (Items 51-100)**
Q1: Primary Stress on Correct Syllable?	93.12%	89.96%
Q2: Primary Stressed Vowel is Correct?	78.43%	87.40%
Q3: Non-Primary Stressed Vowels are Correct?	75%	73.17%
Q4: All Segmentals are Likely Intelligible?	85.89%	88.15%
Q5: Word has Correct Number of Syllables?	97.12%	93.59%

All of these agreement levels were above the binary decision between correct and incorrect in Experiment 1. It is striking that even though vowel quality is the primary criterion used by native English speakers to decide if a syllable is stressed [14], decisions about vowel quality were associated with lower agreement levels between raters. This may be due to two of the raters being near-native speakers of English from different language families (Romance vs. Tai-Kadai). Although it would be almost impossible to recognize a foreign accent in their spoken language, their L1 perceptual systems are likely different from that of R1, the native English-speaking rater, and from each other. The lowest agreements for vowels in non-primary stressed positions may have occurred because, in 3+ syllable words, raters were asked to evaluate multiple vowels. If any vowel was incorrect in any way, this meant the answer would have to be "No." In contrast, the decision for primary-stressed vowels always involved only one vowel. Again, this suggests that the simpler decision is the better one.

5. Conclusion

Word stress is central to comprehensibility assessment [6], making the reliable identification of stress errors essential to any research or teaching endeavor. To use an MPD system's ability to evaluate vowel accuracy in identifying likely word stress errors, a first step may be to evaluate only the primary stressed vowel in a multisyllabic word. If correct, the word would not be flagged for stress. If incorrect, the system could call attention to stress by giving feedback that the vowel should be pronounced strongly with a certain quality.

For teaching and learning, it is critical to attend to the most important cues for stress identification in order to scaffold learning to identify stress more effectively. Finally, for research into the effects of word stress on intelligibility, we must have agreed-upon ways to measure the accuracy of stress production. Simpler criteria are better. Using binary decisions to evaluate stress without embedding secondary decisions into the process appears to be a much better approach than assuming that raters know how cues are weighted in evaluations of stress placement.

6. Acknowledgements

This study was funded by the National Science Foundation grant 2016984.

7. References

[1] H. Dechert, "Individual variation in speech," *Second language productions*, pp. 156–185, 1984.

[2] J. Field, "Intelligibility and the listener: The role of lexical stress," *TESOL quarterly*, vol. 39, no. 3, pp. 399–423, 2005.

[3] A. Cutler and A. Jesse, "Word stress in speech perception," *The handbook of speech perception*, pp. 239–265, 2021.

[4] M. Ghosh and J. M. Levis, "Vowel quality and direction of stress shift in a predictive model explaining the varying impact of misplaced word stress: Evidence from english," *Frontiers in Communication*, vol. 6, p. 628780, 2021.

[5] M. Benrabah, "Word-stress: A source of unintelligibility in english," *IRAL. International review of applied linguistics in language teaching*, vol. 35, no. 3, pp. 157–165, 1997.

[6] T. Isaacs and P. Trofimovich, "Deconstructing comprehensibility: Identifying the linguistic influences on listeners' l2 comprehensibility ratings," *Studies in Second Language Acquisition*, vol. 34, no. 3, pp. 475–505, 2012.

[7] K. Saito, K. Macmillan, M. Kachlicka, T. Kunihara, and N. Minematsu, "Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies," *Studies in Second Language Acquisition*, vol. 45, no. 1, pp. 234–263, 2023.

[8] C. Lewis, "Word stress in english as a lingua franca: Evidence from asean interactions," *Unpublished doctoral dissertation). Universiti Brunei Darussalam*, 2023.

[9] L. Sippel and I. A. Martin, "Immediate and long-term improvement in lexical stress perception: the role of teacher and peer feedback," *International Review of Applied Linguistics in Language Teaching*, vol. 61, no. 3, pp. 1173–1195, 2023.

[10] M. G. Richards, "Not all word stress errors are created equal: Validating an english word stress error gravity hierarchy," Ph.D. dissertation, Iowa State University, 2016.

[11] J. Levis and G. M. Levis, "This is how a gondolier gallops"–pronunciation and unintelligibility in international teaching assistant (ita) presentations." presented at Pronunciation in Second Language Learning and Teaching, University of Utah, Salt Lake City, UT, 2017.

[12] A. Cutler, "Lexical stress," in *The Handbook of Speech Perception*, D. Pisoni and R. Remez, Eds. Wiley Blackwell, 2005, pp. 265–289.

[13] Y. Zhang and A. Francis, "The weighting of vowel quality in native and non-native listeners' perception of english lexical stress," *Journal of Phonetics*, vol. 38, no. 2, pp. 260–271, 2010.

[14] A. Cutler, "Lexical stress in english pronunciation," in *The Handbook of English Pronunciation*, M. Reed and J. Levis, Eds. John Wiley & Sons, Ltd, 2015, ch. 6, pp. 106–124.

[15] L. Bruggeman, J. Yu, and A. Cutler, "Listener adjustment of stress cue use to fit language vocabulary structure," in *Speech Prosody 2022*, 2022, pp. 264–267.

[16] S. Peperkamp and E. Dupoux, "A typological study of stress 'deafness'," *Laboratory phonology*, vol. 7, pp. 203–240, 2002.

[17] N. H. De Jong, M. P. Steinel, A. F. Florijn, R. Schoonen, and J. H. Hulstijn, "Facets of speaking proficiency," *Studies in Second Language Acquisition*, vol. 34, no. 1, pp. 5–34, 2012.

[18] M. W. Tanner and M. M. Landon, "The effects of computer-assisted pronunciation readings on esl learners' use of pausing, stress, intonation, and overall comprehensibility," *Language Learning & Technology*, vol. 13, no. 3, pp. 51–65, 2009.

[19] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.