PERCEPTUALLY-MOTIVATED ASSESSMENT OF AUTOMATICALLY DETECTED LEXICAL STRESS IN L2 LEARNERS' SPEECH

Kun Li and Helen Meng

Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
Shun Hing Institute of Advanced Engineering
The Chinese University of Hong Kong, Hong Kong SAR, China
{kli, hmmeng}@se.cuhk.edu.hk

ABSTRACT

This paper presents a method of automatic lexical stress assessment for L2 English speech. Syllable stress can be labeled at three levels – primary (P), secondary (S) and no (N) stress, but secondary stress may vary among word pronunciations within and across accents and present difficulties for human perception. Hence, evaluation of lexical stress based on all three levels (i.e., the P-S-N criterion which requires that all syllables in a word must be correctly classified in terms of stress) may be too strict, and we may consider relaxing it to either the P-N or A-P-N criterion – the former only requires the correct placement of primary stress, while the latter relaxes further to allow for confusion between primary and secondary stress. An automatic syllable stress detector is applied to L2 learners' speech. Its output for all the syllables in a word is evaluated in terms of the P-S-N, P-N or A-P-N criterion. Comparisons between automatic and manual assessments of lexical stress patterns suggests that the A-P-N criterion can strike a good balance between accommodating variability and screening out problematic patterns, giving an average word accuracy of 79.6%.

Index Terms— stress assessment, stress perception, stress detection

1. INTRODUCTION

Suprasegmental phonology plays an important role in the perceived proficiency of the second language (L2) spoken by a learner [1]. Lexical stress is associated with syllable prominence in a word. Faithful production of lexical stress is important for the perceived proficiency of L2 English. In some cases, it also serves to disambiguate lexical terms by proper placement of primary stress, e.g., "*'insert*" vs. "*in'sert*". This paper focuses on the assessment of lexical stress in a word, i.e. evaluate the stress patterns of L2 learners' speech as generally right or wrong.

Previous research mainly focused on automatic stress detection, i.e. to identify the syllable carrying Primary Stress (PS), Secondary Stress (SS), or no stress (NS) at all. In the study of syllable stress detection for German and Italian, Tepperman [2] used the mean values of fundamental frequency (F0), syllable nucleus duration, energy and other features related to F0 slope and RMS energy range. Imoto [3] developed Hidden Markov Models to detect stress in English sentences read by Japanese students. Tamburini [4] combined the detection of lexical stress and pitch accents into a task of prominence detection. Stress detection was based on syllable nucleus duration and high-frequency features. In this paper, we make use of the automatic lexical stress detector in [5] for syllable stress classification, whose syllable-based accuracy is 78.6% (in classification of PS, SS and NS) or 89.8% (in determining the presence or absence of PS).

Few research has been conducted in perceptually-motivated evaluation of lexical stress, especially in the L2 learners' perception. In [6], a perceptual test conducted with 58 Mandarin speakers. 21 words covering different stress patterns were recorded by a native American English speaker, and were presented to each speaker. The speakers had poor performance in the perceptual test.

For assessment of lexical stress in L2 English speech by a computer-assisted pronunciation training (CAPT) system, we may start in the identification of syllables with PS, SS, or NS. The results of lexical stress detection can be compared with the model stress patterns of dictionaries. Due to the following challenges, this method is not considered to be good.

(1) For certain words, dictionaries may give different lexical stress patterns, e.g., *"re'frige,rator"* in US English vs. *"re'frigerator"* in British English [7].

(2) Inaccuracies in automatic lexical stress detection will affect automatic lexical stress assessment. Assuming the syllable-based accuracy of a lexical stress detector is 80%, the word-based accuracy for 4-syllabic words will be decreased to about 40% ($0.8^4 \approx 0.4$).

(3) Sometimes even humans may not be able to correctly identify the stress patterns in speech with high accuracy – a previous perceptual test [6] conducted with 58 Mandarin speakers who were phonetically trained showed that only 36% of the speakers could correctly identify the whole stress patterns in native English speech.

To design a better scheme to assess the lexical stress of L2 learners' speech, we first study the results of human perception of lexical stress patterns in native US English speech. These results are used in the design of automatic lexical stress evaluation of non-native English speech.

2. HUMAN PERCEPTION OF LEXICAL STRESS IN NATIVE US ENGLISH

This perceptual test aims to elicit human perception of different lexical stress patterns. As shown in Table 1, 30 words are selected to cover a variety of stress patterns. Bi-syllabic words are excluded due to their simplicity in our study. We present stress patterns for both US and British English, with reference to [7, 8]. A native US English speaker was invited to record in a natural speaking style – hence US English patterns are treated as canonical patterns in this work.

Table 1: Lexical stress patterns in US and British English.

	US	**British**	Words
3	● – –	● – –	hospital, processing
	– ● –	– ● –	department, tomorrow
	● – ○	● – –	autograph
4	● – – –	● – – –	admirable
	● – ○ –	● – ○ –	millisecond, motorcycle
		● – – –	activator, elevator
	– ● – –	– ● – –	available, experience
	– ● – ○	– ● – –	accumulate, facilitate
	○ – ● –	– – ● –	transportation
	○ – – ● ○ ○ – ●	○ – – ●	misunderstand
5	– ● – ○ –	– ● – ○ –	refrigerator
	○ – ● – –	○ – ● – –	transformational, unacceptable
		– – ● – –	anniversary, interchangeable
	○ – – ● –	○ – – ● –	documentation
	– ○ – ● –	– ○ – ● –	examination, experimental pronunciation
		– ○ – ● – ○ – – ● –	participation
6	○ – ● – – –	○ – ● – – –	intellectually, unambiguously
	○ – – ● – –	○ – – ● – –	eligibility
	○ – – – ● –	○ – – – ● –	characterization

Note: ‘●’ *denotes primary stress,* ‘○’ *secondary stress and* ‘–’ *unstressed.*

We recruited three groups of listeners: 58 listeners with mother tongue in Mandarin (ML), 25 in Cantonese (CL) and 25 in US English (EL). The recordings were played word by word. For a given word, each subject was asked to mark “1” under the syllable with primary stress; “2” under the syllable with secondary stress; or check under “I don’t know” if he/she does not know the stress position(s) of the word. An excerpt of the questionnaire is shown in Table 2.

Table 2: An excerpt of the questionnaire in perceptual test.

Words	Syllables				I don’t know
elevator	e	le	va	tor	
	1		2		

2.1. Preliminary analysis

Fig. 1 shows the results of human perception (with ML, CL and EL) of speech recorded in native US English. Note that a word is considered correct if its entire stress pattern is correct – including the primary (P), secondary (S) and no (N) stress in syllables, referred as the **P-S-N** criterion.

We observe that stress identification accuracies decrease dramatically as the syllable length of the word increases. This is probably because more stress patterns are possible for longer words. The overall average accuracies in ML, CL and EL are only at 37%, 25% and 21% respectively.

In particular, for words with five to six syllables, the Cantonese and even native US English listeners achieve less than 10% identification accuracies. This may be due to the long syllable length of the words: 70% syllable-based accuracy is only equivalent to about 10% word-based accuracy ($0.7^6 \approx 0.12$). And it also suggest that the P-S-N criterion is too harsh for evaluating the lexical stress patterns.

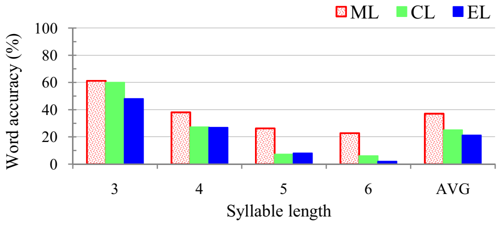

Fig. 1: Average stress identification accuracies for words with different syllable lengths. The overall averages are shown by the rightmost bars.

2.2. Alternative criteria

Detailed analysis of Table 1 shows that among the 30 words studied, 8 of them have different stress patterns between US and British English, e.g., “*’auto,graph’* vs. “*’autograph’*, “*,transport’tation”* vs. “*transpor’tation”*. Even within the same accent type, a word may have more than one stress patterns, e.g., “*,misunder’stand”* vs. “*,mis,under’stand”* in US English, and “*par,tici’pation”* vs. “*,partici’pation”* in British English. This indicates that we cannot assess an L2 learner’s articulation of word stress pattern by a straightforward comparison with a single dictionary entry. Another important point to note from Table 1 is that variations in the stress patterns of a given word are primarily due to the presence or placement of secondary stress. Hence we propose to relax the P-S-N criterion as follows:

(1) the **P-N** criterion: a word is considered correct if the subject gives correct placement of the primary stress.

(2) the **A-P-N** criterion: a word is considered correct if the subject gives correct placement of the primary stress; or the subject places the primary stress in a secondary-stress-carrying syllable and places the secondary stress in a primary-stress-carrying syllable. This is referred to as the adjusted P-N criterion (A-P-N).

Table 3: Illustration of applying different criteria to different stress patterns with respect to the word “*’milli,second*”.

	P-S-N	P-N	A-P-N
● – ○ –	✓	✓	✓
● – – –	×	✓	✓
● – – ○	×	✓	✓
○ – ● –	×	×	✓
– – ● –	×	×	×
– ● – ○	×	×	×

Table 3 illustrates the results of different stress patterns with respect to an example of “*millisecond*” under different criteria. Our objective is to come up with a criterion that human perception finds acceptable, which can be used later for automatic assessment. We present an analysis of the perceptual test results in the following.

2.3. Analysis of results from human perceptual test

We group the words into 3 categories: words with a **single stressed** syllable, words with primary stress followed by secondary stress (**PS + SS**), words with secondary stress followed by primary stress (**SS + PS**).

2.3.1. Words with a single stressed syllable

There are 7 words with a single stressed syllable, e.g., *"hospital"*, *"processing"*, etc. Table 4 shows the accuracies based on the P-S-N and P-N criteria. In words with only one stressed syllable, the A-P-N criterion is equivalent to the P-N criterion. Our observations indicate that even for words with a single stressed syllable, the identification accuracies under the P-S-N criterion are low. The P-N criterion seems to give more reasonable assessments based on the perception across all three groups of listeners. It is also interesting to note that the American listeners got lower scores under the P-S-N criterion; while under the P-N criterion, the performance of the three groups was more consistent.

Table 4: Average stress identification accuracies for words with a single stressed syllable.

	ML	CL	EL
P-S-N	65.8 ± 6.0	58.9 ± 18.6	46.3 ± 14.3
P-N	83.3 ± 8.1	86.9 ± 14.3	78.9 ± 13.1

2.3.2. Words with PS + SS patterns

These words include *"autograph"*, *"millisecond"*, etc. Among these eight words, five do not carry secondary stress based on British English (see Table 1). Results in Table 5 indicate that about 70% of the listeners gave correct placement of PS, but only 30% could correctly identify stress pattern of the entire word (P-S-N). This means that 40% of the listeners could not give correct placement of secondary stress, but this is actually "acceptable" based on the British pronunciation of some of the words.

Table 5: Average stress identification accuracies of words with primary stress followed by secondary stress (PS + SS).

	ML	CL	EL
P-S-N	31.3 ± 8.8	26.0 ± 11.4	26.5 ± 9.1
P-N	70.5 ± 8.9	68.0 ± 22.6	71.0 ± 10.9
A-P-N	78.2 ± 7.8	74.0 ± 20.7	75.5 ± 10.8

2.3.3. Words with SS + PS patterns

These words include *"transportation"*, *"misunderstand"*, etc., 15 words in total. We note from Table 6 that PS placement was correct for less than half of the words perceived by the Mandarin listeners, and for about one fifth of the words perceived by Cantonese and American listeners. The differences between the accuracies under the A-P-N criterion and the P-N criterion suggests that confusion between PS and SS occurred in about 25% of the words as perceived by Mandarin listeners, and in about 40% of the words as perceived by Cantonese and native American listeners.

Table 6: Average stress identification accuracies of words with secondary stress followed by primary stress (SS+PS).

	ML	CL	EL
P-S-N	24.5 ± 9.8	5.1 ± 4.0	6.1 ± 5.4
P-N	47.5 ± 15.0	18.7 ± 11.0	20.5 ± 5.4
A-P-N	73.2 ± 14.4	56.8 ± 16.4	58.5 ± 10.7

2.4. Key findings

The key findings in this section indicate that even for native US English speech, listeners are unable to identify syllable stress with ease. As explained in Section 2.2, lexical stress for a given English word may vary both within and across accent types. Humans (including native listeners) may have difficulty identifying the lexical stress patterns of a word uttered by a native speaker, especially if the word involves secondary stress. Results from the perceptual tests indicate that stress identification accuracies decrease dramatically as the number of syllables in a word increases. For words with five or more syllables, listeners almost always wrongly identify the stress patterns under the P-S-N criterion. Subjects perform better in general under the P-N criterion. This implies that listeners find it hard to identify secondary stress accurately. Furthermore, for words with SS followed by PS, listeners show difficulty even in placing the PS. The A-P-N criterion relaxes this constraint and achieves higher consistency among the three subject groups.

3. PERCEPTUALLY-MOTIVATED AUTOMATIC STRESS ASSESSMENT OF L2 ENGLISH SPEECH

In this section, we migrate to a more challenging analysis of stress in non-native (L2) English speech, as compared with native English speech in the previous section(s). Our goal is to investigate how we may perform automatic lexical stress assessment of L2 English speech in the context of computer-aided pronunciation training (CAPT) applications. We refer to the evaluation of a word-level stress pattern as lexical stress assessment, which is different from the identification of a single syllable's stress level (at primary/secondary/no stress).

3.1. L2 English corpus with manual lexical stress assessment

Our experiments are based on a subset of a suprasegmental corpus in [9], which contains English speech recording from 100 Mandarin speakers and 100 Cantonese speakers (both groups are gender-balanced). Each speaker utters 28 words, which results in 5,600 words in total. A trained linguist labeled all syllables in the corpus with PS/SS/NS, and another trained linguist rated all words with a lexical stress assessment score as follows:

Score 4 : Near native (*Pass*).
– Distinct stress assigned to the correct syllables, explicit distinction among PS, SS and NS.

Score 3 : Acceptable (*Pass*).
– PS is assigned to the correct syllable.
– PS is assigned to the syllable carrying secondary stress

and SS is assigned to the syllable carrying primary stress, but the difference between PS and SS is not prominent.

Score 2 : Unclear (*Fail*).

– PS is assigned to the secondary stress syllable, while: *a*) no SS is assigned to any syllable; *b*) SS is assigned to the syllable not carrying primary stress; *c*) SS is assigned to the syllable with primary lexical stress, but the difference between SS and NS is not prominent.

Score 1 : Wrong (*Fail*).

- PS is assigned to the syllable not carrying stress.

The above rating system is based on the observation in Section 2: humans may have difficulty identifying the entire lexical stress patterns of a word uttered by a native speaker, especially if the word involves secondary stress; while under the A-P-N criterion, subjects perform much better.

No consideration is given to segmental correctness. Words were ignored if they contain serious segmental errors, which results in 5,480 word tokens for investigation. Essentially, a word is perceived as largely correct if it receives Score 3 or 4 from the annotator. On the other hand, it is regarded as generally problematic with Score 1 or 2.

Table 7 shows the performance of applying different criteria to "automatically" assess the annotated lexical stress patterns. It shows that the P-N and A-P-N criteria are much better than the P-S-N criterion: the accuracies for the P-S-N, P-N and A-P-N criterion are about 73%, 95% and 90%, respectly. This means that we may apply the P-N or A-P-N criterion to automatically assess the detected lexical stress patterns.

Table 7: Performance of applying different criteria to assess the annotated lexical stress patterns.

Sys.	P-S-N		P-N		A-P-N	
Ann.	✓	×	✓	×	✓	×
Pass	2,981	1,473	4,185	269	4,427	27
Fail	17	1,009	31	995	499	527

3.2. Automatic syllable stress detector

We have previously developed a syllable stress detector [5] for L2 English learners' speech. The detector extracts three prosodic features of each syllable – syllable nucleus duration, maximum loudness [10] and differential pitch value. Then a prominence model is applied to these three prosodic features. The detector was trained and tested on the same corpus described in Section 3.1. Performance evaluation based on 10-fold cross-validation shows that its syllable-based accuracy is 78.6% (in classification of PS, SS and NS) or 89.8% (in determining the presence or absence of PS).

3.3. Automatic lexical stress assessment

We ran our automatic syllable stress detector on our L2 English corpus, as described in Section 3.1. We then perform automatic lexical stress assessment for each word by analyzing it in terms of the P-S-N, P-N or A-P-N criterion. Results are shown in Table 8 and Table 9.

Results show that using the P-S-N criterion to assess the L2 learners' lexical stress only achieves an accuracy of about 44%; while using the P-N or A-P-N criterion can achieve an accuracy of about 77%, or about 80% respectly. For the A-P-N criterion, its F-measure is about 86%, which outperforms the F-measure of the P-N criterion by about 5%. Note that the rate of words manually assessed as *pass* is about 81%, i.e. the amount of words manually assessed as *pass* is much larger than that of words manually assessed as *fail*.

Table 8: Performance of applying different criteria to assess the detected lexical stress patterns.

Sys.	P-S-N		P-N		A-P-N	
Ann.	✓	×	✓	×	✓	×
Pass	1,411	3,043	3,326	1,128	3,531	923
Fail	42	984	139	887	195	831

Table 9: Accuracies, recall rates, precisions and F-measures of using different criteria to automatically assess the recognized lexical stress patterns.

	Accuracy	Recall	Precision	F-measure
P-S-N	43.70%	31.68%	97.11%	47.77%
P-N	76.88%	74.67%	95.99%	81.00%
A-P-N	79.60%	79.28%	94.77%	86.33%
All	81.28%	100.00%	81.28%	89.67%

4. CONCLUSIONS

This paper presents an investigation in automatic lexical stress assessment for L2 English speech. We describe the challenges in automatic lexical stress assessor, due to variations of stress patterns for a given word within and across accents, as well as inexactness in human perception of lexical stress even in native English speech, especially for secondary stress. Consequently, it seems impractical to require an L2 learner to produce the exact stress pattern of a word based on the P-S-N (primary/secondary/no) stress criterion. Hence we propose to relax the P-S-N criterion to either the P-N or A-P-N criterion – the former only requires correct primary stress placement, while the latter relaxes further to allow for confusion between primary and secondary stress. We developed an automatic syllable stress detector that can be further developed for automatic lexical stress (word-level) assessment. This can be achieved by taking the outputs of the automatic syllable stress detector for every syllable in a test word, and examining the overall word-level stress pattern under one of the P-S-N/P-N/A-P-N criteria. Results show that the A-P-N criterion can strike a good balance between accommodating variability and screening out problematic patterns, giving an average word accuracy of 79.6%.

5. ACKNOWLEDGEMENTS

The work is jointly supported by the Shun Hing Institute of Advanced Engineering and the NSFC/RGC Joint Research Scheme (Project No. N_CUHK 414/09).

6. REFERENCES

[1] J. Anderson-Hsieh, R. Johnson and K. Koehler, "The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody and syllable structure," *Language Learning*, vol. 42, pp. 529–555, 1992.

[2] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," in *Proc. of ICASSP*, 2005.

[3] K. Imoto, Y. Tsubota, A. Raux, T. Kawahara and M. Dantsuji, "Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system," in *Proc. of ICSLP*, 2002.

[4] F. Tamburini, "Prosodic prominence detection in speech," in *Proc. of Signal Processing and its Applications*, 2003.

[5] K. Li, S. Zhang, M Li, W. Lo and H. Meng, "Prominence model for prosodic features in automatic lexical stress and pitch accent detection," in *Proc. of INTERSPEECH*, 2011.

[6] S. Zhang, K. Li, W. Lo and H. Meng, "Perception of English suprasegmental features by non-native Chinese learners," in *Proc. of Int. Conf. on Speech Prosody*, 2010.

[7] *Oxford English Dictionary*, http://www.oed.com/.

[8] *Merriam-Webster Dictionary*, http://www.merriam-webster.com/.

[9] M. Li, S. Zhang, K. Li, A. Harrison, W. Lo and H. Meng, "Design and collection of an L2 English corpus with a suprasegmental focus for Chinese learners of English," in *Proc. of ICPhS*, 2011.

[10] E. Zwicker and H. Fastl, *Psychoacoustics Facts and Models 2nd Updated Edition*, Springer, 1999.