Avoidance of Stress Clash in Perception of Conversational American English

Amelia E. Kimball, Jennifer Cole
Department of Linguistics and Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign, Urbana, Illinois, United States

akimbal2@illinois.edu; jscole@illinois.edu

Abstract

We examine evidence for a regularity bias in the perception of sentence-level stress patterns, asking to what degree listeners perceive speech as *metrically regular*, with few or no occurrences of stress clash. We assess regularity through a stress perception task carried out by untrained listeners annotating transcripts of recorded speech, with sentences designed to have regular stress, and sentences drawn from a corpus of spontaneous conversational speech. Results show listeners report perceiving fewer stress clashes than predicted by random placement of stresses or by concatenating the citation form stress patterns of each individual word in a given sentence, though some incidence of stress clash is reported for both the regular and irregular speech materials. These findings suggest that listeners perceive English speech in accordance with a weak regularity bias. Inter-transcriber agreement rates also reveal substantial disagreement in perceived stress patterns at the sentence level, for regular and irregular sentences alike, suggesting variability in the perception of acoustic cues to stress at these levels.

Index Terms: stress clash, meter, stress perception, rhythm, metrical regularity

1. Introduction

Metrical patterns in speech arise from the sequencing of stressed and unstressed syllables across words and phrases [1]. In English, regular patterns occur when words are sequenced such that stressed syllables occur at regular intervals. For instance, the sentence 'HEIdi SOMEtimes SAW the JUry LEAving' has the stress pattern [SW SW S W SW SW], with a recurring pattern of stressed (strong) and unstressed (weak) syllables. However, due to the frequency of mono-syllabic words and the variety of patterns of word-level stress in English, metrically irregular phrases and sentences are also very common, e.g., 'JILL LIKES to SKI CAREfully' with the stress pattern [S S W S SWW]. To avoid confusion with notions of isochrony based on acoustic duration (discussed below), here we use the terms regular and irregular to refer to *metrical patterns of phonological stress* in sentences.

Despite the fact that sentences and phrases in English are not necessarily—or even typically—regular in their stress patterning, there is evidence that listeners are biased to perceive stress in terms of such regular patterns, and that more generally, regular stress patterns are privileged in speech processing. Early evidence for a regularity bias comes from studies of English phrasal stress. In phrases where stress clash results from the sequencing of word-level stresses, (e.g. *thirTEEN MEN*), speakers have the option of resolving the clash in favor of an alternating stress pattern (*THIRteen MEN),* or a pattern with only a single stress (*thirteen MEN*). When asked to identify the stressed syllables in such instances, listeners report hearing the alternating, clash-free pattern (THIRteen MEN) when listening to the entire intact sentence [2,3], but do not reliably perceive the resolved stress (THIRteen or thirteen) in the first word in the sequence when it is extracted and presented by itself [3]. These findings suggest that English speakers are biased to perceive stress patterns in phrases or sentences as regular, even when the acoustic evidence is not particularly strong.

The bias for listeners to perceive phrasal stress as regular may reflect a more general bias for regular stress patterns in speech processing. Studies on the perceptual processing of speech show that sentences with regular stress patterns yield faster and more accurate phoneme and word recognition [4-6]. In addition, ERP studies have shown that regularity modulates the amplitude of the n400 response [7], suggesting that speech perception and semantic integration are made easier by predictable, alternating stress patterns. In speech production, strings of non-words with regular stress patterns are easier to produce than irregularly patterned strings of the same non-words [8]. These experimental results point to a basis for a regularity bias in the mechanisms of speech processing, e.g., in neural oscillatory processes, as claimed by [9].

The studies cited above investigated stress regularity in speech production and perception with experimenter-controlled, read speech materials. This leaves us to wonder about the production and perception of speech that is produced under more natural conditions, e.g., conversational speech. This paper represents an initial step in the investigation of the effects of stress regularity in everyday speech, with an experimental approach combining experimenter-designed sentences read aloud by a model speaker with speech samples from a corpus of conversational speech that are re-enacted by the same speaker. Our focus in this paper is on **perceived** stress patterns. The goal is to compare the observed patterning of listeners' reported stress perception to random placement of stresses and to word stresses as reported in the dictionary. If there is a bias towards regularity, we expect observed patterns will be more regular than predicted patterns.

Note that in this paper metrical regularity is defined as an alternating pattern of strong and weak syllables. This is distinct from temporal measures of regularity as defined over acoustic intervals [e.g., 10,11]. In other words, in this paper we are interested in whether listeners report adjacent stressed syllables, regardless of when these syllables happen in time. Acoustic measures of our sample are not within the scope of the present paper, but are the subject of our ongoing investigation.

2. Experiment

This experiment tests the hypothesis that listeners' perception of stress patterns in naturally occurring sentences of English will be biased towards regular patterns. Specifically, we predict that listeners will report fewer instances of stress clash (stress on adjacent syllables) than are expected based on the location of stress within each content word, and also fewer than expected by three other calculations of chance occurrence, described further below. Listeners' perception of stress is assessed through a beat annotation task presented in a web survey format.

2.1. Stimuli

The test materials consisted of twenty sentences of conversational speech from the Buckeye Corpus of Conversational American English [12]. Also, twenty

10.21437/SpeechProsody.2014-86

sentences designed to have regular stress patterns, selected from previous experiments by the first author and from published studies, were included in this experiment as a control condition where no stress clash is expected to be perceived. To minimize the effects of speaker-dependent variability in speech rate, in patterns of phonetic reduction, or in other aspects of the phonetic realization of stress, all speech materials used in this study were re-enacted by a model speaker who is a native speaker of American English trained in linguistics, but who had no knowledge of the research goals of this study. A text transcript of the speech excerpts was presented to the model speaker, and the speaker was instructed to repeat the utterances in a natural and conversational style.

The form of the Buckeye corpus is informal sociolinguistic interviews with 40 speakers in the Columbus, Ohio area. Buckeye sentences were taken from interviews with four different speakers, chosen randomly from the larger corpus. Sentences were selected by the first author and chosen for the absence of disfluencies or major internal prosodic phrase breaks. For each of the four interviews the first five prosodically-demarcated utterances of the target duration and with no significant internal prosodic breaks were taken. Excerpts were approximately 1-2 seconds in duration. The 20 Buckeye excerpts taken together consisted of 159 words with a total of 205 syllables. The sentences with regular stress patterning (prepared by the experimenter) consisted of 140 words with a total of 198 syllables.

The regular sentences conformed to three metrical patterns: trochaic (SWSW…), iambic (WSWS…), or dactylic (SWWSWW…). The model speaker read all sentences of one stress pattern together. Examples of each sentence type are listed in table 1 below

Buckeye	My grandmother's from Ireland I go to Northland high school right now
Trochaic	Read a bedtime story. Heidi sometimes saw the jury leaving.
Iambic	Michelle foresees mistakes. My shoes are beige and black.
Dactylic	Sally is hoping to travel to Canada. Thomas has already taken geography.

Table 1. *Stress patterns for sample sentences*

2.2. Participants

55 participants total (31=Female) were recruited on Amazon Mechanical Turk, an online marketplace for human intelligence tasks. Participants' age ranged from 19 to 55 (mean =32.5, s.d.=8.8). Built-in Mechanical Turk screening tools ensured that the posting was displayed only to those workers who reported their location as in the United States. Only data from the 48 native English speakers with no reported hearing problems were eligible for inclusion in this study.

2.3. Procedure

Participants were presented with a display via an online survey built with Qualtrics survey tools [13]. Instructions stated that they would listen to a series of sentences and should "mark the beats" in a sentence by checking boxes. They listened to an example sentence ("I like to run and jump") and wrote the last word of that sentence in an answer box, in order to confirm that their audio was working. All participants correctly identified the last word in the example sentence. Participants were then shown an example of checked boxes for that sentence, and told "this person thinks the beats are on *like, run*, and *jump*." An example of the user interface is below in Figure 1.

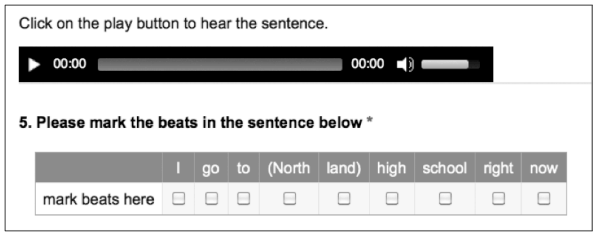

Figure 1: *User interface for the experiment.*

Participants were explicitly told to mark beats as the speaker said them in the audio recording, and not how the participant would say them. They were told that if they were unsure, they should "make their best guess", and that experimenters expected that listeners "may differ in where they mark beats." The listeners could play the audio as many times as they wanted, and were required to check at least one box. They were told that if they could not hear the audio they should check all the boxes. Two catch questions were presented in which the speech was purposefully obscured. Data was analyzed only from the 46 participants who (in addition to reporting as native speakers with no hearing problems) followed directions and identified the catch questions by marking all boxes.

3. Results

This experiment tested the hypothesis that listeners have a regularity bias in the perception of stress in conversational speech. If this hypothesis is true, we expect that when asked to mark stressed syllables in a corpus, listeners will report fewer clashes than predicted by chance or than are predicted by the concatenated stresses from the dictionary entry for each individual word. Results showed 7 out of 8 predictions of clash frequency were greater than the observed rate of clash, meaning that listeners hear fewer clashes than predicted.

3.1. Clash measures

For both the regular and Buckeye sentences five different frequency counts of the number of clashes were conducted: the observed number of clashes, plus four measures of the expected number of clashes.

Observed clashes were calculated based on participant responses. For every *n* sequential checkmarks, *n*-1 clashes were counted, such that four beats in a row would be marked as three clashes. This same counting method was used for the expected clashes.

The first rate of expected clashes was calculated based on the rate at which an individual participant marked syllables as beats for each sentence. For example, if a participant marked 2 beats out of every 4 syllables in a given sentence, their rate for that sentence was 2/4. This rate was squared to get the probability that two adjacent syllables would be marked as beats (a clash) and then the rate was multiplied by the number of adjacent syllable pairs in a given sentence, which represents the number of clashes possible for that sentence. Equation 1 below shows the method. Where C is the number of checks and N is the number of syllables:

$$\left(\frac{C}{N}\right)^2 (N-1) = \text{Number of expected clashes} \quad (1)$$

The second measure of expected clashes was created through a random sampling simulation using the R statistical computing language [14]. Separate simulations were run for dummy sentences with the same number of syllables as the test sentences, and for all possible numbers of beat marks within those sentences. 10,000 trails were run for each sentence length. Each trial looped through the total number of syllables, randomly selecting either a check or no check for each syllable, without replacement. Clashes were counted for each trial, and a mean clash occurrence was calculated across trials and compared with observed values.

The third measure of expected clashes was the number of clashes predicted by the dictionary entry of the citation form of the word. This was done by marking the primary stress of each word as reported in the Oxford English Dictionary and then counting occurrences of adjacent stressed syllables. Though this method was expected to overestimate because of the preponderance of monosyllabic words, it was included nonetheless because all words, including function words, may be variably stressed in conversational speech.

The last measure of expected clashes was also calculated based on the stresses marked in the dictionary for each word, but this time marking only the stressed syllables of content words as beats, with no beats marked on function words.

Figure 2 below compares the total clashes counted by each measure across sentences. Clashes are reported in clashes per syllable, to normalize for the differing number of syllables in the two samples.

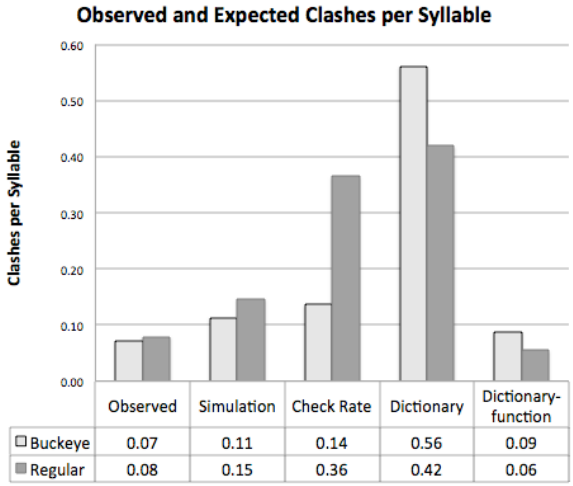

Figure 2: *Observed vs. expected clashes per syllable in both regular and conversational sentences*

Paired sample t-tests comparing the observed number of clashes vs. expected number of clashes under each calculation showed that for both samples the observed number of clashes was statistically significantly smaller than the expected number of clashes based on participant check rate, simulation, or the dictionary with and without stressing function words. The dictionary prediction that includes function words proved to be an especially poor model, grossly overestimating the amount of observed clashes in both regular and conversational sentences. However, the expected values based on dictionary stress markings *not including function words* provided the closest approximation of participants' responses, though it still differed significantly from the observed clash rate—in the case of the *regular* sentences, dictionary stress without function words predicted *fewer* clashes than observed, in the case of *Buckeye* sentences dictionary stress without function words predicted *more* clashes than observed. Table 2 below lists the *t* statistic, degrees of freedom, mean of the differences, and *p* value for each measure of expected clash frequency as compared to observed clash frequency.

REGULAR	t	df	mean of the differences	p
Check rate	-54.4861	919	-2.827624	<.001
Simulation	-24.1348	919	-0.6640634	<.001
Dictionary	-28.7383	919	-3.365217	<.001
Dictionary- function	3.9725	919	0.2347826	<.001

BUCKEYE	t	df	mean of the differences	p
Check rate	-28.0077	919	-0.6749942	<.001
Simulation	-18.1389	919	-0.4359521	<.001
Dictionary	-48.3472	919	-5.026087	<.001
Dictionary- function	-3.532	919	-0.176087	<.001

Table 2. *Results of paired sample t-tests comparing observed and expected clash rates. Negative mean differences indicate that expected measures were higher than observed.*

3.2. Agreement

In addition to a comparison of observed clashes to expected clashes, we also calculated the inter-transcriber agreement for each syllable. Each syllable received a rating based on the number of listeners that marked it as a beat. Agreement scores are proportions ranging from 0 to 1. If listeners were in total agreement, the distribution of these scores would be bimodal, with peaks at 0 (meaning many syllables were marked by no listeners) and 1(meaning many syllables were marked by all listeners). Instead, as is clear from figure 3 below, the distribution was spread within the conversational Buckeye sentences. Relatively few syllables were marked by a majority of listeners, though many syllables were left unchecked by all listeners.

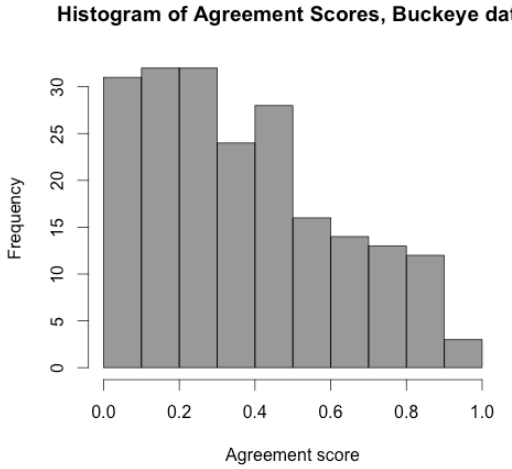

Figure 3: *A histogram of the distribution of agreement scores in the Buckeye sentences.*

It was expected that stress patterning would be more salient in the regular sentences, making them easier for listeners to annotate, and yielding higher levels of inter-transcriber agreement. Contrary to this expectation, we find that in both samples listeners fail to agree on the transcription for a majority of syllables.

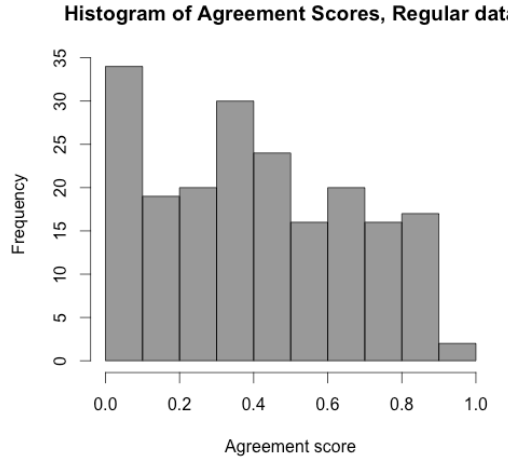

Figure 4: *A histogram of the distribution of agreement scores in the Regular sentences.*

4. Discussion

Given that metrically regular speech is advantaged in production and perception, we predicted that listeners would hear speech as regular and report few clashes. Our results bear this out—participants mark significantly fewer clashes than would be expected if syllables were randomly marked as beats, and also fewer than are expected based on the dictionary stress of the citation form of the words. This finding is consistent with the experimental hypothesis, suggesting the influence of a regularity bias in the perception of sentence-level stress. However, despite this apparent bias participants *do* report stress clashes, even in the regular sentences which were designed to have no clashes. This finding was surprising and points to the difficulty English speakers have in identifying and reporting lexical stress. Overall, these results support a 'soft' regularity bias [15,16], which favors regular alternating stress patterns at the sentence level, but which also allows for deviation from the preferred regular pattern in speech perception (and, we hypothesize, in speech production).

An important finding of this study is that listeners do not achieve a high level of agreement in their responses. Phonological theory which projects stress from accented syllables to the phrase level [1] would lead us to assume that patterns of word stresses are straightforwardly determined by the words of the utterance. However, our results show that individual listeners report different stress placement after being exposed to the same acoustic stimuli.

Some of this variation my be due to listeners defining "beat" differently, or interpreting the task somewhat differently. Then too the current results do not address individual differences in listeners, who varied in age (and no doubt in various cognitive measures). However, we believe that despite these potential sources of noise, the variability in the reported results point to variability in the perception of the acoustic cues to stress at the sentence-level. If listeners were uniform in their stress perception, we would expect high levels of agreement when a group of listeners were presented with the same acoustic stimuli, designed to be metrically regular. Our results are not consistent with this prediction.

The tendency to minimize clash in the perceived pattern of stresses in a sentence is shown to be all the more robust when inter-transcriber agreement in the marking of stress beats is taken into consideration. Though listeners did not agree in their placement of "beats," they nonetheless as a group avoided marking clashes. The low incidence of reported stress clash, together with the low rates of agreement in the marking of stress beats suggests a regularity bias in perception. These results motivate a thorough investigation of acoustic correlates of stress in these stimuli as predictors of listener responses, and individual differences in stress perception. This investigation is currently underway.

5. Conclusions

There are three main conclusions of this study. First, the comparison of observed rates of perceived stress clash with expected rates suggests a regularity bias in the perception of sentence-level stress, in that listeners report perceiving fewer clashes than would be expected. Second, despite the fact that listeners as a group report fewer stress clashes than expected, individual listeners disagree on stress placement for a given sentence, suggesting that the perception of stress may vary from listener to listener. Lastly, this study shows that concatenating citation form stress as marked in a dictionary provides a poor model of listeners' perception of sentence-level stress. Though citation stress without function words is a better model of perceived stress patterns, it still differs significantly from listeners' reported perception.

Further research is called for to determine which measure of stress is most accurate as a representation of stress as produced by a speaker, or as perceived by a listener, and whether the two measures converge on a common stress pattern. Finally, we note that inter-transcriber variability in the perception of stress beats, as reported here, is similar to the variability in pitch accent perception reported in studies of prosodic transcription [e.g., 17]. This parallel is expected if stress beats at the sentence level are equated with prominence-lending pitch accents in prosodic transcription systems.

6. Acknowledgements

Thanks are due to Cody T. Johnson for data collection and analysis, and Evangeline Reynolds for simulation coding. This project is supported by a Cognitive Science and Artificial Intelligence Award to the first author from the Beckman Center for Advanced Science and Technology at the University of Illinois at Urbana-Champaign. The second author's contribution was supported by NSF BCS 12-51343.

7. References

[1] Selkirk, E. 0. *Phonology and syntax: the relation between sound and structure*. Cambridge, Mass.: MIT Press. 1984.

[2] Vogel, I., Bunnell, T, and Hoskins, S. "The Phonology and Phonetics of the Rhythm Rule." In B. Connell and A. Arvaniti [Ed.], *Papers in Laboratory Phonology IV*, Cambridge: University of Cambridge Press. 1995.

[3] Grabe, E. and Warren, P. "Stress Shift: Do Speakers Do It or Do Listeners Hear It?", In B. Connell and A. Arvaniti [Ed.], *Papers in Laboratory Phonology IV*, Cambridge: University of Cambridge Press. 1995.

[4] Zheng, X., and Pierrehumbert, J, "The Effects of Prosodic Prominence and Serial Position on Duration Perception." *Journal of the Acoustical Society of America* 128 (2): 851. 2011 doi:10.1121/1.3455796.

[5] Quené, H., and Port, R.F. "Effects of Timing Regularity and Metrical Expectancy on Spoken-word Perception.", *Phonetica,* 62 (1): 1–13. 2005.

[6] Brown, M., Salverda, A. P., Dilley, L. C., Tanenhaus, M. K "Metrical expectations from preceding prosody influence spoken word recognition." Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX. 2012.

[7] Rothermich, K., Schmidt-Kassow, M.,and Kotz, S. A. "Rhythm's gonna get you: Regular meter facilitates semantic sentence processing." *Neuropsychologia*, *50*(2), 232–244. 2012. doi:10.1016/j.neuropsychologia.2011.10.025

[8] Tilsen, S. "Metrical Regularity Facilitates Speech Planning and Production." *Laboratory Phonology* 2 (1) Jan. 2011. doi:10.1515/labphon.2011.006. http://www.degruyter.com/view/j/labphon.2011.2.issue-1/labphon.2011.006/labphon.2011.006.xml.

[9] Peelle, J. E., and Davis, M.H. "Neural Oscillations Carry Speech Rhythm through to Comprehension." *Frontiers in Psychology* 3. 2012. doi:10.3389/fpsyg.2012.00320.

[10] Dauer, R.M. "Stress-timing and Syllable-timing Reanalyzed." *Journal of Phonetics* 11 (1): 51–62. 1983.

[11] Low, L.E., Grabe,E. and Nolan,F. "Quantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English." *Language and Speech.* Dec. 2000

[12] Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler- Lussier, E. (2007) Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

[13] Qualtrics survey software. www.Qualtrics.com

[14] R Core Team, "R: A Language and Environment for Statistical Computing" Vienna, Austria, 2013. http://www.R-project.org

[15] Beckman, M.E. "Evidence for Speech Rhythms Across Languages." In Y. Tohura, E. Vatikiotis-Bateson, and Y. Sagisaka [Eds.] *Speech Perception, Production and Linguistic Structure*, 457–63. Tokyo: IOS Press. 1992.

[16] Laver, J. *Principles of Phonetics*. Cambridge: Cambridge University Press. 1994.

[17] Pitrelli, J.F., Beckman, M.E., & Hirschberg, J. "Evaluation of prosodic transcription labeling reliability in the ToBI framework." In Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, 123-126, 1994.