INTERSPEECH 2009 BRIGHTON

Perception of English Compound vs. Phrasal Stress: Natural vs. Synthetic Speech

Irene Vogel, Arild Hestvik*, H. Timothy Bunnell*+, Laura Spinu**

*Linguistics and Cognitive Science, University of Delaware, USA
+Speech Research Laboratory, Nemours Biomedical Research, USA
{ivogel,arild,bunnell,lspinu}@udel.edu

Abstract

The ability of listeners to distinguish between compound and phrasal stress in English was examined on the basis of a picture selection task. The responses to naturally and synthetically produced stimuli were compared. While greater overall accuracy was observed with the natural stimuli, the same pattern of greater accuracy with compound stress than with phrasal stress was observed with both types of stimuli.
Index Terms: Phonetics, phonology, prosody, speech synthesis

1. Introduction

The prosodic distinction between compounds and phrases in English has received much attention from various perspectives since the Compound Stress and Nuclear Stress Rules were explicitly formulated in *The Sound Pattern of English* [1]. The issue has been addressed not only from a theoretical perspective, but also from the perspectives of perception, acoustics and language acquisition. In the present paper, we focus on the perception of the distinction between compound and phrasal stress, comparing responses to natural and synthetic speech. We also consider the acoustic properties of the compounds and phrases that are used in the perception study, and briefly touch on the question of the acquisition of the relevant prosodic patterns.

2. Properties of compound and phrasal stress

While it turns out that the well-known pattern of compound vs. phrasal stress in English exemplified by a minimal pair such as ***green*** *house* vs. *green* ***house*** is somewhat variable (see [2]), there is nevertheless a large percentage of compounds that exhibit prominence on the first member of the string, as opposed to phrases that exhibit prominence on the last member. The present research addresses this typical distinction, and makes use of compound stimuli that are uncontroversially stressed on the first word. Descriptively, the compounds and phrases can be distinguished in terms of their syllabic stress pattern. That is, in compounds, the first element is the primary stressed syllable while in phrases both elements bear primary stress (e.g. [2], [3]). Thus, when realized in an utterance, these stress differences may afford differences in accentual patterns. Acoustically, these differences are typically associated with variations in the pitch, duration and amplitude of syllables associated with the compound or phrase elements.[1,2] In perceptual studies of minimal pairs, it has been found that the ability to reliably distinguish compound from phrasal stress patterns develops surprisingly late: both [5] and [6] observed that children only reach adult performance levels at 11 to 12 years of age.

Given that this perceptual distinction is apparently subtle, we sought to compare perception of the compound/phrasal distinction in natural versus synthetic speech tokens. This comparison might (a) lay the groundwork for future studies using only synthetic speech where stimulus properties can be more tightly controlled, (b) provide additional insight into perceptual features for natural speech, and (c) serve to measure the extent to which the contrast is correctly rendered by the specific TTS system being used.

3. Perception of compound and phrasal stress with natural and synthetic speech

3.1. Methodology

The present investigation continues the line of research involving the perception of minimal pairs in which subjects are presented with an auditory stimulus with either compound or phrasal stress. They also see two pictures, one corresponding to the compound meaning and one corresponding to the phrasal meaning. The task is then to indicate which of the two pictures best matches what they have heard.

3.1.1. Stimuli

While the stimuli in [6] were produced in a somewhat child-directed voice (the youngest children were 5-6 years old), the natural stimuli used in the present study were prepared in a colloquial adult-directed speech style.[3] The synthesized stimuli were generated using the ModelTalker TTS system [7], a concatenative synthesizer that also affords control of intonation and timing in its synthesized speech, and thus seemed to offer a good comparison with the natural voice stimuli.[4] In the laboratory version of the ModelTalker system, both intonation and timing are controlled via PSOLA processing. Consequently, F0 and timing effects associated with pitch and phrase accents are synthesized. However, other factors associated with prominence in natural speech, including changes in amplitude and voice quality or spectral tilt would be represented in the synthetic output only to the

[1] Spectral tilt has also been considered recently in the study of stress (e.g. [4]).
[2] See also [2] and references therein for discussion of the acoustic properties of stress.
[3] The first author read the natural stimuli here and in [6], so the only difference is in the style of speech used in the two cases.
[4] The ModelTalker voice used in the present study was based on an extremely sparse corpus comprising less than 150 short sentences; in further research, these findings will be compared with a voice based on the normal 1650 utterance corpus.

extent they were present in the concatenation units the system selected from its speech datatbase.

Twelve ambiguous minimal pairs differing only in the compound vs. phrasal stress patterns were used (e.g. *red head* vs. *red head*). (See appendix for full list.) In addition, twenty-four pairs of control items, which also served as distracters, were included. These were unambiguous pairings of a compound and a phrase where one of the words was kept constant (e.g. *cat food* vs. *fat cat*) so that the difference between these items and the target stimuli was less apparent. The items were all presented in the frame: "Show me where the X is." Following the presentation of the auditory stimulus, the subjects saw a pair of pictures. For the experimental items, each picture corresponded to either the phrasal or the compound meaning. That is, for the pairing above, one picture would show a head colored red, and another picture would show a female head with red hair. For the controls, pictures were presented corresponding to the items named. A practice session administered prior to the actual experimental procedure contained four ambiguous pairs and eight control pairs of items. These were different from the stimuli used in the actual experiment.

3.1.2. Procedure

The experiment was presented individually to each subject in a sound-attenuated booth. E-prime software was used to present a subject with one of two versions of the test, such that for each pair of items, only one appeared in a given test. For each picture pair, half the subjects heard the compound version and half the phrasal version. The items appeared in a different random order for each subject. First the auditory stimulus was presented, and then the subject saw a screen with pictures corresponding to the two meanings of the ambiguous pairs, or pictures corresponding to the two items in the control pairs. The subjects responded by pressing "1" or "0" if the appropriate picture was on the left or right side, respectively. The phrasal and compound pictures were counterbalanced for left-right presentation, crossed with counterbalancing of whether the left or right picture was the correct choice.

3.1.3. Subjects

The subjects were all undergraduates at the University of Delaware who participated in return for extra credit in a linguistics course in which they were enrolled. Since all students were given the opportunity to participate, we were not able to use the responses of a number of the participants. Specifically, we only considered the results of participants who were right-handed monolingual speakers of English, with English-speaking parents. We excluded from consideration the responses of left-handed participants and participants with any history of speech or language problems. Thus, our analysis is based on the responses of 34 participants for the natural speech condition and 28 participants for the synthetic speech condition

3.2. Results

With regard to overall accuracy of picking the picture corresponding to the compound or phrasal stress of the auditory stimuli, we observe a significant difference for speech type. That is, the performance was consistently better with the natural speech stimuli than with the synthetic speech stimuli. The overall accuracy is shown in Figure 1, which also presents the results of the adult participants in [5] and [6], as a basis for comparison.

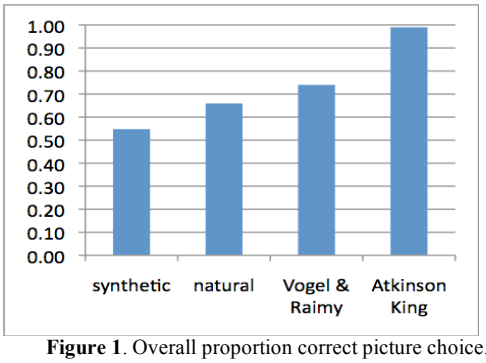

Figure 1. Overall proportion correct picture choice.

Focusing on the results of the present study, we find a further significant effect, that of the stress pattern. That is, when the subjects heard a compound stress pattern, they were more likely to choose the correct picture than when they heard the phrasal stress. This difference was observed in both the natural and synthetic speech conditions, as seen in Figure 2.

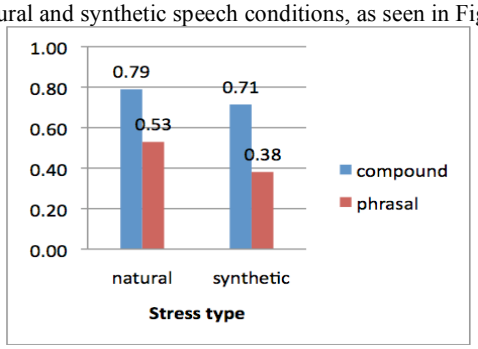

Figure 2. Proportion correct picture choice by speech type and stress type.

Accuracy and reaction time data were analyzed with two-level hierarchical linear modeling [8]. The repeated measures, within-subject variable stress was modeled at the trial level, using the compound stress means as intercept, and PHRASAL as a coefficient for stress at the trial level. The between-subjects variable voice type was modeled at the subject level, with natural voice means as the intercept and SYNTHETIC as coefficient. The mixed model for accuracy is given in (1), where the probability of correct judgment for a single trial in a single subject is expressed as the logarithm of the odds of correct judgment (the logit).

(1) logit = g_{00} + g_{01}(SYNTHETIC) + g_{10}(PHRASAL) + g_{11}*SYNTHETIC*PHRASAL + u_0 + u_1*PHRASAL + r

The coefficient for synthetic speech was not significant (-0.39, t(60) = -1.384, p=.17). The coefficient for phrasal stress was significant (-1.19, t(60) = -6.38, p < .001). The interaction term was not significant (t(60) = -0.81), i.e., the decrease in accuracy for phrasal stress was the same for synthetic and natural speech. Using the same model for reaction time data (using data from only correctly judged trials), with RT on each trial as the predicted outcome, the main effect of voice type was not significant (synthetic speech, 108ms slower, t(60) = 0.96, p = .34); the coefficient for phrasal stress was marginally significant (also 108ms; t(60) = 1.971, p = .053), and the interaction term was not significant (t(60) = -0.95, p = .34). Thus, the RT data match the accuracy data, in that subject were both less accurate and generally slower in the phrasal trials.

4. Discussion

At first glance, it seems surprising that [5] observed close to 100% overall accuracy among the adults, compared to the results in the other studies. Closer examination of the methodologies reveals a possible explanation for this difference. In [5], the subjects were explicitly trained with the items that they were subsequently tested on. By contrast, in the other studies different items were used for the practice sessions. It was felt that the responses would thus more accurately represent the perception of the distinction between compound and phrasal stress in general. If this is, in fact, a robust distinction of English, subjects should recognize it without training on specific stimuli.

Comparing the studies that did not train the subjects on the targets (i.e. [6] and the present study), we nevertheless observe substantial differences. First, we see a substantial difference in overall accuracy between the two studies when we just compare the two sets of natural stimuli (i.e. those in [6] and those in the first condition in the present study). A further significant difference was observed between the natural and synthetic stimuli in the present study. Since [6] made use of child-directed speech—the focus was on the acquisition of the stress distinction—it seemed possible that there were specific properties of the stimuli that may have made the perception task easier or harder. We thus examined certain aspects of the acoustic properties of the different sets of stimuli.

Our acoustic analysis focuses on the durations of the two words (or stressed syllable in bisyllabic items) in the compounds and phrases, and the change in F0 within the rhyme of the two words (or stressed syllable in bisyllabic items). These properties were chosen since these are the ones reported in [6], and could thus be compared in the different studies. In [6] it was reported that the durations of the words in a pair such as *green house* are essentially the same in both the compounds and phrases, with one exception. That is, significant final lengthening was observed in the second word (*house* in this case) when it was at the end of the phrasal stress pattern, but not the compound stress pattern. A similar pattern was observed in the natural stimuli in the present study, although to a slightly lesser extent. Interestingly, the same duration patterns were not found in the synthetic stimuli where, instead, we observe substantial lengthening at the end of the compound, but not the phrasal stress stimuli. The comparison of the duration properties is shown in Figure 3; W1 and W2 are the first and second words of the stimuli.

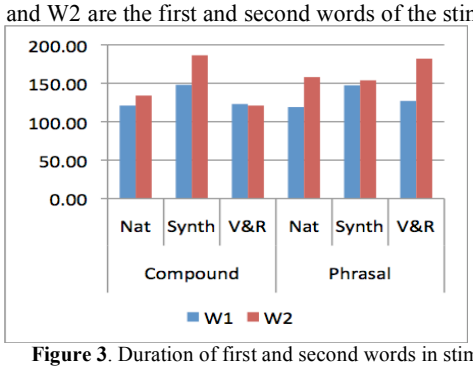

Figure 3. Duration of first and second words in stimuli.

With regard to F0, the change from the beginning to the end of the rhyme portion of the stressed syllables of the stimuli was examined. In [6] it was observed that in the compounds, F0 rose on the first word, and fell on the second; in the phrases, F0 fell somewhat on the first word, and fell more substantially on the second. A similar pattern was observed, but to a much smaller extent in the natural and synthetic stimuli in the present study, as shown in Figure 4.

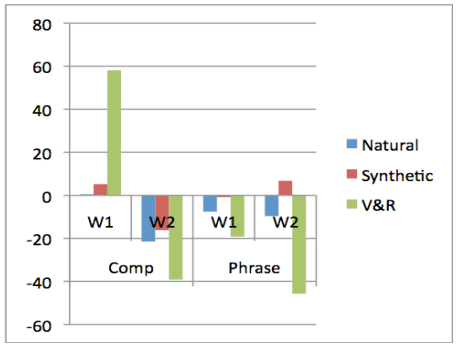

Figure 4. F0 change from beginning to end of syllable rhyme of first and second words in stimuli.

Both the timing and intonation effects are further illustrated in Figure 5, which shows examples of the tokens for *green house* and *green house* for natural speech (lower panels) and synthetic speech (upper panels). These figures show waveforms and F0 contours for both the compound and phrasal tokens. Note in the natural speech compound case (right panel) the relatively sharp drop in F0 following the end of the stressed first syllable, presumably associated with the low phrase accent that follows the nuclear pitch accent on *green*. By contrast, intonation is relatively flat with a slight rise in F0 over the final word *house* of the phrase (left panel). In this case, the high nuclear pitch accent is falling on the last stressed syllable of the Intonation Phrase.

A similar, but less salient pattern can be seen for the corresponding synthetic utterances shown in the upper panels. Note, for example, that while F0 is falling throughout the *house* portion of the compound (right panel), the drop is not nearly as well marked as with the natural speech. This points to an area where the intonation model of the TTS system could be improved.

Thus, while the overall accuracy results with regard to compound and phrasal stress show the same patterns with both the natural and synthetic speech stimuli, we now find an account for the different degrees of success across the studies. The most successful response pattern was observed in [6], which used child-directed speech, a fact that is seen in the more differentiated pitch patterns in particular. Between the two conditions in the present study, we find a more successful response pattern with the natural stimuli, which show quite similar duration properties to those of [6], while the synthetic stimuli show rather different duration properties.

A question that will be addressed in future research is whether an improved version of the synthesized stimuli will significantly improve the accuracy of the subjects' responses, and whether the improvement will continue to be distributed across the compounds and phrases in the same pattern we have observed with the stimuli considered in this study. As noted earlier, the synthetic stimuli in the present experiment were derived from a highly restricted speech corpus. This led to greater than usual spectral discontinuities and may in turn have disrupted the sentence processing, forcing the perception of phonetic and prosodic factors to compete. Since the synthesizer will employ exactly the same intonation and timing regardless of the phonetic units being concatenated, we will be able to directly test this possibility.

The comparison of the compound and phrasal stress patterns also reveals an interesting bias. That is, it appears that the compound interpretation is favored over the phrasal interpretation. This can be seen in the better performance with the compound stress stimuli, a pattern that was also observed in the oldest (12-year-old) children in [6]. That is, while the subjects were frequently able to use the phrasal stress pattern in the selection of the appropriate pictures, they made more

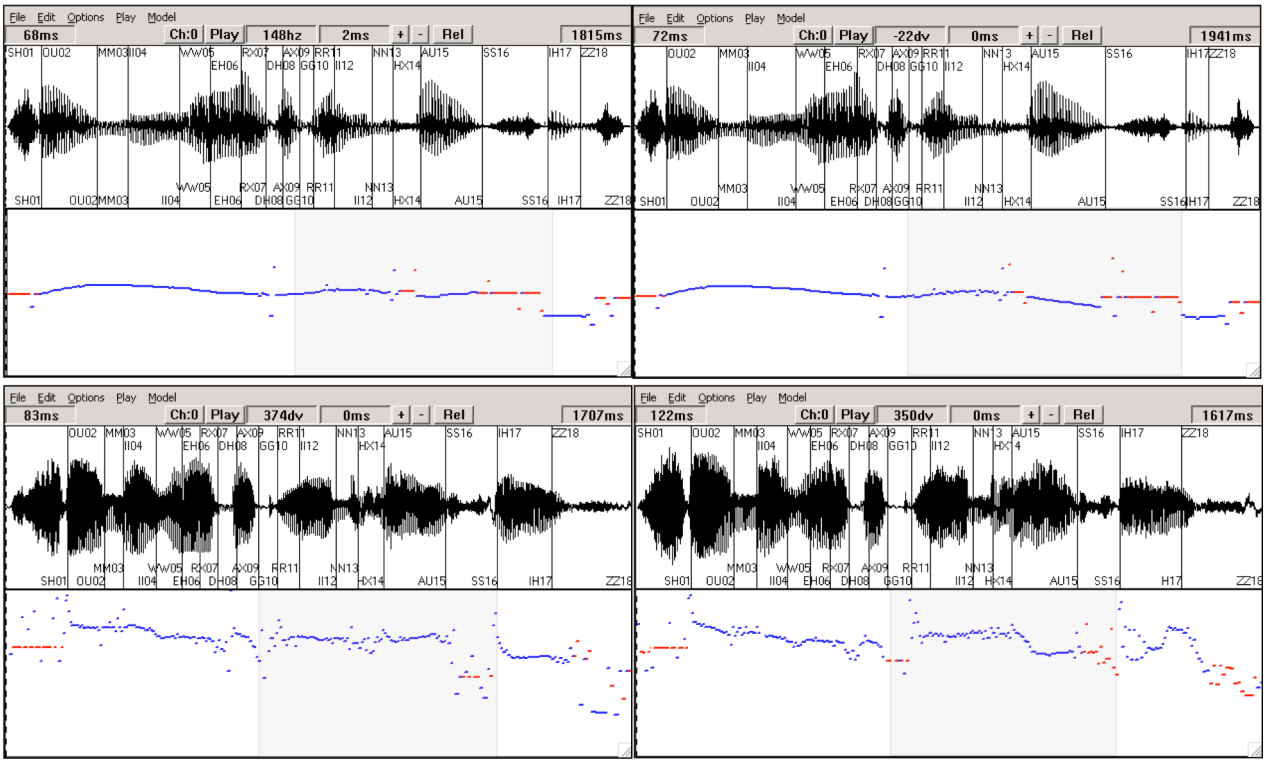

Figure 5. Waveform and F0 traces for synthetic (top panels) and natural (bottom panels) examples of 'green house' as both phrasal (left panels) and compound (right panels) elements. Shaded areas indicate location of 'green house' in each panel.

errors with this prosodic pattern, picking the picture corresponding to the compound interpretation. An even stronger bias in favor of the compound interpretation was seen in the younger children in the acquisition studies. That is, when the younger children heard either *green house* or *green house*, they consistently picked the picture corresponding to the former, regardless of the stress pattern.

5. Conclusions

The ability of subjects to choose the correct picture when presented with either a compound or phrasal production of items such as ***green*** *house* and *green* ***house*** was examined with natural and synthetic voice stimuli. While there was higher overall accuracy with the natural stimuli, there was the same pattern of greater accuracy with the compounds than with the phrases with both types of stimuli. This apparent preference for the compound stress pattern seems to reflect a bias in the acquisition process as well, since younger children very strongly favor the compound interpretation over the phrasal interpretation, a pattern that gradually decreases up to the age of twelve years.

6. References

[1] Chomsky, N. and M. Halle (1968). *The Sound Pattern of English*. New York: Harper and Row.

[2] Plag, I., G. Kunter, S. Lappe and M. Braun (2008). The role of semantics, argument structure, and lexicalization in compound stress assignment in English. *Language*. 84: 760-794.

[3] Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

[4] Sluijter, A., V. van Heuven and J. Pically(1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustic Society of America*. 101:503-513.

[5] Atkinson-King, K. (1973). Children's acquisition of phonological stress contrasts. UCLA: Unpublished Ph.D. Diss.

[6] Vogel, I. and E. Raimy (2002). The acquisition of compound vs. phrasal stress: the role of prosodic constituents. *Journal of Child Language*. 29:225-250.

[7] Bunnell, H.T., Pennington, C., Yarrington, D., and Gray, J. (2005). Automatic personal synthetic voice construction. Proceedings of the Eurospeech 2005, Lisbon, Portugal. September 4-8, 2005.

[8] Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage Publications, Inc.

Appendix: Target stimulus items

a. Experiment

big birds	black belt	black board
blue jay	green house	high chair
hot dog	orange tree	paper boys
red head	soft ball	yellow jacket

b. Practice

black top lady bug toy store white house