Cue Equivalence in Prosodic Entrainment for Focus Detection

Martin Ho Kwan Ip, Anne Cutler

The MARCS Institute and ARC Centre of Excellence for the Dynamics of Language (CoEDL), Western Sydney University, Penrith South, NSW 2751, Australia

m.ip@westernsydney.edu.au, a.cutler@westernsydney.edu.au

Abstract

Using a phoneme detection task, the present series of experiments examines whether listeners can entrain to different combinations of prosodic cues to predict where focus will fall in an utterance. The stimuli were recorded by four female native speakers of Australian English who happened to have used different prosodic cues to produce sentences with prosodic focus: a combination of duration cues, mean and maximum F_0, F_0 range, and longer pre-target interval before the focused word onset, only mean F_0 cues, only pre-target interval, and only duration cues. Results revealed that listeners can entrain in almost every condition except for where duration was the only reliable cue. Our findings suggest that listeners are flexible in the cues they use for focus processing.

Index Terms: prosody, entrainment, focus, speech perception

1. Introduction

Humans use prosody to signal information structure, and possibly universally [1]. Speech perception involves a number of mental challenges where listeners not only need to process the segmental features that make up the words and phrases in the speech stream, but also the prosodic features that determine the wider discourse structure and the speaker's intended message. On this view, attending to prosody may be a useful strategy for finding the most important highlighted part of the utterance, and research has indeed shown that prosodically focused words are more perceptible [2], are recognised more rapidly [3], are processed more deeply in lexical activation [4], and are better retained in memory [5, 6].

However, it remains unclear whether some prosodic cues (e.g., F_0 versus duration) may prove more informative to listeners' processing of information structure. In earlier experiments [7, 8], Cutler and colleagues discovered that listeners could anticipate an upcoming accented word by entraining to various features in the utterance prosodic contour. Using a phoneme detection task, Cutler and colleagues asked participants to listen to a series of sentences and respond as fast as they could to words that began with a specified phoneme stop target (e.g., /d/ in "duck"). Listeners responded faster to the target in sentences where the preceding intonation contour predicted high stress on the target-bearing word, compared to sentences where the intonation predicted low stress. Importantly, response times were still faster for sentences with predicted high stress contexts, even when the original target words in both contexts were replaced by an acoustically identical neutral version of the same words. Since the only difference was in the preceding intonation, it was concluded that listeners can already entrain with the cues in the preceding prosody to anticipate an upcoming focus before they receive the acoustic signals of the focused word.

Subsequent experiments [9] using the same phoneme detection paradigm revealed that listeners can still forecast an upcoming focused word even when the F_0 information in the preceding prosody is rendered uninformative (by being monotonised). Similarly, listeners can still process upcoming focus when the duration of the closure before the burst of the target stop phoneme is controlled. Building on these findings, the present paper seeks to further examine the role of different prosodic information by using natural speech from sentences recorded by different speakers who happened to have used different prosodic cues in producing the same set of stimuli.

2. Experiment 1

2.1. Method

2.1.1. Participants

The sample consisted of 22 native speakers of Australian English (M_{age} = 24.23 years, SD = 8.76 years; 15 females). All of the participants reported that they were born and raised in Australia.

2.1.2. Materials

Twenty-four unrelated experimental sentences were recorded in three versions by a female native speaker (see Figure 1). In the first version, the target-bearing word received emphatic stress. In the second version, emphatic stress was instead placed on a word that occurred later in the sentence than the target-bearing word, which, as a result, received very reduced stress. In the third version, the target-bearing word and the sentence as a whole were produced in a neutral manner. In all of the experimental sentences, the phoneme target was a voiceless aspirated bilabial stop [p^h] occurring at the start of the target word's first syllable (e.g., [p^hi:nʌts] "peanut").

Using Praat [10], the target-bearing words were excised from all three versions of each experimental sentence. The high- and low-stressed target-bearing words from the first and second versions were replaced by an acoustically identical token of the same target word from the neutral version. Thereby, two experimental conditions were constructed, each containing one version of each of the 24 spliced experimental sentences, plus an additional set of 24 filler sentences. The experimental sentences with predicted high versus predicted low stress were counterbalanced across the two conditions. To avoid interference between the sentences, sentence beginnings were varied and semantic content that could be associated with another sentence in the set was avoided. In addition, apart from the target-bearing word, none of the sentences had any additional occurrence of the target phoneme or any other stop phonemes similar to the target phoneme (e.g., [b]). All of the sentences were produced at a natural fast-normal rate.

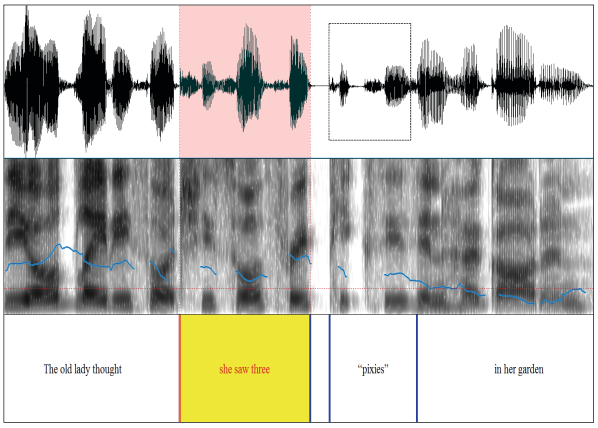
Target: [p^h]

(a) The old lady thought she saw three **[PIXIES]** in her garden.

(b) The old lady thought she saw three pixies in her **[GARDEN]**.

(c) The old lady thought she saw three pixies in her garden.

(a)



(b)

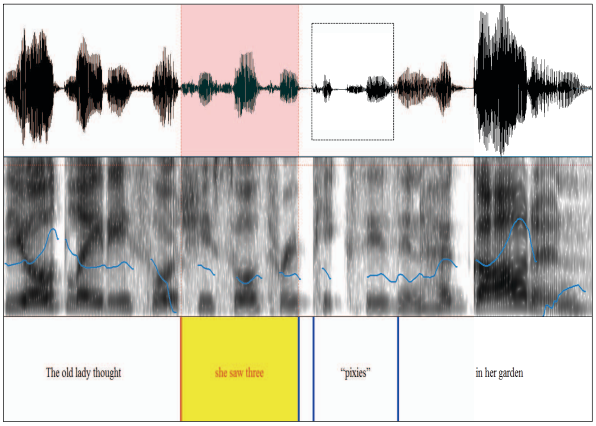


Figure 1: *Waveforms and pitch contours of an example experimental sentence in predicted high (a) and low (b) contexts; text (c) gives the neutral context. The shaded portion – three syllables preceding the target-bearing word – was analysed acoustically (section 2.1.2).*

We conducted acoustic analyses of the stimuli experimental sentences based on simultaneous inspection of the waveform and the spectrogram in Praat. Segments consisting of around three to four or five syllables before the onset of the target-bearing word were annotated and duration, mean F_0, maximum F_0, and F_0 range were measured. We also measured temporal signals such as the pre-target interval, the part of the utterance between the onset of the target-bearing word and the offset of the word before it (usually around 60 to 100 milliseconds). Results show that, for all measurements, the preceding intonation contours of sentences with predicted high stress contexts were significantly higher than the sentences with predicted low stress.

2.1.3. Procedures

Participants were tested in a sound-attenuated booth at the MARCS Institute, Western Sydney University. The phoneme-detection task was administered using E-Prime software on a laptop computer, with attached to it a set of headphones and a Chronos USB-based device for button pressing. Participants were told that the experiment aimed to examine listeners' memory and language comprehension. All participants were told that they would listen to a series of sentences and had two tasks: first, pay careful attention to the meaning of each sentence, and second, press the button as soon as they heard a word that began with the target phoneme. Participants received two practice trials and feedback before starting the actual experiment. At the end, all participants completed a follow-up recognition test in which they were asked to judge whether or not each of the 20 sentences in the list was from the experiment. We only included data from participants who scored 65 percent or above in the test.

2.2. Results and Discussion

Response times (RT) longer than 2500 milliseconds were excluded from final analyses, because such a delayed response may indicate a reprocessing of the sentence [7]. A two-tailed within-subjects t-test with an alpha threshold of .05 was conducted to assess the difference in RT between the predicted high versus low stress sentences. RTs were significantly faster in predicted high stress sentences (M = 414.92, SD = 71.68) compared to sentences with predicted low stress (M = 447.09, SD = 59.81), $t(21) = 2.83$, $p = .010$ (see Figure 2).

With respect to detection accuracy, we performed a two-tailed binomial sign test to determine whether participants were more likely to miss a button press to the phoneme target in sentences with predicted low stress than in predicted high stress. In total, there were one miss in predicted high stress contexts and five misses in low stress contexts, which was not statistically different from chance, $p = .219$ (see Table 1).

Consistent with previous studies, the results revealed that Australian English speakers can entrain with the preceding contour to forecast an upcoming focused word. However, because the acoustic analyses of the stimuli revealed significant differences for all measurements, it remains unclear as to whether some types of cues are more informative than others. Therefore, we conducted a second experiment using the same sentences produced by a different speaker.

3. Experiment 2

3.1. Method

3.1.1. Participants

We recruited a new sample of 23 native speakers of Australian English (M_{age} = 22.16 years, SD = 5.37 years; 17 females).

3.1.2. Materials and Procedures

The procedures and sentences were identical to those in the previous experiment, only this time, the sentences were recorded by another female native speaker. Acoustic analyses of the experimental sentences only revealed significantly higher mean F_0 in the predicted high stress sentences. It is important to note that no explicit instructions were given for the speaker to produce the sentences in any particular way (e.g., produce the preceding prosody with higher pitch).

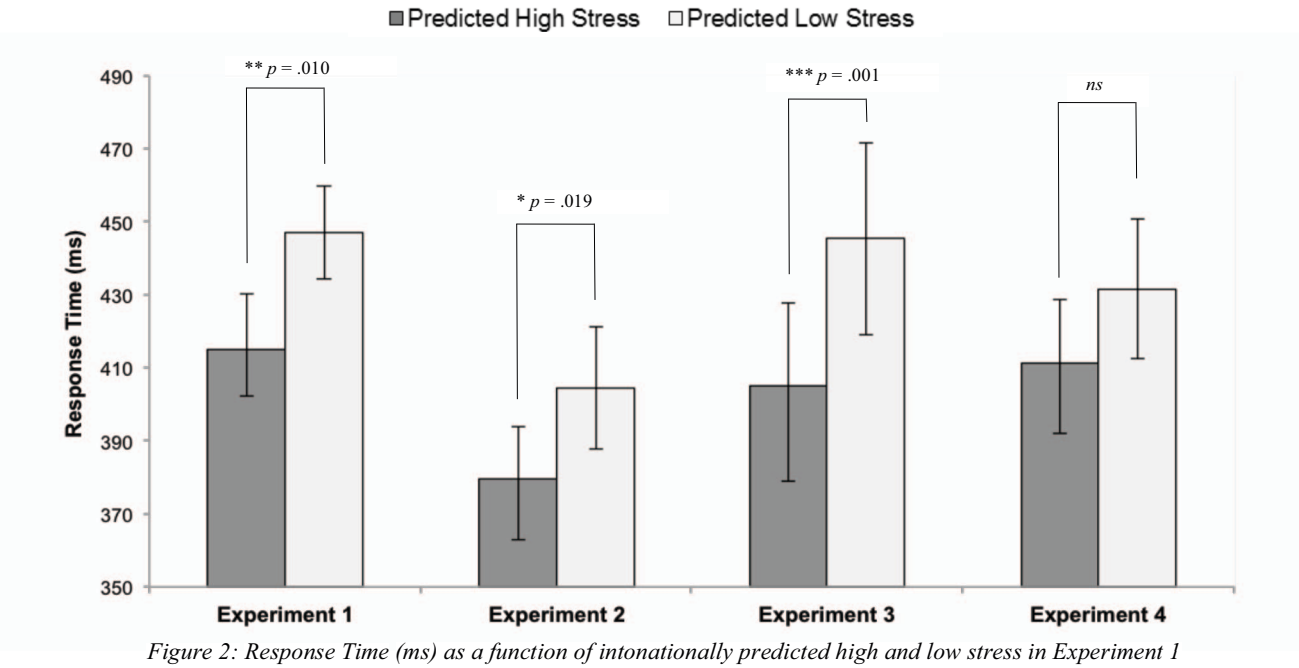


Figure 2: Response Time (ms) as a function of intonationally predicted high and low stress in Experiment 1 (with significant acoustic differences in mean F_0, maximum F_0, F_0 range, pre-target interval, and overall duration), Experiment 2 (significant acoustic difference only in mean F_0), Experiment 3 (significant difference only in the pre-target intervals), and in Experiment 4 (with only significant difference in overall duration). Error bars indicate standard error of the mean.

3.2. Results and Discussion

Consistent with the results from Experiment 1, participants' RT in Experiment 2 was faster for predicted high stress sentences (M = 379.65, SD = 68.12) compared to low stress sentences (M = 404.52, SD = 80.44), $t(22) = 2.54$, $p = .019$. In terms of accuracy, there was one miss and one false alarm (i.e. responding before the target phoneme occurred) for the predicted high stress sentences and only one false alarm for the low stress contexts.

The results indicate that listeners are as likely to use the cues from the preceding intonation regardless of whether there is a combination of many different cues (as in Experiment 1) or whether there is only one type of cue (Experiment 2). However, it is still an open question whether listeners of Australian English can still entrain if the most informative cue in the preceding prosody is not F_0-based. In the following experiments, we used the same set of sentences recorded by speakers who happened to have signalled upcoming focus using mostly duration-based cues.

4. Experiment 3

4.1. Method

4.1.1. Participants

There were 23 native speakers of Australian English (M_{age} = 22.04 years, SD = 6.80 years; 19 females).

4.1.2. Materials and Procedures

All sentences and procedures were identical to the previous experiments. Acoustic analyses of the experimental sentences recorded by the third female native speaker only revealed significantly longer pre-target intervals before the target words in the predicted high stress sentences.

4.2. Results and Discussion

Consistent with the results from the previous experiments, RT was faster for predicted high stress sentences (M = 405.19, SD = 108.50) compared to low stress sentences (M = 445.37, SD = 126.41), $t(22) = 3.96$, $p = .001$. In terms of accuracy, there were only two misses and one false alarm for the predicted low stress sentences.

5. Experiment 4

5.1. Method

5.1.1. Participants

These were 22 college-aged native speakers of Australian English (16 females).

5.1.2. Materials and Procedures

We used the same procedures and sentences from the previous experiments using stimuli produced by a fourth female speaker. Acoustic analyses of the experimental sentences from this speaker only revealed significant differences in duration, such the preceding part of predicted high stress sentences three to five syllables before the onset of the target-bearing word were longer (i.e. produced slower) than the preceding parts of the low stress sentences. There were no significant differences in the pre-target intervals or in any of the F_0 measures.

5.2. Results and Discussion

In striking contrast to the previous experiments, there was no significant RT difference between the predicted high versus low stress sentences, $t(21) = 0.96$, $p = .346$, although in the same direction. In terms of accuracy, both the predicted high and low stress sentences had an equal number of misses and false alarms (i.e. three misses and one false alarm).

Table 1. *Number of misses as a function of predicted high versus low stress contexts in Experiments 1 to 4.*

	Predicted High Stress	Predicted Low Stress
Experiment 1	1	5
Experiment 2	1	0
Experiment 3	0	2
Experiment 4	3	3

6. General Discussion

The present series of experiments provides a useful insight into how listeners use different prosodic information to detect an upcoming focused word. Consistent with previous findings, we demonstrate that listeners of Australian English can entrain with a variety of prosodic cues to forecast the location of an upcoming focused word in the utterance intonation contour. Results from Experiments 1 and 2 show that sentences that were recorded by the speaker who only consistently produced one type of cue (e.g., mean F_0) to distinguish predicted low and high stress contexts were just as likely to facilitate prosodic entrainment as the sentences produced by the speaker who produced a variety of cues. Further, in Experiment 3, having only pre-target interval as a significant temporal cue further supports the view that F_0 is not a necessary component of the preceding prosody for focus detection. However, Experiment 4 revealed that preceding prosody with longer duration (i.e. slower speech) before the predicted focus can be insufficient to support listeners' prosodic entrainment.

Overall, our findings indicate that although speakers can differ in their prosodic production, listeners are generally flexible in their use of the various prosodic information. Prosodic entrainment to locate focus may be justified by its value as listening strategy for everyday communication and semantic processing [11]. Irrespective of language or culture, holding a conversation presents a number of mental challenges. For one thing, conversational utterances tend to be fragmentary and elliptical [12]. At the same time, there is much uncertainty with respect to how a dialogue will unfold, and listeners often need to constantly organise and update their current discourse model. Given that accented words are generally the semantically most central part of the sentence, entraining to intonation contours to detect focus may therefore provide a headstart for listeners in navigating the utterance information structure early on, making it a strategy useful for all listeners for maintaining a socially effective conversation. On this view, prosodic entrainment could be understood as a comprehension process where listeners could attend to whatever cue they encounter in the speech stream to process the semantically highlighted part of speaker's message.

Of particular note are the results of Experiments 3 and 4, where listeners could successfully forecast an upcoming focused word when the length of the pre-target interval was informative, but not when there was a difference in overall duration of the preceding syllables. We speculate that one of the reasons for the lack of entrainment in Experiment 4 could be because the duration cues were in conflict with other prosodic information (e.g., preceding prosody having longer duration but low F_0) [13]. The pre-target intervals in Experiment 3 may be informative temporal cues because they represent an intake of breath or pausing before the focused word, which is in line with previous research showing that speakers tend to pause to single out new information [e.g., 14].

Future research can also assess whether listeners' flexibility in prosodic entrainment could also partly be based on a statistical learning mechanism. For example, one way in which listeners can use the different cues is by extracting the statistical information about the types of prosodic cues that are characteristic of a particular speaker.

7. Conclusion

Our findings provide evidence that (1) individual speakers within a given language (i.e., Australian English) can differ in the prosodic cues they display, (2) despite these differences, listeners can entrain with almost any cue or combination of cues in the speech signal to efficiently anticipate an upcoming focused word, and (3) it is unlikely that there is a hierarchy of cues in terms of how well they facilitate prosodic entrainment.

8. Acknowledgements

We acknowledge financial support from the ARC Centre of Excellence in the Dynamics of Language (CE140100041**).** We thank Mark Antoniou and Chris Carignan for technical advice. We also thank Matthew Stansfield for his support when we set up a student club to recruit research participants.

9. References

[1] Bolinger, D. L., "Intonation across languages", in J. Greenberg [Ed], Universals of Human Language II: Phonology, 471-524, Stanford University Press, 1978.

[2] Lieberman, P., "Some effects of semantic and grammatical context on the production and perception of speech", Lang. Speech., 6(3): 172-187, 1963.

[3] Cutler, A. and Foss, D. J., "On the role of sentence stress in sentence processing", Lang. Speech., 20(1): 1-10, 1977.

[4] Norris, D., Cutler, A., McQueen, J. M. and Butterfield, S., "Phonological and conceptual activation in speech comprehension," Cognit. Psych., 53(2):146-193. 2006.

[5] Fraundorf, S., Watson, D. G. and Benjamin, A. S., "Recognition memory reveals just how CONTRASTIVE contrastive accenting really is", J. Mem. Lang., 63(3): 367-386, 2010.

[6] Kember, H., Choi, J. Y. and Cutler, A., "Processing Advantages for Focused Words in Korean", Speech Prosody Proc., 702-705, 2016.

[7] Cutler, A., "Phoneme-monitoring as a function of preceding intonation contour", Percep. Psychophys., 20(1): 55-60, 1976.

[8] Akker, E. and Cutler, A. "Prosodic cues to semantic structure in native and nonnative listening", Biling: Lang. Cogn., 6(2): 81-96, 2003.

[9] Cutler, A. and Darwin, C. J., "Phoneme-monitoring and preceding prosody: Effects of stop closure duration and of fundamental frequency", Percept. Psychophys., 29(3): 217-224, 1981.

[10] Boersma, P., "Praat, a system for doing phonetics by computer", Glot Intl., 5(9/10): 1381-3439, 2002.

[11] Cutler, A. and Fodor, J., "Semantic focus and sentence comprehension", Cogn., 7(1): 49-59, 1979.

[12] Garrod S. and Pickering, M. J., "Why is conversation so easy?", Trends Cognit. Sci., 8(1): 8-11, 2004.

[13] Cutler, A. "Components of prosodic effects in speech recognition," Proc. 11th Intnl. Cong. Phon. Sci., 84-87, 1987.

[14] Gee, J. P. and Grosjean, J. "Empirical evidence for narrative structure", Cognit. Sc., 8(1): 59-85, 1984.