The interpretation and phonetic implementation of !H* in American English

Jill C. Thorson1 & Rachel Steindel Burdin1

1University of New Hampshire
rachel.burdin@unh.edu, jill.thorson@unh.edu

Abstract

Downstep in American English has been understudied relative to other types of pitch accents. Our aim is to investigate both the interpretation and the phonetic implementation of !H*. The first experiment investigated participant’s preference for H* vs. !H* pitch accents in new vs. accessible contexts. Results from this experiment show that participants showed a preference for H* in both new and accessible contexts, but that they choose !H* relatively more in the accessible ones. Additionally, participants were more likely to select H* when it had a smaller fall onto the stressed syllable. The second experiment explored whether participants could distinguish between !H* pitch accents with larger and smaller falls. The results showed that participants were more accurate discriminating between stimuli that were further apart in pitch; however, this effect was mediated by the stimuli’s f0 range, with lower stimuli being easier to discriminate than higher ones. Together, these experiments reveal the complexities of downstep in American English in both where it occurs pragmatically and how it is phonetically produced and perceived.

Index Terms: intonational phonology, pragmatics, auditory perception, acoustic phonetics

1. Introduction

In Mainstream American English (MAE), pitch accents have been claimed to mark information structure. While the interpretation and phonetic implementation of high pitch accents (H* and L+H*) has been well-studied ([1], [2]), downstepped pitch accents (!H*) have been less so. In MAE, the downstepped pitch accent has been described as being used to relay information that is inferable [3]. Beyond this definition, there has been little exploration of the pragmatic interpretation(s) of the downstepped high pitch accent. In the related West Germanic language of German, !H* has been suggested to express accessibility versus that of newness or givenness [4]. Corpus-based research in MAE has proposed that like most intonational contours, there can be multiple functions for downstep contours [5]. In particular, the !H* has been hypothesized to have at least two functions related to discourse structure. Supporting [3], it was shown that the downstepped pitch accents relayed information that was inferable from the shared beliefs of the speaker and hearer (accessible). In contrast, H* marks items that are new in the discourse, that is, items to be added to the hearer’s belief space. Critically, [5] concluded that downstep may be predicted by a number of factors including the discourse structure, information status, the speakers themselves, and other features such as the length of the noun phrase. Thus, the meaning of the downstep contour has remained elusive with additional research needed.

In terms of phonetic implementation, the H* pitch accent has a (rise to a) peak aligned with the stressed syllable of a word [6]–[8]. Meanwhile, the !H* is realized with a lower peak than the preceding H*, but not so low to be considered at the bottom of the speaker’s range [9]. The !H* cannot occur as the first pitch accent in a phrase but must occur after the presence of a previous H*, which triggers the downstep. Phonetically, the stereotypical pitch range between the first H* and the !H* has been reported to be about three semitones [10]. Interestingly, past work has debated the legitimacy of the downstepped pitch accent as its own prosodic category [11], while later work using acoustic analysis and machine learning maintains a categorical distinction between H* and !H* [12].

Our goal is to explore the phonetic implementation of downstep and how it is interpreted relative to a given context. The primary research questions of this study are:

1) What is the meaning of the !H* pitch accent? Specifically, in what contexts do listeners prefer the !H* compared to the H*?

2) To what extent do listeners treat downstep falls (that is, the size of the step down from H* to the following !H*) as members of same category? Are listeners able to differentiate different degrees of downstep (i.e., more or less of a fall)?

To address these questions, we conducted two experiments. In experiment 1, we explore the interpretation of !H* in comparison to H*. After reading a particular context, listeners were asked which type of contour (one with and one without downstep) best followed. Based on previous research, we hypothesized that one of the meanings of downstep is to relay information that is accessible given the context and that the accessible contexts would see a bias in !H* responses [3], [4].

The examination of the acoustic realizations of the !H* stimuli in experiment 1 led to a second experiment. In experiment 2, we delve deeper into the phonetic realization of !H* and ask what the phonetic, and potential categorical boundaries are, of downstep. Here, we hypothesized that listeners would be able to perceptually differentiate the different degrees of downstep (i.e., a larger step versus a smaller step), but that there may be response variability due to the size of the step, the order presentation of the stimuli, and where in the range the step occurred. This set of predictions for experiment 2 are based on work by [13], which examined a similar effect in the L*-H category. Together, experiments 1 and 2 address the complexities in determining the interpretation and perceptual boundaries of downstep in American English.

2. Experiment 1: Meaning of !H*

2.1. Participants

A total of 61 (33 female; 27 male; 1 agender) participants were recruited via the online platform Prolific. An online platform was preferred due to COVID-19 restrictions in collecting in person data. Participants were all native speakers of American English with an average age of 34.7 years; the majority (50)
were white. They were from a range of locations across the United States.

2.2. Procedure

Twelve scenarios were constructed, with two versions for each scenario. In the first, the target utterance (1) was intended to be new (scenario 1a), and in the second, the target utterance was intended to be accessible (scenario 1b).

(1) Target utterance: He ran a marathon in town.
(1a-New): You and your friend are discussing what your other friend, Paul, did over the weekend.
(1b-Accessible): Your friend Paul is very active, and nearly every weekend, he goes and runs a race. You and another friend are discussing what Paul did over the weekend.

Two versions of each target utterance were recorded: a H* H* L-L% (high star) guise and a H* !H* L-L% (downstep) guise. Both had two pitch accents (indicated by bolding above); we included unaccented material after the nuclear pitch accent to allow for time to produce the L-L% fall. The utterances were recorded by 4 trained speakers (2 male, 2 female (the authors); all white). The experiment was blocked, so that participants saw either the new or accessible context for each scenario. Participants were then presented with both audio guises and asked which one sounded better in the context that they had read. Thirty-eight filler items were also included. Each participant saw a total of 40 contexts (12 targets and 38 fillers).

2.3. Results & Discussion

Overall, the high star guises were preferred (selected 64% of the time) regardless of the context, but less so in the accessible contexts (57%) than in the new ones (71%), as predicted. Initial examination of the data suggested a large range of response variability depending on both the scenario and the speaker. We decided to further explore both the extent to which these two factors affected the selection of either the high star or the downstep guise, as well as acoustic measurements for both guises. For the H* H* guise, we examined (1) the peak of the second H*, in f0; for the H* !H* guise, we examined (2) the f0 of H*, (3) the f0 of the high point before the !H*, and (4) the span of the fall onto the downstep (i.e., the difference between (2) and (3); for both guises, we examined (5) the duration of the stressed vowel, and (6) the overall pitch range of the utterances.

For (1)- (4) in addition to the raw f0 of these points, we also looked at a relative measure of f0. Following PoLaR guidelines, the overall pitch range of the utterance was measured, and divided into quintiles; the f0 of the point was then assigned the number of the quintile it fell into, with 1 being the lowest and 5 being the highest [14].

Due to the high correlation between these variables, we created a random forest model to explore the importance of the variables for predicting the selection of either the high star or downstepped guise, following the procedure in [15]. Figure 1 shows the relative importance of the predictors to the model (overall accuracy, C=0.77). A higher number indicates a greater degree of contribution to model accuracy; numbers near or at zero were not important to the model. Scenario was the most important, followed by Context (new vs. accessible). After these two variables, there were a range of measurements which were related to the production of the downstep guise that influenced responses: the initial f0 of the !H* in both relative (DownstepEnd_Levels) and absolute (Downstep_End_Hz) terms, and the duration of the stressed vowel in both the H* and !H* guises (HStar_Duration and Downstep_Duration). A conditional inference tree grown on the data (Fig. 2) helps visualize the interaction between these factors. Scenario, Context, and Speaker all occupy the highest branches. Below that, downstep guises were generally dispreferred if the fall was too large, in both absolute (falling to the 2nd quintile or lower; compare nodes 9 and 10 in figure 2) and relative terms (fall span of greater than 42 Hz; compare nodes 4 and 5). We theorize this may be due to these downsteps not sounding enough like an expected downstep, and rather sounding too much like a fall to a L* or L-L%. Following this finding, we created experiment 2 to further investigate the auditory perception of different downstepped contours that vary by the amount of fall.

![Figure 1: Conditional permutation variable importance](image1.png)

![Figure 2: Conditional inference tree](image2.png)

3. Experiment 2: 2IAX

3.1. Participants

A total of 41 participants (28 female, 12 male, 1 gender fluid) were recruited from an introductory linguistics class at a mid-sized university in New England, and from the wider community. Students received extra credit for participation; community members received a $10 gift card to either Target or Amazon. The majority (39) identified as white and were between the ages of 18 and 22 (M = 22.1). All reported spending the majority of their childhood in the United States, with most (31) participants being from New England. Data from one participant were lost due to a technological malfunction.
3.2. Procedure

To explore the acoustic perception of downstep, we ran a 2IAX perceptual experiment. Participants completed a 2IAX experiment: that is, they heard pairs of stimuli, one after another, and asked if they sounded the same or different. Three sentences were recorded by two speakers, one female (the 1st author) and one male. There were two versions of each sentence recorded: one H* H* L-L%, and one H* !H* L-L%; as in experiment 1. The H* !H* L-L% stimuli were then annotated using PoLaR [14]. The H* !H* L-L% stimuli all had a drop from a high point to the downstep, followed by a slight fall from the !H* syllable to the end of the phrase, as can be seen in figure 3. Six different versions of these drops were created: one from the top of the speaker’s pitch range in the utterance to the bottom (5-0), one from the top to the top of the 1st quintile (5-1), one from the top to the top of the 2nd quintile (5-2), one from the top to the top of the 3rd quintile (5-3), the top of the 4th (5-4), and finally, a flat version with no fall until after the stressed syllable (5-5) (see Figs. 4-6 for examples). The size of the steps was thus dependent on the original pitch range of the utterance and ranged from 4 Hz to 12 Hz. Note in the plots below, the H* H* guise is represented as 5-6 (Figs. 7-10).

![Figure 3: Original utterance](image)

![Figure 4: 5-0 guise](image)

![Figure 5: 5-3 guise](image)

![Figure 6: 5-5 guise](image)

![Figure 7: Percent correct (“same” for identical guises; “different” for different guises) by step size between guises](image)

Participants heard all possible combinations of stimuli pairs for each utterance, including the H* H* L-L% version, cross-balanced for order of presentation, so participants would hear both a 5-1 followed by a 5-2 and a 5-2 followed by a 5-1. The former, where a larger drop (e.g., 5-1) followed by a smaller drop (e.g., 5-2) will be called AB order; the latter, a smaller drop (e.g., 5-2) followed by a larger drop (e.g., 5-1) will be called BA order.

3.3. Results & Discussion

Figure 7 presents the overall correct rates by the step size between the stimuli, from 1 step to 6 steps, with 6 steps representing the difference between the 5-0 fall and the H* guise. We can see a few overall trends. First, participants were very accurate at identifying both identical utterances, and distinguishing the H* H* from the H* !H* guises; part of this might be due to the fact that the H* H* guises were different original recordings, and were thus generally quite distinct from the H* !H*. Second, there is a very clear linear trend, where the larger the size between the stimuli, the more accurate the participants were, ranging from around 27% for stimuli separated only by 1 (e.g., a 5-1 and a 5-2) step to 81% percent for stimuli separated by 5 (a 5-0 and a 5-5) steps. Participants were also overall more accurate when a smaller drop was followed by a larger drop (BA order). This mirrors the finding in [13], which found greater accuracy in distinguishing between two peaks when the higher peak was presented second: the opposite order (a high peak, than a low) may have resulted in greater perceptual similarity due to general declination effects. Here, a smaller fall is heard as more equivalent to a larger one when the larger one is presented first; again, declination may be the explanation.
Figure 8 shows the stimuli separated by 1 step: 5-0 and 5-1, 5-1 and 5-2, etc. Two patterns can be seen here. First, the overall accuracy was quite low and decreased the higher in frequency the stimuli were: a 5-4/5-5 pair was more difficult to distinguish than a 5-0/5-1 pair. Similar patterns can be seen in figures 9 (stimuli separated by two steps) and 10 (stimuli separated by three steps). This may be as a result of the non-linear perception of f0, with the, e.g., 7 Hz difference separating the two steps being perceptually smaller in the 5-4/5-5 pair compared to the 5-1/5-2 pair. The second is that the general pattern of BA pairs being easier to distinguish than AB pairs decreases the higher in the range one gets, and even switches in some cases, particularly in pairs involving the 5-5 stimuli. One potential explanation for this is overall information about the size of the pitch range which may affect declination patterns may be harder to extract from the 5-5 contours, as they are high and flat for a longer period of time.

Figure 8: Percent correct (“same” for identical guises; “different” for different guises) for guises with 1 step difference

![Figure 8: Percent correct (“same” for identical guises; “different” for different guises) for guises with 1 step difference](image)

Figure 9: Percent correct (“same” for identical guises; “different” for different guises) for guises with 3 step difference

![Figure 9: Percent correct (“same” for identical guises; “different” for different guises) for guises with 3 step difference](image)

4. Discussion & Conclusion

Our study demonstrates the variability of how \( \text{!H}^* \) is produced (experiment 2) and how it is perceived (experiment 1). Our first experiment showed that listeners exhibit different rates of preference for \( \text{H}^* \) versus \( \text{!H}^* \) contours depending on whether the item was accessible or new to the context. This parallels findings from German [4]. This preference appeared to be mediated by the phonetic implementation of the downstep. Downsteps with smaller falls were preferred over ones with larger ones, suggesting that prototypical downstep may be one with a smaller fall onto the downstep pitch accent. This led to experiment 2 where we explored the extent to which listeners could distinguish between different degrees of downstep. In general, we found that listeners showed improved accuracy when the phonetic difference between the falls was greater. When the steps were lower in the pitch range accuracy was improved. A pattern thus emerges in figures 8, 9, and 10, where pairs involving mid-range falls (5-2 and 5-3) were the most confusable, lending support to the findings from experiment 1, that larger falls (e.g., a 5-1 or 5-0) may not be perceived as prototypical downstep. There was also a strong ordering preference discovered, showing that listeners were more accurate when a smaller drop was followed by a larger drop. These results are similar to other work on intra-category of pitch accents [13].

Future work plans to include eye-tracking experiments to investigate online interpretations between \( \text{H}^* \) and \( \text{!H}^* \). Additionally, we plan to examine the ability of listeners to match context to contour in the opposite manner of experiment 1 (given a given contour, which context is preferred; following the experimental design of [13]).

5. Acknowledgements

Thank you to Gabrielle Gagnon, Lydia Sack, Sabrina Salmela, and all of the members of the CAT Lab at UNH for their help in data collection. Thank you to Stefanie Shattuck-Hufnagel for her helpful discussion and feedback, and special thanks to all the participants in experiments 1 and 2.

6. References

