Published in final edited form as:
Lang Cogn Process. 2010 January 1; 25(7-9): 1178–1200. doi:10.1080/01690961003783699.

Accent detection is a slippery slope: Direction and rate of F0 change drives listeners' comprehension

Angela M. Isaacs and
Department of Psychology, University of Illinois at Urbana-Champaign

Duane G. Watson
Department of Psychology, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign

Abstract

The present study tests whether listeners use F0, duration, or some combination of the two to identify the presence of an accented word in a short discourse. Participants' eye movements to previously mentioned and new objects were monitored as participants listened to instructions to move objects in a display. The name of the target object on critical trials was resynthesized from naturally-produced utterances so that it had either high or low F0 and either long or short duration. Fixations to the new object were highest when there was a steep rise in F0. Fixations to the previously mentioned object were highest when there was a steep drop in F0. These results suggest that listeners use F0 slope to make decisions about the presence of an accent, and that F0 and duration by themselves do not solely determine accent interpretation.

How you say something can be just as important as what you say. A pitch accent (hereafter simply 'accent') is an acoustic prominence associated with a word. It has been argued that prominence is signaled by a number of acoustic changes, which include a rise in fundamental frequency (F0), an increase in duration, and an increase in intensity (Ladd, 1996). Accents can affect the meaning of the utterance by signaling whether information is new, important, in focus, or non-predictable (Bolinger, 1972; Brown & Yule, 1983; Büring, 2001; Halliday, 1967; Levy & Jaeger, 2007; Schwarzschild, 1999; Selkirk, 1995). It is clear that accents play an important role in signaling discourse structure. However, it is less clear which acoustic feature(s) listeners use to determine whether a word is accented. The present study explores listeners' use of F0 and duration in accent detection.

Many theorists have argued that F0 underlies the perception of accenting, though they differ in how they characterize its role. For example, in Pierrehumbert's influential theory (Pierrehumbert, 1980; Pierrehumbert & Hirschberg, 1990), which has been instantiated in the Tone and Break Index labeling system (ToBI; Silverman et al., 1992), accents are composed of tonal targets that are either high (**H**) or low (**L**) and these are defined in relation to the speaker's F0 range. Different targets and different sequences of targets convey different meanings. F0 also plays a critical role in other theories of prominence (Bolinger, 1986; Cruttenden, 1997; Halliday, 1967). For example, Cruttenden (1997) proposed that pragmatic meaning is conveyed by individual F0 contours such as rises, falls, plateaus, and combinations of these features. Two words both containing a rise in F0 may be treated differently depending on the extent of the rise and the path it takes to reach the

Corresponding Author: Angela M. Isaacs, aisaacs2@uiuc.edu, 603 East Daniel St., Champaign, IL 61820, Phone: (217) 721-2631, Fax: (217) 244-5876.

maximum. Critically, although proponents of these different approaches would argue that duration and intensity correlate with accenting, F0 plays a central role within these theories.

Work in the computational literature also suggests that F0 is used in accent perception. Grabe, Kochanski, and Coleman (2007) tested whether computer-calculated polynomial equations could be used to detect differences between the F0 contours of seven different types of accents in a corpus of hand-labeled speech. The accents included (in their modified ToBI notation) fall (**H*L**,%), fall-rise (**H*L**,**H**%), high rise (**H***,**H**%), rise-plateau (**L*H**, %), rise (**L*H**,**H**%), late-rise (**L***,**H**%), and rise-plateau-fall (**L*H**,**L**%). These equations were first fit with one set of utterances and then tested using a different set of utterances from the same labeled corpora. Comparisons of the coefficients of these seven equations revealed numerous differences between accent types, suggesting that different accent types can be distinguished based on F0 contour. However, in a follow-up experiment, Grabe et al. trained several computer algorithms to classify accents as belonging to one of the seven accent types previously identified by the experimenters. Each classifier was trained to distinguish between two of the seven accent types using the F0 contours of the words. The classifier then categorized novel words as being one of the two accent types on which the classifier had been trained. The classifiers were successful in distinguishing between some, but not all, accent types, suggesting that F0 alone may not be sufficient in distinguishing one accent type from another.

An earlier experiment by Kochanski, Grabe, Coleman, and Rosner (2005) provides some evidence that intensity and duration may fill this gap. Kochanski et al. developed and tested classifiers designed to identify words in spoken speech that had been labeled as accented by trained labelers. Each classifier was based on a different acoustic measure including F0, intensity, and duration, among others. In contrast to Grabe et al. (2007), the classifier based on F0 contour was one of the worst performers. The most successful classifiers were those based on intensity and duration. Because Kochanski et al. only compared accented to unaccented words (regardless of accent type), this experiment should not be taken as evidence against Grabe et al.'s results showing the importance of F0 contour. However, it does suggest that duration and intensity may be important for listeners to distinguish accented from unaccented words.

Bard and Aylett (1999) found additional evidence for the importance of duration in prominence by examining speech in a referential communication task. Words that were repeated across the discourse were coded using a variant of ToBI and were measured for intelligibility and duration. Bard & Aylett found that, compared to its first realization, the second realization of a word was only rarely coded as de-accented but was frequently produced with reduced intelligibility and duration. This is contrary to the well-documented findings that 1) given words are de-accented in speech and 2) reduced intelligibility and duration of given words is due solely to de-accentuation (Bolinger, 1972; Brown & Yule, 1983; Halliday, 1967). This illustrates that for production at least, duration may have an important role to play in conveying givenness.

As a whole, these studies suggest that duration and F0 play an important role in listeners' ability to identify accented words, but it is unclear what role each plays. Research in speech perception and prosody perception provide some possible ways in which F0 and duration might interact. In phoneme perception multiple, redundant phonetic cues exist and sometimes the presence of any one cue is sufficient to indicate the correct phoneme. Repp (1985) termed this type of relationship "cue trading". Beach (1991) found a cue trading relationship between F0 and duration in the detection of intonational boundaries while resolving locally ambiguous sentences. Listeners may also use a similar relationship between F0 and duration to make inferences about accenting.

Bartels and Kingston (1994) found evidence for a different relationship between F0 and duration. Bartels and Kingston were interested in exploring differences between presentational (**H***) and contrastive (**L+H***) accents types. They synthesized multiple versions of the phrase “Amanda had a banana”. For the word “banana”, they systematically varied the F0 height on the peak, the F0 height on the dip prior to the peak, the alignment of the F0 peak, and the onset of the F0 peak (i.e. how late the F0 began to rise). Listeners rated whether each utterance had occurred in a discourse context consistent with a contrastive or presentational accent. They found that the cue which best distinguished presentational from contrastive utterances was the F0 slope, which is the direction and rate of F0 change. This suggests that listeners are capable of using F0 slope to distinguish between accent types and raises the possibility that listeners also use F0 slope to distinguish between accented and unaccented words.

The research cited above suggests that listeners may use both duration and F0 to determine when a word is accented. However, these studies relied upon intuitions and data from tasks that are unlike natural comprehension. Although this work has served as an important first step in understanding the acoustics of accents, directly testing the effect of acoustic factors on accent perception in real time language processing may provide deeper insights into the role F0 and duration play in accent detection.

To investigate this question, we employed a visual world paradigm (Allopenna, Magnuson, & Tanenhaus, 1998; Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In this paradigm, the location of participants’ gaze within a visual display is recorded over time as participants engage in normal language comprehension or production. Changes in looking preferences over time provide information about the time course of processing. Many studies have used the visual world paradigm to test questions relating to prosody (e.g. Arnold, 2008; Chen, Os, & De Ruiter, 2007; Dahan, Tanenhaus, & Chambers, 2002; Ito & Speer, 2008; Weber, Braun, & Crocker, 2006; Watson, Tanenhaus, & Gunlogson, 2008; for an overview of the visual world paradigm and prosody see Watson, Gunlogson, & Tanenhaus, 2006).

Dahan, Tanenhaus, & Chambers (2002) conducted an important, early study on prominence using the visual world paradigm that serves as a launching point for the work presented here. Dahan et al. found that participants’ gaze location was reliably related to the prosody and given/new status of the critical word. Participants listened to short discourses (see [1] below, capital letters indicate accenting) consisting of two instructions to move objects within a visual display. On critical trials, two of the possible targets were cohort competitors, words which share initial word onsets and are therefore temporarily lexically ambiguous (Allopenna, Magnuson, & Tanenhaus, 1998; Marslen-Wilson, 1987). In the first instruction, participants were told to move one of these cohort objects. In the second instruction participants were told to move either the given cohort object (anaphoric condition) or the new cohort object (non-anaphoric condition).

[1] Example discourse from Dahan et al. (2002)

a. Instruction 1 Move the *candy/camel* below the square.

b. Instruction 2 NOW move the *CANDY/candy* above the star.

Dahan et al. (2002) measured participants’ fixations to given and new cohort objects during the ambiguous region of the critical word in the second instruction. Consistent with previous research showing that listeners associate accenting with new information (e.g. Birch & Clifton, 1995; Terken & Nooteboom, 1987), they found that when the critical word was produced with an accent, listeners were biased to look at the new cohort object and when the critical word was produced without an accent, they were biased to look at the given cohort object. This suggests that 1) listeners are sensitive to prosodic information, 2) listeners use this information to make rapid predictions about referents in a discourse context, and 3) there is a link between listeners' processing of accents and looking preferences.

The present study will use a modified version of the design used by Dahan et al. (2002). The critical word in the second instruction (i.e. "candy" in [1]) was resynthesized to have either high or low F0 and either long or short duration. Resynthesis was used in this study because of the inherent drawbacks associated with synthesized speech and natural speech. Synthesized speech offers a high degree of control but the resulting audio file does not sound natural. On the other hand, human produced speech sounds very natural but is difficult to control. Resynthesized speech combines the strengths of both approaches. Resynthesis is the process of artificially altering the acoustic properties (like F0, duration, and intensity) of naturally-produced audio to produce audio that is both relatively natural-sounding and highly controlled. This has made resynthesis a popular tool for exploring questions in prosody and phonology (e.g. Dilley & Brown, 2007; Ladd & Morton, 1997; Ladd & Schepman, 2003; Welby, 2006).

Participants' interpretations of the resynthesized word will be reflected in their looking preferences to the given and new cohort objects. Dahan et al.'s (2002) results suggest that if listeners detect an accent on a word referring to a cohort object, then they should be biased to look at the new cohort object. In Dahan et al.'s study, the speaker produced high F0 peaked accents (**H*** and **L+H***). These accents are associated with a high F0 and a relatively long duration. If listeners detect de-accenting, then they should be biased to look at the given cohort object. Unaccented words are characterized by the absence of a rise in F0 and a short duration. This pattern was confirmed by an analysis of the recordings used for re-synthesis, which we present below: accented words are long and are produced with a rise in F0 while unaccented words are short and are produced with a fall in F0.

Several outcomes are possible. If F0 excursion alone is critical for signaling accenting, High F0 should result in more looks to the new cohort object and fewer looks to the given cohort object, independent of duration.

It is also possible that listeners will ignore all F0 information and, instead, use duration alone as a cue to accenting. This would result in a main effect of duration such that long duration, which has been claimed to correlate with accenting (e.g. Kochanski et al., 2005), would result in more looks to the new cohort object than the given cohort object.

Alternately, both F0 and duration may be important. Research suggests two possible interactions: a cue trading relationship or the use of F0 slope. Beach (1991) found a cue trading relationship between F0 and duration: either was sufficient for indicating the presence of a boundary tone. Therefore, if a cue trading relationship exists between F0 and duration in accent perception, then either high F0 or long duration, which are both associated with high F0 peaked accents, will result in more looks to the new cohort object and fewer looks to the given cohort object. The condition in which neither of these cues is present will result in a bias to the given cohort object.

A second possibility is that F0 slope is critical for accent detection. Bartels and Kingston (1994) found that listeners used F0 slope to detect the difference between accent types. In the present study, F0 slope is determined by a combination of both the F0 and duration manipulations. In an initial analysis of stimuli for this experiment that were produced in accented and unaccented contexts, we found that accented words had a positive slope and unaccented words had a negative slope, the latter most likely due to declination that occurs naturally over a sentence. Positive sloped conditions (high F0) are a better match for accented words and should result in more looks to the new cohort object. On the other hand, negative sloped conditions (low F0) are a less good match and should result in more looks away from the new cohort object, and possibly more looks to the given cohort object. In addition, because the F0 contour for the two F0 conditions were set to the same height, conditions with short duration had a steeper slope than conditions with long durations although the absolute F0 excursion is constant within each F0 condition. Bolinger (1986) argues that more abrupt changes in F0 are more salient. Therefore, if slope is important for perceiving the presence of an accent, then there should be more looks to the new cohort object for steep, positive slopes (high F0, short duration conditions).

Methods

Participants

Forty-eight native speakers of American English were recruited from the University of Illinois at Urbana-Champaign. Participants received either class credit or $7 compensation for their participation.

Materials

Trials—Participants completed 20 critical trials, 30 filler trials, and 4 practice trials for a total of 54 trials. Each trial consisted of two separate instructions as in [2] below to move one of the four possible objects. The information status (anaphoric/non-anaphoric) of each trial was determined by the object moved in the first instruction. In the anaphoric condition the same (given) object was moved in both instructions. In the non-anaphoric condition a different (new) object was moved in the second instruction. The F0 and duration manipulations were realized on the critical word in the second instruction (i.e. ‘camel’). Four different versions of the critical word were resynthesized (described below) resulting in a complete cross of F0 (high/low) and duration (long/short). In both the anaphoric and non-anaphoric conditions the object that served as the destination marker (i.e. “circle” and “triangle” in [2]) was different in the two instructions so that accenting the destination would be felicitous. Because English sentences typically have at least one pitch accent, accenting the destination marker was necessary so that sentences in which the target word might be interpreted as unaccented would still be felicitous.

[2] Instruction 1:

Anaphoric: Move the camel below the circle.

Non-anaphoric: Move the candle below the circle.

Instruction 2:

Now move the *camel* above the triangle.

Information status (anaphoric/non-anaphoric) was a between subjects factor while the acoustic variables, F0 (high/low) and duration (long/short), were within subjects factors. Two cohort items and two distractor items were included on each critical trial for a total of four pictured items (Appendix). All cohort pairs were picturable nouns that shared initial phonemes, had equal numbers of syllables, and were matched for lexical frequency (paired t = 1.373, df = 19, p > .05) as reported in the CELEX English database (1993). One member of each pair was randomly chosen to be the critical item in the second instruction. In addition, four stationary geometric figures were present on all trials (Figure 1). Distractor items were not semantically related to the cohort pair and did not share word onsets with the cohort pair.

Filler trials included anaphoric trials in which the same object was moved in both instructions and non-anaphoric trials in which different objects were moved in each instruction. Filler trials included 10 trials in which no cohort pair was present (5 anaphoric and 5 non-anaphoric), 8 trials in which a cohort pair was present but not mentioned (5 anaphoric, 3 non-anaphoric), and 2 trials in which a cohort pair was present but only one of the cohort objects was mentioned in only one instruction (both non-anaphoric). In addition, there were 10 trials in which a cohort pair was present and was used in both instructions. To help balance the overall information status of all trials, these 10 trials were anaphoric for subjects whose critical trials were non-anaphoric and these 10 trials were non-anaphoric for subjects whose critical trials were anaphoric. One of the filler trials, including the cohort pair 'cat'-'cap', was dropped for some participants because earlier participants consistently mistook the target word ('cap') for the cohort competitor ('cat').

Visual Stimuli—The 216 pictures used in the visual presentation were selected from the original Snodgrass and Vanderwart (1980) normed picture set, the colorized version of this picture set by Rossion and Pourtois (2001), and from commercial picture databases. All pictures were colored line drawings or were black and white line drawings which were colorized using computer software.

Resynthesis of audio files—The critical word for each experimental trial was modified using the Pitch Synchronous Overlap Add (PSOLA) algorithm (Moulines & Verhelst, 1995). This process allowed us to alter the F0 and duration of the critical words in the audio files while leaving all other acoustic features of the soundwave unchanged. Resynthesis took place in three steps, which are discussed below in detail: 1) recording naturally-produced audio stimuli (source audio), 2) measurement of median F0 and duration of the critical word across items, and 3) altering F0 and duration values in the critical word of each item, using the measures calculated in step 2 as targets.

Source recordings—It is difficult to know *a priori* what the F0 and duration targets should be for naturally-produced accented and unaccented words. Therefore, the source audio had a dual purpose: it served both as the audio which would be resynthesized and the audio from which target acoustic values were calculated. By taking the aggregate duration and F0 of accented and unaccented words in their appropriate discourse contexts, we could estimate appropriate target values for the accented and unaccented words. To this end, we recorded naturally-produced accented and unaccented words in non-anaphoric and anaphoric contexts, respectively. Median values of F0 and duration were calculated and these values were used in resynthesis as described below.

For each critical trial, a set of four instructions (see [3] below) was recorded by a male native speaker of American English. Two version of the second sentence were produced, one in which the critical word was accented and one in which the critical word was unaccented. The information status (anaphoric/non-anaphoric) of the trial was determined by the cohort object moved in the first instruction. Two versions of the first instruction were recorded: in the anaphoric condition the target cohort object (e.g. ‘camel’) was moved and in the non-anaphoric condition the competitor cohort (e.g. ‘candle’) was moved.

[3] Instruction 1:

Anaphoric: Move the camel below the circle.

Non-anaphoric: Move the candle below the circle.

Instruction 2:

Unaccented: Now move the camel above the triangle.

Accented: Now move the CAMEL above the triangle.

The acoustic features of the critical words in the second instructions were analyzed to determine the acoustic values to be used in resynthesis. Acoustic comparisons were made between the unaccented items, which tend to have short duration and low F0, and accented items, which tend to have long duration and high F0.

The accented versions (mean = 598ms, SD=90.94) were significantly longer than the unaccented versions (mean = 497ms, SD = 79.75), paired t(19) = 7.91, $p < .001$, one-tailed. A ratio of the duration was measured (unaccented duration to accented duration) for each item. The mean ratio for all 20 items was .84.

For each word, F0 was measured by dividing the word into 20 segments of equal duration and measuring the average F0 over each segment. The absolute F0 at any point of a given word is a function of both the accenting on that word and the relative height of the entire utterance in the speaker's range, the F0 baseline. The current experiment used a baselining procedure to isolate the F0 change due to accenting. A constant F0 baseline was measured at the onset of the word "move" for each item and this F0 baseline was subtracted from the F0 means calculated for each segment of the corresponding critical word. This process extracted the F0 contour from the word and corrected for differences in the speaker's range across productions. The algorithms used to extract F0 can occasionally provide spuriously high or low F0 values. Rather than hand-correcting the errors (which may introduce additional error), median values were used to reduce the impact of these outliers on the F0 contour.

The F0 contour for the high F0 condition was generated by calculating the median F0 across all words for each segment. These points are plotted in Figure 2 as grey dots connected by lines. A 6-term polynomial equation (black lines in Figure 2) was generated to fit these 20 F0 points using a line-of-best-fit algorithm in Microsoft Excel. The F0 contour for the low F0 condition was modeled using the same process but with the unaccented items.[1]

Resynthesis—Experimental audio files for each trial were created in Praat (Boersma & Weenink, 2008) using the PSOLA algorithm. The F0 and duration of each critical word was resynthesized to create four distinct versions, all originating from the same source utterance. For this audio set, we found that lengthening a word degraded the word's naturalness. Therefore, resynthesizing from an unaccented sound file, which requires lengthening to produce the sound file for the long duration conditions, resulted in consistently less-natural audio quality for long conditions than short conditions. For this reason, all critical utterances were resynthesized from accented source audio.

The critical word from each accented item was spliced out of the original utterance into a separate audio file. The F0 of each critical word was altered first. The word was divided into 20 segments. For each segment, a new F0 value was created that was equal to the sum of the baseline value for that carrier sentence (calculated as described above) and the target F0 value generated from the polynomials discussed above.

From each raw version of the critical word, two separate audio files were created: one with high F0 and one with low F0. Because these were created from accented sound files (which have naturally long duration), both versions already had long duration. The two audio files for the two short duration conditions were created by shortening the long versions, resulting in four versions of the critical word. The calculated ratio of 0.84 duration difference between

[1] F0 values were measured across the entire critical word including both mono- and disyllabic words. Visual comparisons of the F0 contours collapsed across mono- and disyllabic words separately reveal no difference in contour shape. Separate analyses of the fixation data for mono- and disyllabic words show the same pattern of results, so they are collapsed in the results section.

unaccented and accented audio files was deemed by the authors to be too small a difference to produce a sufficient contrast. Other work done in our lab has shown that there are large individual differences in lengthening for accented words (Isaacs & Watson, 2009) but that differences as large as 60% are still naturally produced (Isaacs & Watson, 2007). Therefore, the short duration audio file was shortened to a value equal to 70% of the original (long) duration.

Figure 3 contains plots of the functions used to generate the F0 contours resynthesized for critical words. To illustrate the difference between short and long duration, each equation has been plotted twice: once normally and once with x-axis units that are 70% of the original length.

The four resulting audio files, each containing one version of the critical word (low F0-short duration, low F0-high duration, high F0-short duration, and high F0-long duration), were then spliced into a carrier sentence. The carrier sentence consisted of a preamble (e.g. “Move the”) and a destination (e.g. “above the star”). The carrier sentence was taken from the version of each item which had originally held the unaccented version of the critical item. After splicing, the preamble and destination of the carrier sentence were stylized. Stylization identifies key points in the F0 contour (local maximums, minimums, and locations of large changes in F0 trajectory) and then interpolates the intermediate F0 values using the PSOLA algorithm. The process maintains the general shape of the F0 contour but introduces a small amount of noise into the recording similar to the type of noise that is present in resynthesized portions of the audio file. The stylization is generally not noticeable compared to unaltered natural recordings, but the process reduced the contrast between the resynthesized and unresynthesized portions of the utterance, improving the overall coherence of the critical utterances.

Procedure

On each trial, participants listened to two separate instructions as in [2] above. Audio files were played over computer speakers. After each instruction, pictured objects were moved by clicking and dragging the critical object to the specified location.

At the onset of the second instruction, the participants' eye movements were monitored using an Eyelink II eyetracking system. Fixations were compared relative to the onset of the critical word. Because it takes roughly 200ms to program an eye movement (Allopenna et al., 1998; Henderson & Ferreira, 2004), we expected prosody driven fixations to appear 200ms after the onset of the critical word.

Participants were not able to move objects while audio was being played. If a participant made an error on the first instruction, either by selecting the wrong object or moving it to the wrong location, all objects returned to their original location and the instruction was repeated. Errors made in the second (critical) instruction were not corrected. All errors were noted for later exclusion in data analysis.

Results

After completing the experiment, participants took a survey asking them 1) if they noticed anything unusual and 2) what they thought the experiment was testing. Of the 41 subjects who completed the survey, 8 stated that the audio sounded strange or mentioned the audio manipulation in their description of the experiment’s purpose.

A total of 6 trials (0.625% of all trials) were excluded because the participant made an error in the first instruction by selecting the wrong object (2 trials) or moving the object to the wrong location (4 trials). On an additional 15 trials (1.56% of all trials), participants selected the wrong object in the second instruction. In 14 of these cases, the participants chose the competitor object suggesting that the preferences induced by the manipulation persisted through the end of the trial. Analyses were performed both with and without these 14 trials, both analyses yielded the same result, therefore these trials were retained in the present analyses because it was felt that these errors were related to the manipulation.

Results are reported as Target Advantage Scores (TAS), which are the proportion of fixations to the target minus the proportion of fixations to the competitor over a specified time region (e.g. Arnold, Fagnano, & Tanenhaus, 2003). The resulting value is the difference in the proportion of fixations between the target and competitor. A positive TAS indicates that the participant was looking at the target more than the competitor and a negative TAS indicates more looks to the competitor. Thus, in the anaphoric condition positive values indicate more looks to the given cohort object and negative values indicate more looks to the new cohort object. In the nonanaphoric condition positive values indicate more looks to the new cohort object and negative values indicate more looks to the given cohort object. Figure 4 shows TAS values plotted over time starting at the onset of the critical word (0 ms).

There was a higher overall TAS for trials in non-anaphoric conditions (Figure 4). In the critical time regions from 200 to 600ms after critical word onset, non-anaphoric conditions had significantly higher TAS than anaphoric conditions, t1(46) = −4.40, p < .001, t2(38) = −3.97, p < .001. This bias was even significant in the 400ms preceding the critical region, t1(46) = −5.84, p <.001, t2(38) = −4.85, p <.001, which is before participants' eye-movements should be affected by the manipulation. This bias replicates the bias found by Dahan et al. (2002).[2]

In the critical time region from 200 to 600ms after word onset, a reliable Information status (Anaphoric/Non-anaphoric) × F0 (High/Low) × Duration (Long/Short) interaction was found, F1(1,46) = 8.02, p < .01, F2(1,38) =11.72, p < .005. Figure 5 illustrates this interaction in the mean TAS values for the critical time region. When duration was long, there was no difference between high and low F0 conditions as shown by a nonsignificant Information Status (Anaphoric/Nonanaphoric) × F0 (High/Low) interaction for long duration conditions, F1(1,46) = 1.13, p>.05, F2(1,38) = 1.39, p>.05. This was true in both the anaphoric condition, t1(23) = 0.52, p > .05, t2(19) = 0.63, p>.05, and the non-anaphoric condition, t1(23) = −0.99, p > .05, t2(19) = −1.01, p>.05. In contrast, the same analysis performed on short duration conditions shows that when the duration was short, there was a difference between the high and low F0 conditions, F2(1,46) = 9.51, p<.005, F2(1,38) = 13.18, p<.005. Again, this was true for both the anaphoric condition, t1(23) = −2.16, p < .05, t2(19) = −2.08, p < .10, and the non-anaphoric condition, t1(23) = 2.23, p < .05, t2(19) = 3.16, p < .01. In anaphoric conditions, TAS scores were higher if the word was produced with low F0 indicating a looking preference to the given cohort object when the critical word was produced with low F0 as compared to high F0. In contrast, TAS scores were higher in the non-anaphoric condition when the critical word was produced with high F0 indicating a looking preference to the new cohort object when the critical word was produced with high F0 as compared to low F0.

Furthermore, when long and short duration conditions are compared, different patterns are found for the anaphoric (Figure 5A) and non-anaphoric (Figure 5B) conditions. In the

[2] Although we find a new bias in this data, replicating Dahan et al. (2002), Arnold (2008) did not find this bias using a slightly different task. This suggests that this bias may be due to task demands rather than a general listener preference to fixate the new item within a contrast set.

anaphoric conditions, when F0 was low there was a difference between the long and short duration, t1(23) = −2.42, $p < .05$, t2(19) = −2.75, $p < .05$, but when F0 was high there was no difference between the duration conditions, t's<1. In the non-anaphoric conditions, by comparison, when F0 was low there was no difference between the long and short duration conditions, t's<1, but there was a significant difference in the high F0 condition, t1(23) = −2.79, $p < .05$, t2(19) = −3.78, $p < .005$. Thus, both conditions show higher TAS values in the short duration condition: there was a preference for the given cohort object when F0 was low and a preference for the new cohort object when the F0 was high.

Discussion

Dahan et al. (2002) found that when the critical word was accented, participants looked more at the new cohort object and when the critical word was unaccented, participants looked more at the given cohort object. In this study, we find that accenting is signaled by an interaction between duration and F0 information. The interaction described above rules out the possibility that listeners rely primarily on either F0 or duration alone to make decisions about accenting. As we discuss below, the data here are more consistent with an F0 slope account of accenting than a cue trading account.

The F0 slope hypothesis predicts that listeners use slope to detect accents. F0 slope is here defined as the direction and rate of F0 change. Accented words in this experiment have high F0 peaks (Figure 3). This positive slope makes listeners more likely to look at the new cohort object. Unaccented words lack this definite peak and, instead, show a general F0 declination over the word. This negative slope makes listeners more likely to look at the given cohort object. Duration affects the rate of this change. Since the absolute amount of F0 change is the same in all conditions within the same F0 condition, a short duration results in steeper slopes than a long duration. Thus, when duration is short, the high F0 condition will have steep positive slope and the low F0 condition will have steep negative slope.

Recall that TAS scores are composed of the difference between the proportion of looks to the target and competitor with positive values indicating more looks to the target and negative values indicating more looks to the competitor. Thus, in the anaphoric condition, positive values indicate more looks to the given cohort object and negative values indicate more looks to the new cohort object. In the nonanaphoric condition, positive values indicate more looks to the new cohort object and negative values indicate more looks to the given cohort object.

In the short duration conditions, therefore, the F0 slope hypothesis correctly predicts that there should be higher TAS scores for the low F0 condition when the target is anaphoric and higher TAS scores for the high F0 condition when the target is non-anaphoric.

When duration is long, slopes will be less steep both in the high and low F0 condition which should reduce or eliminate looking preferences both in the anaphoric and non-anaphoric conditions. The F0 slope hypothesis correctly predicts that when duration is long there should be less of a difference in TAS in both the anaphoric and non-anaphoric conditions. Thus, the F0 slope hypothesis correctly predicts the outcome of comparisons between high and low F0.

Since a positive slope leads listeners to look at the new cohort object and negative slopes lead listeners to look more at the given cohort object, when the slope is in the wrong direction for the target condition, listeners should be biased away from the correct target regardless of duration. Thus, the F0 slope hypothesis correctly predicts that there should be no difference in the TAS values for long and short durations in the anaphoric condition when the F0 is high or in the non-anaphoric condition when the F0 is low. However, when the slope is in the correct direction for the information status of the condition, then the looking preference should be largest in the condition in which the slope is steepest, the short duration condition. Therefore, there should be higher TAS values for the short duration condition when the F0 is low in the anaphoric condition and when the F0 is high in the non-anaphoric condition which is the pattern observed in Figure 5. The F0 slope hypothesis correctly predicts the results of all 8 comparisons between conditions.

In comparison, the cue trading hypothesis predicts that either of the two cues that indicate accenting, long duration or high F0, should be sufficient to bias listeners toward an accented interpretation. Thus, any condition in which either duration is long or F0 is high should result in a higher proportion of looks to the new cohort object which would mean higher TAS values for non-anaphoric conditions, and lower TAS values for anaphoric conditions.

Since cue trading predicts that either long duration or high F0 will result in a bias to look at the new cohort object, then any comparison between two items with long duration or high F0 should result in a non-significant difference. As predicted, there was no significant difference between F0 conditions with long duration in either the anaphoric condition or the non-anaphoric condition and no significant difference between duration conditions with high F0 in the anaphoric condition. However, contrary to the prediction of cue trading, in non-anaphoric conditions there was a significant difference between duration conditions when F0 was high. Not only is there a significant difference when cue trading predicts there should be none, but there is actually a preference for the new cohort object when the duration is short. Unlike the F0 slope hypothesis, the cue trading hypothesis cannot account for this pattern of results.

Cue trading further predicts that any comparison between high and low F0 should result in a larger bias to look at the new cohort object for the high F0 condition. Therefore, in anaphoric conditions, the low F0 condition should have larger TAS values than the high F0 condition and in non-anaphoric conditions the high F0 condition should have larger TAS values than the low F0 condition. In addition, any comparison between long and short duration should result in a larger bias to look at the new cohort object for the long duration condition. Thus, in anaphoric conditions the short duration condition should have larger TAS values than the long duration condition and in non-anaphoric conditions the long duration condition should have larger TAS values than the short duration condition.

As predicted, when duration was short, TAS values were higher in the high F0 condition when the target was non-anaphoric and higher in the low F0 condition when the target was anaphoric. However, cue trading predicts that when F0 is low, TAS values should be higher in the long duration condition than in the short duration condition when the target is non-anaphoric, but no significant difference was found. Furthermore there were higher TAS values in the high F0 condition in the non-anaphoric condition when the duration was short than when it was long.

Thus, while cue trading can largely explain the pattern of effects observed in Figure 5, it cannot explain the significant difference between the high F0 conditions in the non-anaphoric condition or the direction of the effect: that there was a greater looking preference to the new cohort object in the high F0 conditions when the duration was short.

Arguably, the version of cue-trading that we have outlined above is an extreme version in which any cue to prominence (i.e. high F0 or long duration) is independently sufficient for signaling prominence. In a weaker cue based theory, cues to prominence might be interpreted additively such that having two cues that signal prominence leads to stronger expectations of prominence than having only a single cue, and the presence of a single cue leads to a stronger perception of prominence than having no cue. Under such a theory, one would predict a difference between conditions in which one cue indicates accenting and conditions in which both cues indicate accenting. However the difference is in the wrong direction. Even a weaker version cannot explain why looks to the target were greater in the new cohort object condition when F0 was high and duration was short compared to the condition in which F0 was high and duration was long. Under the strong version of the cue trading hypothesis, these two conditions should be equal because both contain at least one cue that indicates accenting. Under the weaker, additive version of the cue theory, the long duration condition should have more looks because it has two cues that indicate accenting (high F0 and long duration) while the short duration condition only has one (high F0). This suggests that F0 slope and not cue trading, best explains the pattern of looking preferences observed in the present study.

Bartels and Kingston (1994) found that F0 slope was important for listeners to distinguish presentational from contrastive accents and the data presented here suggest that F0 slope may be a general cue used by listeners for accent detection. Although we can only speculate as to why slope might be important, it could simply be that F0 is a salient, easily detectable cue for listeners. Future research will need to explore the degree to which listeners are sensitive to slope differences. The fact that listeners only relied upon slope information when the slope was relatively steep may provide insight into the process used by listeners to detect accents. Several possibilities exist. First, this imbalance may result from limitations in listeners’ ability to detect slope. Bolinger (1986) has pointed out that steepness of F0 change makes F0 change more salient. Therefore, listeners’ lack of preferences in the less-steep sloped conditions may reflect difficulty in determining slope direction when it is not steep. When they are unsure of the directionality, they fail to posit the presence of an accent and, therefore, do not use prosodic information to disambiguate the cohort object. Alternately, listeners may be sensitive to the fact that less-steep slopes are more likely to be the result of random variation in F0 and, therefore, only use slope as a cue when it is more extreme. Lastly, listeners may make decisions about accenting by attempting to match the F0 contour (the “shape” of the F0 change) to one of a set of prototypes. Steep-sloped conditions might provide a better match for the correct prototype, resulting in an increased likelihood of deciding that an accent is present

These results also raise questions about accent interpretation. It is unclear what is driving the preference for the given cohort object in the short – low F0 condition. One possibility is that a steep drop in F0 signals de-accenting resulting in an explicit representation of the word as being deaccented which biases listeners towards the given cohort object. Another possibility is that a steep drop in F0 is a reliable indicator of the absence of the high-peaked accent. In this case, the effects seen above reflect fixations away from the new cohort object. This is a subtle difference but one that has important implications for the way that listeners interpret accented and unaccented words. The data here do not differentiate between these two possibilities and either possibility is consistent with the three explanations outlined above for listeners’ lack of a looking preference in the less-steep conditions.

Another potential concern is how the non-accent condition is characterized in this experiment. A drop in F0 is consistent with an L* accent, which Pierrehumbert & Hirschberg (1990) argue signals information that is new or salient, but should not be instantiated in the listener’s discourse model. This raises a puzzling question: why did listeners not interpret the low F0 conditions as signaling new information since a drop an F0 is consistent with an L* accent? We point out that we derived the F0 falling contour from natural productions of non-accented words. Although L* accent may correlate with falling F0, the contour here is consistent with naturally produced non-accented words, and clearly, the data suggest listeners interpret this contour as cueing given information. Future work will need to determine how the drop in F0 associated with non-accented words differs from the contour associated with L*.

Additional questions remain. These results suggest that listeners are using F0 slope to detect the presence of an accent but Ladd, Faulkner, Faulkner, and Schepman (1999) found no consistent relationship between accenting and F0 slope in production. Ladd et al. tested whether F0 contours could be defined in terms of tonal targets that are anchored to points in the speech signal, an idea known as segmental anchoring. Speakers produced a passage under slow, normal, and fast conditions. If segmental anchoring indeed drives accenting, then the presence of tonal targets and their relative alignment should remain the same across speed conditions. This would result in differences in absolute duration and in the F0 slope. On the other hand, if either duration or F0 slope remained constant across the speed conditions then this would be evidence that slope and duration drive accenting. Ladd et al. found no consistent relationship between F0 slope and accenting, supporting the segmental anchoring hypothesis. They conclude that F0 contour in production is determined by F0 alignment and anchoring of tonal targets.

In the present study, the alignment of the F0 peak occurred at the same location within each word in long and short duration conditions since the F0 contour was stretched to match the length of the word. Thus, in the present study at least, differences in the duration conditions are not attributable to differences in F0 alignment. Ladd et al. (1999) found that speakers do not produce consistent F0 slope and yet listeners in our experiment used F0 slope to make decisions about accenting. This raises an interesting question. Are listeners using a cue (F0 slope) that speakers do not produce consistently? Further research needs to explore speakers' production of F0 alignment and slope and listeners' use of these features in comprehension.

Acknowledgments

This work was supported by NIH grant R01DC008774-01A1. The first author was supported by NIH grant T32HD055272. We would like to thank our undergraduate assistants Kathleen Eberhardt and Ahmad Parvinian for help with stimulus preparation and data collection.

References

Allopenna PD, Magnuson JS, Tanenhaus MK. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language. 1998; 38:419–439.

Arnold JE. THE BACON not the bacon: How children and adults understand accented and unaccented noun phrases. Cognition. 2008; 108(1):69–99. [PubMed: 18358460]

Arnold JE, Fagnano M, Tanenhaus MK. Disfluencies signal theee, um, new information. Journal of Psycholinguistic Research. 2003; 32(1):25–36. [PubMed: 12647561]

Bard, EG.; Aylett, MP. The dissociation of deaccenting, givenness, and syntactic role in spontaneous speech. San Francisco: ICPhS-99; 1999.

Bartels C, Kingston J. Salient pitch cues in the perception of contrastive focus. The Journal of the Acoustical Society of America. 1994; 95(5):2973.

Beach CM. The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. Journal of Memory and Language. 1991; 30(6):644–663.

Birch S, Clifton C. Effects of varying focus and accenting of adjuncts on the comprehension of utterances. Journal of Memory and Language. 2002; 47(4):571–588.

Boersma, P.; Weenink, D. Praat: Doing phoenetics by computer. Amsterdam: Institute of Phoenetic Sciences; 2008.

Bolinger D. Accent is predictable (if you're a mind-reader). Language. 1972; 48(3):633–644.

Bolinger, D. Intonation and its parts: Melody in spoken English. Stanford: Stanford University Press; 1986.

Brown, G.; Yule, G. Discourse analysis. Cambridge, UK: Cambridge University Press; 1983.

Buring, D. Intonation, semantics and information structure. In: Ramchand, G.; Reiss, C., editors. The Oxford handbook of linguistic interfaces. Oxford: Oxford University Press; 2007.

CELEX English database. [Retrieved 09/20, 2006] 1993. from http://www.mpi.nl/world/celex

Chen A, Os ED, De Ruiter JP. Pitch accent type matters for online processing of information status: Evidence from natural and synthetic speech. Linguistic Review. Special Issue: Prosodic Phrasing and Tunes. 2007; 24(2–3):317–344.

Cruttenden, A. Intonation. 2nd ed.. Cambridge: Cambridge University Press; 1997.

Dahan D, Tanenhaus MK, Chambers CC. Accent and reference resolution in spoken-language comprehension. Journal of Memory and Language. 2002; 47:292–314.

Dilley LC, Brown M. Effects of pitch range variation on f-sub-0 extrema in an imitation task. Journal of Phonetics. 2007; 35(4):523–551.

Eberhard KM, SpiveyKnowlton MJ, Sedivy JC, Tanenhaus MK. Eye movements as a window into real-time spoken language comprehension in natural contexts. Journal of Psycholinguistic Research. 1995; 24(6):409–436. [PubMed: 8531168]

Grabe E, Kochanski G, Coleman J. Connecting intonation labels to mathematical descriptions of fundamental frequency. Language and Speech, 2007, 50, 3, Oct. 2007; 50(3) Oct.

Halliday, MAK. Intonation and grammar in British English. Paris: Mouton; 1967.

Henderson, JM.; Ferreira, F. Scene perception for psycholinguists. In: Henderson, JM.; Ferreira, F., editors. The interface of language, vision, and action: Eye movements and the visual world. New York, NY: Psychology Press; 2004. p. 1-58.

Isaacs AM, Watson DG. Accenting is more than pitch: Word duration and listeners' preferences for discourse referents. 2007 Unpublished manuscript.

Isaacs, AM.; Watson, DG. Speakers and listeners don't agree: Audience design in the production and comprehension of acoustic prominence. Presented at CUNY Sentence Processing Conference; Davis, CA. 2009.

Ito K, Speer SR. Anticipatory effects of intonation: Eye movements during instructed visual search. Journal of Memory and Language. 2008; 58(2):541–573. [PubMed: 19190719]

Levy, R.; Jaeger, TF. Speakers optimize information density through syntactic reduction. In: Schlökopf, B.; Platt, J.; Hoffman, T., editors. Advances in neural information processing systems (NIPS). Vol. 19. Cambridge, MA: MIT Press; 2007. p. 849-856.

Kochanski G, Grabe E, Coleman J, Rosner B. Loudness predicts prominence: Fundamental frequency lends little. Journal of the Acoustical Society of America. 2005; 118(2):1038–1054. [PubMed: 16158659]

Ladd, DR. Intonational phonology. Cambridge University Press; 1996.

Ladd DR, Morton R. The perception of intonational emphasis: Continuous or categorical? Journal of Phonetics. 1997; 25:313–342.

Ladd DR, Schepman A. "Sagging transitions" between high pitch accents in English: Experimental evidence. Journal of Phonetics. 2003; 31(1):81–112.

Ladd DR, Faulkner D, Faulkner H, Schepman A. Constant "segmental anchoring" of F0 movements under changes in speech rate. The Journal of the Acoustical Society of America. 1999; 106(3 Pt 1): 1543–1554. [PubMed: 10489710]

Marslen-Wilson WD. Functional parallelism in spoken word-recognition. Cognition.Special Issue: Spoken Word Recognition. 1987; 25(1–2):71–102.

Moulines, E.; Verhelst, W. Time-domain and frequency-domain techniques for prosodic modification of speech. In: Kleijn, WB.; Paliwal, KK., editors. Speech coding and synthesis. New York, NY: Elsevier Science; 1995. p. 519

Pierrehumbert, J. PhD, MIT; 1980. The phonology and phonetics of English intonation.

Pierrehumbert, J.; Hirschberg, J. The meaning of intonational contours in the interpretation of discourse. In: Cohen, PR.; Morgan, J.; Pollack, ME., editors. Intentions in communication. Cambridge: MIT Press; 1990. p. 271

Repp BH. Perceptual coherence of speech: Stability of silence-cued stop consonants. Journal of Experimental Psychology: Human Perception and Performance. 1985; 11(6):799–813. [PubMed: 2934509]

Rossion, B.; Pourtois, G. Revisiting Snodgrass and Vanderwart's object database: Color and texture improve object recognition. 1st Vision Conference; Sarasota, FL. 2001.

Schwarzschild R. GIVENness, AVOIDF and other constraints on the placement of accent. Natural Language Semantics. 1999; 7(2):141–177.

Selkirk, E. Sentence prosody: Intonation, stress and phrasing. In: Goldsmith, JA., editor. The handbook of phonological theory. London: Blackwell; 1995. p. 550-569.

Silverman, K.; Beckman, ME.; Pitrelli, J.; Ostendorf, M.; Wightman, C.; Price, P., et al. ToBI: A standard for labeling English prosody. Paper Presented at the Second International Conference on Spoken Language Processing; Banff, Canada. 1992.

Snodgrass JS, Vanderwart M. A standardized set of 260 pictures: Norms for name agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory. 1980; 6:174–215. [PubMed: 7373248]

Tanenhaus MK, SpiveyKnowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. Science. 1995; 268(5217):1632–1634. [PubMed: 7777863]

Terken J, Nooteboom SG. Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. Language and Cognitive Processes. 1987; 2(3–4):145–163.

Watson, DG.; Gunlogson, CA.; Tanenhaus, MK. Online methods of the investigation of prosody. In: Steube, A., editor. Methods in empirical prosody research. New York: Walter de Gruyter; 2006. p. 259-283.

Watson DG, Arnold JE, Tanenhaus MK. Tic tac TOE: Effects of predictability and importance on acoustic prominence in language production. Cognition. 2008; 106(3):1548–1557. [PubMed: 17697675]

Weber A, Braun B, Crocker MW. Finding referents in time: Eye-tracking evidence for the role of contrastive accents. Language and Speech. 2006; 49(3):367. [PubMed: 17225671]

Welby P. The role of early fundamental frequency rises and elbows in French word segmentation. Speech Communication, 2007, 49, 1, Jan. 2007; 49(1):28–48.

Appendix

Pictures used for each trial

New Cohort Object	Given Cohort Object	Distractor1	Distractor2
Bell	Bed	Apple	Pliers
Clown	Cloud	Skeleton	Rolling Pin
Skate	Scale	Wagon	Rocket
Rabbit	Raccoon	Calculator	Hanger
Sandwich	Sandal	Barrel	Chocolate bar
Candle	Camel	Whistle	Flamingo
Rooster	Ruler	Vest	Guitar
Pencil	Penguin	Dove	Elk
Rope	Road	Flower vase	Umbrella
Chair	Chain	Backpack	Viola
Plate	Plane	Acorn	Snowman
Pillow	Pickle	Helicopter	Watch
Robe	Rose	Strawberry	Gun

New Cohort Object	Given Cohort Object	Distractor1	Distractor2
Doll	Dog	Mountain	Net
Chicken	Chimney	Jellyfish	Flower
Carriage	Carrot	Policecar	Owl
Sheep	Shield	Giraffe	Violin
Cane	Cake	Shirt	Lemon
Cork	Corn	Ketchup	Nail
Lamb	Lamp	Spinning wheel	Light

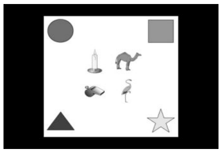

Figure 1.
Display at the onset of a critical trial

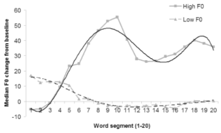

Figure 2.

The median F0 values for Accented (High F0) and Unaccented (Low F0) items in 20 equally-spaced segments are shown in grey. The graph of the polynomial equations generated to fit these points is shown in black.

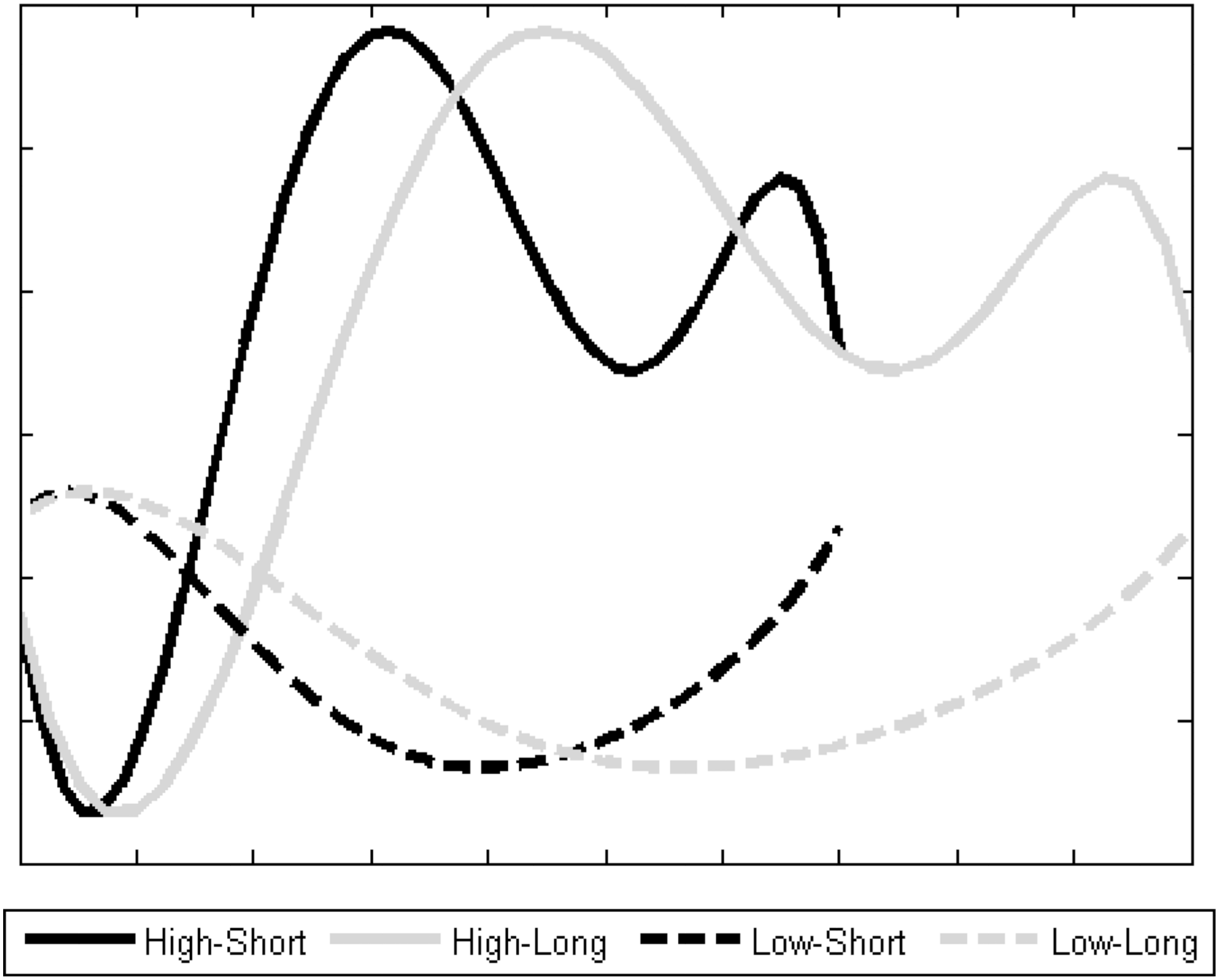

Figure 3.
Graphical representations of the functions used to generate F0 contours for the high and low F0 conditions with either long or short duration

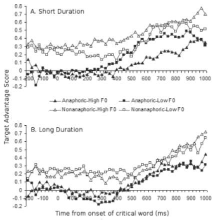

Figure 4.

Time course plot of Target Advantage scores plotted by information status (anaphoric/non-anaphoric) and F0 (high/low) for conditions with Short duration (A) and Long duration (B) starting at the onset of the critical word (0ms)

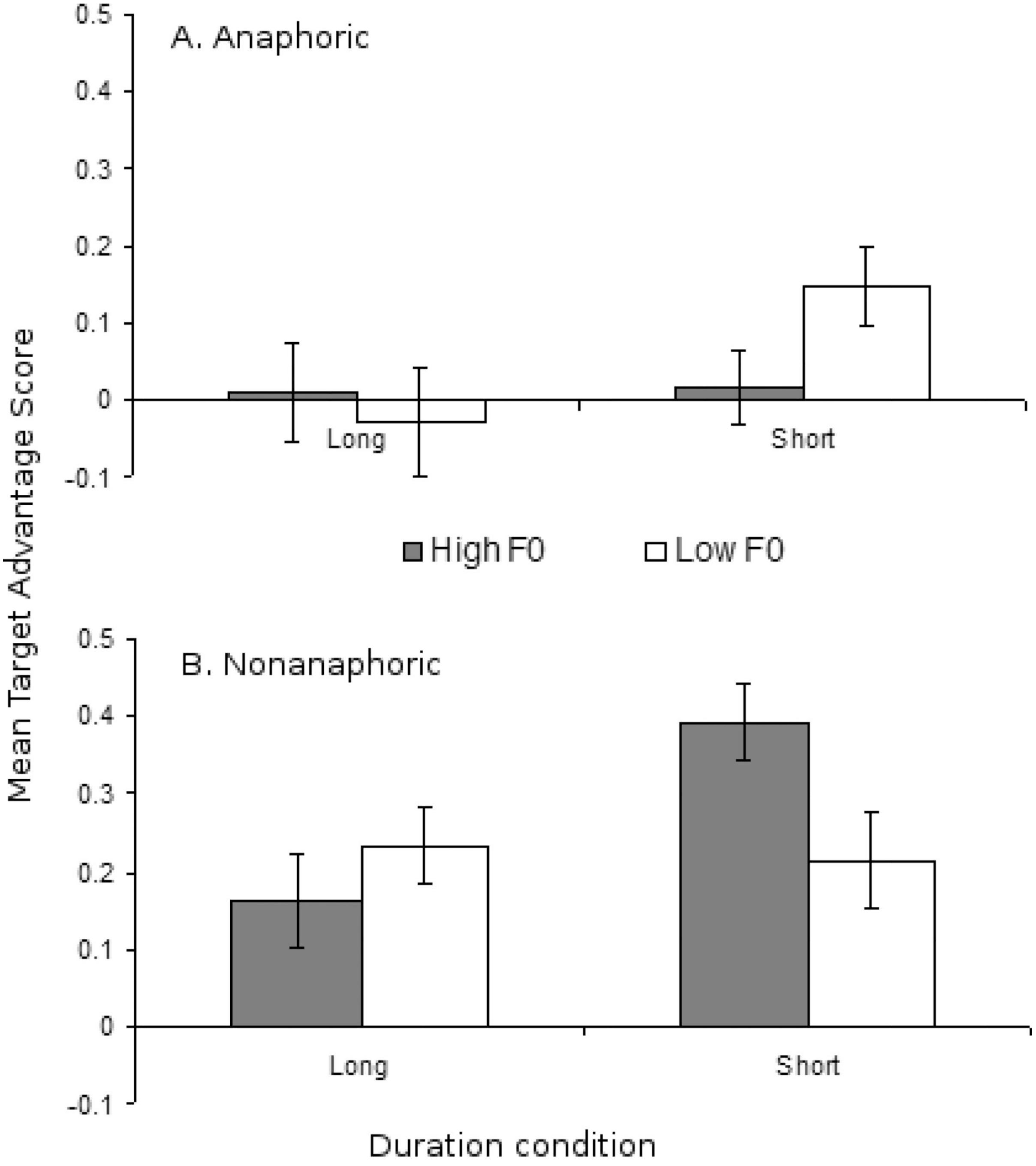

Figure 5.

Mean Target Advantage Scores and standard error values plotted by F0 (high/low) and Duration (long/short) in the Anaphoric (A) and Non-anaphoric conditions (B)