Acoustic correlates of information structure

Mara Breen[a]; Evelina Fedorenko[b]; Michael Wagner[c]; Edward Gibson[b]

[a] Department of Psychology, University of Massachusetts, Amherst, MA, USA [b] Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA [c] Department of Linguistics, McGill University, Montreal, Canada

Online publication date: 06 October 2010

PLEASE SCROLL DOWN FOR ARTICLE

Ψ **Psychology Press**
Taylor & Francis Group

Acoustic correlates of information structure

Mara Breen[1], Evelina Fedorenko[2], Michael Wagner[3], and Edward Gibson[2]

[1]*Department of Psychology, University of Massachusetts, Amherst, MA, USA,*
[2]*Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA,* [3]*Department of Linguistics, McGill University, Montreal, Canada*

This paper reports three studies aimed at addressing three questions about the acoustic correlates of information structure in English: (1) do speakers mark information structure prosodically, and, to the extent they do; (2) what are the acoustic features associated with different aspects of information structure; and (3) how well can listeners retrieve this information from the signal? The information structure of subject–verb–object sentences was manipulated via the questions preceding those sentences: elements in the target sentences were either focused (i.e., the answer to a wh-question) or given (i.e., mentioned in prior discourse); furthermore, focused elements had either an implicit or an explicit contrast set in the discourse; finally, either only the object was focused (narrow object focus) or the entire event was focused (wide focus). The results across all three experiments demonstrated that people reliably mark (1) focus location (subject, verb, or object) using greater intensity, longer duration, and higher mean and maximum F0, and (2) focus breadth, such that narrow object focus is marked with greater intensity, longer duration, and higher mean and maximum F0 on the object than wide focus. Furthermore, when participants are made aware of prosodic ambiguity present across different information structures, they reliably mark focus type, so that contrastively focused elements are produced with greater intensity, longer duration, and lower mean and maximum F0 than noncontrastively focused elements. In addition to having important theoretical consequences for accounts of semantics and prosody, these experiments demonstrate that linear residualisation successfully removes individual differences in people's productions thereby revealing cross-speaker generalisations. Furthermore,

discriminant modelling allows us to objectively determine the acoustic features that underlie meaning differences.

Keywords: Prosody; Information structure.

INTRODUCTION

An important component of the meaning of a sentence is its relationship to the context in which it is produced. Some parts of speakers' sentences refer to information already under discussion, while other parts convey information that the speaker is presenting as new for the listener. Depending on the context, the same sentence can convey different kinds of information to the listener. For example, consider the three contexts in (1a)–(1c) for the sentence in (2):

(1) a. Who fried an omelet?
     b. What did Damon do to an omelet?
     c. What did Damon fry?
(2) Damon fried an omelet.

The event of frying an omelet is already made salient in the context in (1a), and this part of the answer is therefore *given*. Consequently, the sentence *Damon fried an omelet* conveys *Damon* as the *new* or *focused* information.[1] The verb *fried* is the focused information relative to the context in (1b), and the object noun phrase (NP) *an omelet* is the focused information relative to the context in (1c). This component of the meaning of sentences—the differential contributions of different sentence elements to the overall sentence meaning in its relation to the preceding discourse—is called *information structure*.

Three components of information structure have been proposed in the literature: *givenness*, *focus*, and *topic* (see e.g., Féry & Krifka, 2008, for a recent summary). The current paper will be concerned with givenness and focus.[2] Given material is material that has been made salient in the discourse, either explicitly, like the event corresponding to the verb *fried* and the object corresponding to the noun *omelet* in (1a), or implicitly, via inferences based

---

[1] Numerous terms are used in the literature to refer to the distinction between the information that is old for the listener and the information that the speaker is adding to the discourse: *background* and *foreground*; *given* and *new*; *topic* and *comment*; *theme* and *rheme*, etc. In this paper, we will use the term *given* to refer to the parts of the utterance which are old to the discourse, and *focused* to refer to the part of the utterance which is new to the discourse.

[2] Topic, the third component of information structure, describes which discourse referent focused information should be associated with, as in the mention of Damon in "As for Damon, he fried an omelet". The current studies do not address the prosodic realisation of topic.

on world knowledge (e.g., mentioning *omelet* makes the notion of "eggs" given, Schwarzchild, 1999).

Focused material is what is new to the discourse, or in the foreground. The focus of a sentence can often be understood as the part that corresponds to the answer to the wh-part of wh-questions, like *Damon* in (2) as an answer to (1a) (Jackendoff, 1972; Paul, 1880).

There are two dimensions along which focused elements can differ. The first is contrastiveness. A contrastively focused element, like *Damon* in (3b), indicates that the element in question is one of a set of explicit alternatives or serves to correct a specific item already present in the discourse, as in the following:

(3) a. Did Harry fry an omelet yesterday?
     b. Damon fried an omelet yesterday.

Unlike (1a), where there is no explicit set of individuals from which *Damon* is being selected as the "omelet fryer", in (3a) an explicit alternative "omelet fryer" is being introduced: *Harry.* The sentence (3b) in this context thus presents information (i.e., *Damon*) which explicitly contrasts with, or contradicts, some information which has been introduced into the discourse.

There is no consensus in the literature regarding the relationship between noncontrastive focus and contrastive focus. Some researchers have treated noncontrastive focus and contrastive focus as separate categories of information structure (Chafe, 1976; Halliday, 1967; Molnar, 2002; Rochemont, 1986), whereas others have argued that there is no principled difference between the two (e.g., Bolinger, 1961; Rooth, 1985, 1992). According to Rooth (1992), for example, each expression evokes two semantic representations: the expression's actual meaning and a set of alternatives. If a constituent in the expression is focused, then the alternative set contains the expression itself and all expressions with an alternative substituted for the focus-marked constituent; if there is no focus within the expression, the alternative set consists only of the expression itself. Rooth would therefore argue that *Damon* in (1a) is focused and introduces alternative propositions that differ only in the agent of the event ({Damon fried an omelet, Harry fried an omelet, Ada fried an omelet, ...}), even if no alternatives are explicitly mentioned. In (3a), *Damon* also evokes alternative omelet fryers, and therefore has the same focus structure as (1a), but the context makes a specific alternative (*Harry*) more salient than other potential alternatives. Importantly, from Rooth's standpoint, it does not matter whether the alternatives are explicit in the discourse or not: the meaning of the expression is the same.

   The second dimension along which focused elements can vary is focus breadth (Gussenhoven, 1983, 1999; Selkirk, 1984, 1995), which refers to the size of the set of focused elements. *Narrow* focus refers to cases where only a single aspect of an event (e.g., the agent, the action, the patient, etc.) is focused, whereas *wide* focus focuses an entire event. Take, for example, the difference between (5) as an answer to (4a) vs. as an answer to (4b):

   (4) a. What did Damon fry last night?
       b. What happened last night?
   (5) Damon fried an omelet last night.

(4a) narrowly focuses the patient of frying, *omelet* in (5), whereas (4b) widely focuses the entire event of Damon frying an omelet.

   The information structure of a sentence element can be conveyed in at least three ways: (1) using word order (i.e., given information generally precedes focused information; e.g., Birner, 1994; Clark & Clark, 1978); (2) using particular lexical items and syntactic constructions (e.g., using cleft constructions such as *It was Damon who fried an omelet*; Lambrecht, 2001); and (3) using prosody. Prosody—which we focus on in the current paper— refers to the way in which words are grouped in speech, the relative acoustic prominence of words, and the overall tune of an utterance. Prosody is comprised of acoustic features such as fundamental frequency (F0), duration, and loudness, the combinations of which give rise to psychological percepts such as phrasing (grouping), stress (prominence), and tonal movement (intonation).

   The goal of the current paper is to investigate the prosodic realisation of information structure in simple English subject–verb–object (SVO) sentences like (2), with the goal of addressing the following questions:

   (1) First, do speakers mark information structure prosodically?
       a. Do speaker distinguish focused and unfocused elements?
       b. Do speakers distinguish focused elements that have an explicit contrast set in the discourse from those that do not?
       c. Do speakers distinguish sentences in which only the object is focused from those in which the entire event is focused?
   (2) What are the acoustic features associated with these different aspects of information structure?
   (3) How well can listeners retrieve this information from the signal?

Although the current experiments are all performed on English, the answers to these questions will likely be similar for other West Germanic languages. However, the relationship between prosodic features and information

structure across different languages and language groups remains an open question.

In the remainder of the introduction, we briefly lay out two approaches to the study of the relationship between prosody and information structure, and summarise empirical studies which have explored how information structure is realised acoustically and prosodically. We then discuss methodological issues present in previous studies which call into question the general-iseability of the reported findings, and outline how the current methods were designed to better address these questions.

Empirical investigations of prosody and information structure

Two perspectives on the relationship between the acoustics of the speech signal and the meaning associated with various aspects of information structure have been articulated in the literature. According to the *direct-relationship* approach, sets of acoustic features are directly associated with particular meanings (Cooper, Eady, & Mueller, 1985; Eady & Cooper, 1986; Fry, 1955; Lieberman, 1960; Pell, 2001; Xu & Xu, 2005). In contrast, according to the *indirect-relationship* approach (known as the intonational phonology framework), the relationship between acoustics and meaning is mediated by phonological categories (Dilley, 2005; Gussenhoven, 1983; Hawkins & Warren, 1991; Ladd, 1996; Pierrehumbert, 1980). In particular, phonetic prosodic cues are hypothesised to be grouped into prosodic categories which are, in turn, associated with particular meanings. The experiments in the current paper were not designed to decide between these two approaches. However, we will initially discuss our experiments in terms of the direct-relationship approach, because it is more parsimonious. In the general discussion, we will show how the results are also compatible with the indirect-relationship approach.

Turning now to previous empirical work on the relationship between prosody and information structure, we start with studies of focused vs. given elements. Several studies have demonstrated that focused elements are more acoustically prominent than given elements. However, there has been some debate about which acoustic features underlie a listener's perception of acoustic prominence. Some features that have been proposed to be associated with prominence include pitch (i.e., F0; Cooper, Eady, & Mueller, 1985; Eady & Cooper, 1986; Lieberman, 1960), duration (Beckman, 1986; Fry, 1955), loudness (i.e., intensity; Beckman, 1986; Kochanski, Grabe, Coleman, & Rosner, 2005; Turk & Sawusch, 1996), and voice quality (Sluijter & van Heuven, 1996).

In early work on lexical stress, Fry (1955) and Lieberman (1960) argued that intensity and duration of the vowel of the stressed syllable contributed most strongly to the percept of acoustic prominence, such that stressed

vowels were produced with a greater intensity and a longer duration than nonstressed vowels. In experiments on phrase-level prominence, Cooper Eady, and Mueller (1985) and Eady and Cooper (1986) also noted that more prominent syllables are longer than their nonprominent counterparts. Cooper et al. (1985) (see also Lieberman, 1960), Rietveld and Gussenhoven (1985), Gussenhoven, Repp, Rietveld, Rump, and Terken (1997), and Terken (1991) argued that F0 was a highly important acoustic feature underlying prominence. Others have argued that the strongest cue to prominence is intensity (e.g., Beckman, 1986). More recently, Turk and Sawusch (1996) also found that intensity (and duration) were better predictors of perceived prominence than pitch, in a perception task. Finally, in a study of spoken corpora, Kochanski et al. (2005) demonstrated that loudness (i.e., intensity) was a strong predictor of labellers' annotations of prominence, while pitch had very little predictive power.

The question of whether contrastively and noncontrastively focused elements are prosodically differentiated by speakers, and perceptually differentiated by listeners, has also been extensively debated. Some have argued that there is no difference in the acoustic features associated with contrastively vs. noncontrastively focused elements (Bolinger, 1961; Cutler, 1977; 't Hart, Collier, & Cohen, 1990), while others have argued that some acoustic features differ between contrastively vs. noncontrastively focused elements (Bartels & Kingston, 1994; Couper-Kuhlen, 1984; Ito, Speer, & Beckman, 2004; Krahmer & Swerts, 2001). For example, Couper-Kuhlen (1984) reported, on the basis of corpus work, that speakers produce contrastive focus with a steep drop after a high F0 target, while high F0 is sustained after noncontrastive focus (see also Krahmer & Swerts, 2001). However, this finding is in contrast to Bartels and Kingston (1994), who have argued, based on a pair of concurrent perception production studies, that the most salient acoustic cue to contrastiveness is the height of the peak on a contrastive word, such that a higher peak is associated with a greater probability of an element being interpreted as contrastive (see also Ladd & Morton, 1997). Finally, Ito et al. (2004) demonstrated that speakers are more likely to use a $L + H^*$ accent (i.e., a steep rise from a low target to a high target), compared to a H^* accent (i.e., a gradual rise to a high target), to indicate an element that has an explicit contrast set in the discourse.

Krahmer and Swerts (2001) observed that listeners were more likely to perceive a contrastive adjective (e.g., *red* in *red square* preceded by *blue square*) as more prominent than a new adjective when the adjective was presented with a noun compared to when it was presented in isolation. They therefore hypothesised that the lack of a consensus in the literature may be due to the failure of the earlier studies to investigate focused elements in relation to the prosody of the surrounding elements. Consistent with this idea, Calhoun (2005) demonstrated that a model's ability to predict a word's

information structure is significantly improved when information about the acoustics of adjacent words is included in the model. These results suggest that a more consistent picture of the acoustic features associated with contrastively and noncontrastively focused elements may emerge if acoustic context is taken into account.

Finally, prior work has investigated whether speakers prosodically differentiate narrow and wide focus. Selkirk (1995), for example, argued that, through a process called *focus projection*, an acoustic prominence on the head of a phrase or its internal argument can project to the entire phrase, thus making the entire phrase focused (see also Selkirk, 1984; see Gussenhoven, 1983, 1999, for a similar claim). According to Selkirk (1984) and Gussenhoven (1983) then a clause containing a transitive verb whose direct object is acoustically prominent is ambiguous between a reading where the object alone is focused and a reading where the entire verb phrase is focused. This hypothesis has been supported in several perception experiments (Birch & Clifton, 1995; Gussenhoven, 1983; Welby, 2003). Welby (2003), for example, demonstrated that listeners rated a sentence like *I read the DISPATCH* with a single acoustic prominence on *dispatch* as a similarly felicitous response to either a question narrowly focusing the object (i.e., *What newspaper do you read?*), or a question widely focusing the entire event (i.e., *How do you keep up with the news?*). However, Gussenhoven (1983) found that, at least in some productions, there is actually a perceptible difference between narrow and wide focus, although listeners cannot use this information to reliably tell in which context the sentence was uttered (see Baumann, Grice, & Steindamm, 2006, for evidence from German showing that speakers do differentiate between narrow and wide focus, with prosodic cues varying across speakers). In contrast to Gussenhoven's perception results, Rump and Collier (1986) found that listeners can accurately discriminate narrow and wide focus using pitch cues.

Limitations of previous work

Although the studies summarised above provide evidence for some systematic differences in the acoustic realisation of different aspects of information structure, no clear picture has yet emerged with regard to any of the three meaning distinctions discussed above (i.e., focused vs. given elements, noncontrastively focused vs. contrastively focused elements, and narrow vs. wide focus). Furthermore, previous studies suffer from several methodological limitations that make the findings inconclusive. Here, we discuss five limitations of previous studies which the current studies seek to address in an effort to reveal a clearer picture of the relationship between acoustic features and information structure.

First, instead of acoustic features, sometimes only Tones and Break Indices (ToBI)[3] annotations are provided (e.g., Birch & Clifton, 1995; Ito et al., 2004), as in the work of researchers who adopt the intonational phonology framework and who therefore believe that using prosodic annotation offers a useful way to extrapolate away from potentially complex interactions among acoustic features which give rise to the perception of specific intonational patterns. One particular problem concerns H* and L+H* accents. As defined in the ToBI system, these accents are meant to be explicit markers of noncontrastive focus and contrastive focus, respectively (Beckman & Ayers-Elam, 1997). However, H* and L+H* are often confused in ToBI annotations (Syrdal & McGory, 2000), and are, in fact, often collapsed in calculating intercoder agreement (Breen, Dilley, Gibson, Bolivar, & Kraemer, 2006; Pitrelli, Beckman, & Hirschberg, 1994; Yoon, Chavarria, Cole, & Hasegawa-Johnson, 2004). Therefore, it is difficult to interpret the results of studies which are based on the difference between H* and L+H* without a discussion of the acoustic differences between these purported categories. In the current studies, we report acoustic features in order to avoid confusion about what the ToBI labels might mean and in order to not presuppose the existence of prosodic categories associated with particular meaning categories of information structure.

A second limitation concerns the method used to generate and select productions for analysis. A common practice involves eliciting productions from a small number of speakers (e.g., Baumann et al., 2006; Krahmer & Swerts, 2001), which results in a potential decrease in experimental power, and could therefore lead to a Type II error. In addition, several previous experiments have excluded speakers' data from analysis for not producing accents consistently (e.g., Cooper et al., 1985; Eady & Cooper, 1986), which could lead to a Type I error. For the current experiments, we recruited between 13 and 18 speakers. In addition, no speakers' productions were excluded from the analyses based on a priori predictions about potential behaviour (e.g., placing accents in particular locations).

A third limitation concerns the tasks used in perception studies. In particular, some studies asked listeners to make judgements about which of two stimuli were more prominent (Krahmer & Swerts, 2001), what accent is acceptable in a particular context (Birch & Clifton, 1995; Welby, 2003), or with which of two questions a particular answer sounded more natural (Gussenhoven, 1983). The problem with these metalinguistic judgements is that they lack a measure of the participants' interpretation of the sentences. In the current studies, we employ a more natural production–comprehension task, in which speakers are trying to communicate a particular meaning of a

---

[3] The ToBI system was developed in the early 1990s as the standard system for annotation of prosodic features (Silverman et al., 1992).

semantically ambiguous sentence and listeners are trying to understand the intended meaning.

A fourth limitation of previous studies is in how they have dealt with speaker variability. Presenting data from individual subjects separately, as is commonly done, is problematic because it fails to capture the shared aspects of individual productions (e.g., consistent use by most speakers of some set of acoustic features to mark focused elements). In the current studies, we combine data across subjects while simultaneously removing variance due to individual differences using linear regression modelling (e.g., Jaeger, 2008).

A fifth limitation alluded to in the previous section is that many have reported differences between conditions based only on individual acoustic features on single words (Baumann et al., 2006; Cooper et al., 1985; Eady & Cooper, 1986; cf. Krahmer & Swerts, 2001; Calhoun, 2005). If acoustic prominence is perceived in a context-dependent manner, these single-feature/single-word analyses might find spurious differences, or fail to find real differences. In the current studies, we used discriminant modelling on the productions in order to simultaneously investigate the contribution of multiple acoustic features from multiple words in an utterance to the interpretation of information structure of different sentence elements.

EXPERIMENTS: OVERVIEW AND GENERAL METHODS

The current paper presents results from three experiments. Experiment 1 investigated whether speakers prosodically disambiguate focus location (subject vs. verb vs. object), focus type (contrastive vs. noncontrastive), and focus breadth (narrow vs. wide) by eliciting semi-naturalistic productions like that in (3b) (e.g., *Damon fried an omelet this morning*), whose information structure was disambiguated by a preceding question. Experiment 2 investigated whether speakers disambiguate focus location and focus type when the task explicitly required them to communicate a particular meaning to their listeners. Finally, Experiment 3 served as a replication and extension of Experiment 2, in which speakers included an attribution expression (*I heard that*) before the critical sentence.

The acoustic analysis of the productions elicited in all three experiments proceeded in three steps. First, we automatically extracted a series of 24 acoustic features (see Table 2) from the subject, verb, and object of the sentences elicited in Experiments 1–3. Second, we subjected all of these features to a stepwise discriminant function analysis in order to determine which features best discriminated the information structure conditions listed in Table 1 for each of the three experiments. This analysis resulted in a subset of eight acoustic features. Finally, we used discriminant analyses to evaluate whether this subset of eight features could effectively discriminate sets of 2

TABLE 1

Example item from Experiment 1. The target sentence is ''Damon fried an omelet this morning''

Condition	Focus type	Focused argument	Setup question
1	Noncontrastive	Wide	What happened this morning?
2	Noncontrastive	S	Who fried an omelet this morning?
3	Noncontrastive	V	What did Damon do to an omelet this morning?
4	Noncontrastive	O	What did Damon fry this morning?
5	Contrastive	S	Did Harry fry an omelet this morning?
6	Contrastive	V	Did Damon bake an omelet this morning?
7	Contrastive	O	Did Damon fry a chicken this morning?

and 3 conditions for each of the three experiments. Specifically, we tested focus location by comparing the features from productions in which each of the elements (subject, verb, object) was focused vs. given. We tested focus type by comparing the features from sentences in which the focused element was contrastively or noncontrastively focused at each of the three syntactic positions. Last, we tested focus breadth by comparing the features for sentences with wide focus to those with narrow object focus. In addition to the analysis of acoustic features, in Experiments 2 and 3 we investigated whether listeners could correctly determine the intended information structure.

EXPERIMENT 1

Method

Participants

Nine pairs of participants were recorded. All participants were self-reported native speakers of American English. All participants were MIT students or members of the surrounding community. Participants were paid for their participation.

Materials

Each trial consisted of a set-up question and a target sentence, which always had an SVO structure (e.g., *Damon fried an omelet this morning*). The target sentence could plausibly answer any one of the seven set-up questions (see Table 1), which served to focus different elements of the sentence or the entire event described in the sentence. The first condition focused the entire event (i.e., *What happened this morning?*). In the remaining conditions, two factors

were manipulated: (1) the element in the target sentence that was focused by the question (subject, verb, or object) and (2) the presence of an explicit contrast set for the focused element (absent, i.e., noncontrastively focused element, or present, i.e., contrastively focused element).

All subject and object NPs in the target sentences were bisyllabic with first syllable stress, and all verbs were monosyllabic. All subject NPs were proper names, and object NPs were mostly common inanimate objects, such that the events were nonreversible. Furthermore, all words were comprised mostly of sonorant phonemes. These constraints ensured that words could be more easily compared across items, and facilitated the extraction of acoustic features (which is easier for vowels and sonorant consonants). An adjunct prepositional phrase (PP) was included at the end of each sentence so that differences in the production of the object NP due to the experimental manipulations would be dissociable from prosodic effects on phrase-final, or in this case, sentence-final, words, which are typically lengthened and produced with lower F0 compared to phrase-medial words (e.g., Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992).

We constructed 28 sets of materials. Participants saw one condition of each item, following a Latin Square design. A sample item is presented in Table 1. The complete set of materials can be found in Appendix 1.

Procedure

Productions were elicited and prescreened in a two-part procedure. The first part was a training session, where participants learned the intended names for pictures of people, actions, and objects. In the second part, the pairs of participants produced questions and answers for each other. The method was designed to maximise control over what speakers were saying, but also to encourage natural-sounding productions. Pilot testing revealed that having subjects simply read the target sentences resulted in productions with low prosodic variability. After going through the experiment one time, the participants switched roles.

Training session.   In the training session, participants learned mappings between 96 pictures and names, so that they could produce the names from memory during the second part of the experiment. In a PowerPoint presentation, each picture, corresponding to a person, an action, an object, or a modifier, was presented with its intended name (see Figure 1, left). The pictures consisted of eight names of people, which were repeated 3–4 items each in the experimental materials, eight colours (which were used in a concurrently run filler experiment), 34 verbs, 44 objects, and two temporal modifiers (*this morning* and *last night*). The pictures were presented in
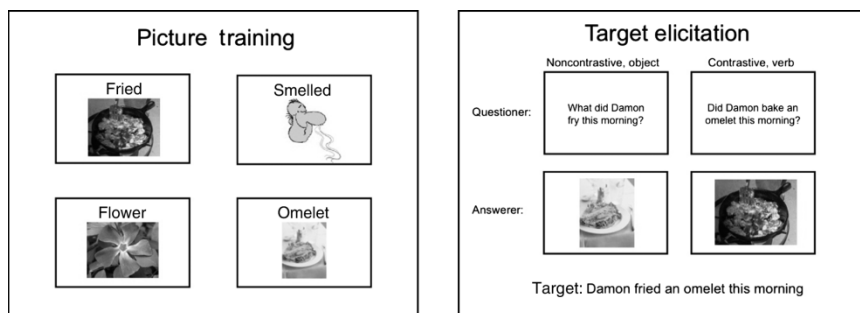
Figure 1. Left: Examples from the picture-training task for Experiment 1. Each square represents a screen shot. Right: Examples of the procedure for the questioner (upper squares) and answerer (lower squares) for Experiment 1. Two conditions are presented: Noncontrastive, object (left), and contrastive, verb (right). The top squares represent screen shots of what the questioner saw on a trial; the bottom squares represent what the answerer saw on a trial.

alphabetical order, to facilitate memorisation and recall. Participants were instructed to learn the mappings by progressing through the PowerPoint at their own pace.

When participants felt they had learned the mappings, they were given a picture-naming test, which consisted of 27 items from the full list of 96. The test was identical for all participants. Participants were told of their mistakes, and, if they made four or more errors, they were instructed to go back through the PowerPoint to improve their memory of the picture–name mappings. Once participants could successfully name 23 or more items on the test, which took between 1 and 3 rounds of testing, they continued with the second part of the experiment. Early in pilot testing, we discovered that subjects had poor recall for the names of the people in the pictures. Therefore, in the actual experiment, subjects could refer to a sheet which had labelled pictures of the people.

Question–answer experiment. The experiment was conducted using Linger 2.92 (available at http://telab.mit.edu/~dr/Linger/), a software platform designed by Doug Rohde for language processing experiments. Participants were randomly paired and randomly assigned to the role of *questioner* or *answerer*. Participants sat at computers in the same room such that neither could see the other's screen. On each trial, as illustrated in Figure 1 (right), the questioner saw a question (e.g., *What did Damon fry this morning?*) which he/she was instructed to produce aloud for the answerer. The answerer was instructed to produce an answer aloud using the information contained in the picture on his/her screen (e.g., *Damon fried an omelet this morning*). The answerer was instructed to produce complete

sentences, including the subject, verb, object, and temporal modifier,[4] and to emphasise the part of the sentence that the questioner had asked about, or that he/she was correcting. On a random 20% of trials, the answerer was asked a comprehension question about the answer s/he produced.

Productions were recorded in a quiet room with a head-mounted microphone at a rate of 44 kHz (Table 2).

Results

Of the 504 speaker productions, 87 (17%) were discarded because (1) the answerer failed to use the correct lexical items; (2) the answerer was disfluent; or (3) the production was poorly recorded. The 417 remaining productions were subjected to the acoustic analyses described below.

Acoustic features

Based on previous investigations of prosody and information structure (Bartels & Kingston, 1994; Baumann et al., 2006; Eady & Cooper, 1986; Cooper et al., 1985; Fry, 1955; Krahmer & Swerts, 2001; Lieberman, 1960), we chose a set of acoustic features to analyse (see Table 2). These features were extracted from the productions automatically using the Praat program (Boersma & Weenink, 2006). The measures of F0 computed over portions of the words (e.g., First quarter F0) were chosen in order to investigate how F0 changes across the syllable might contribute to the differentiation of conditions.

Our first goal was to determine which of the 24 candidate acoustic features mediated differences among conditions. We conducted a series of stepwise linear discriminant analyses (LDA)[5] on all of the data collected in Experiments 1–3 reported in the current paper. In order to determine the features to be used in the analyses of all three experiments, we performed a separate stepwise

---

[4] In the absence of explicit instruction to produce complete sentences, with a lexicalised subject, verb, and object, speakers would likely resort to pronouns or would omit given elements altogether (e.g., *What did Damon fry this morning? An omelet*). A complete production account of information structure meaning distinctions should include not just the prosodic cues used by the speakers, but also syntactic and lexical production choices, as well as the interaction among these different production strategies. However, because we focus on prosody in the current investigation, we wanted to be able to compare acoustic features across identical words. Thus, we required that participants always produce a subject, verb, object, and temporal modifier on every trial.

[5] LDA calculates a function, computed as a linear combination of all predictors entered, which results in the best separation of two or more groups. For two groups, only one function is computed. For three groups, the first function provides the best separation of Group 1 from Groups 2 and 3; a second, orthogonal, function provides the best separation of Groups 2 and 3, after partialling out variance accounted for by the first function. *Stepwise* LDA is an iterative procedure which adds predictors based on which of the candidate predictors provide the best discrimination.

TABLE 2
Acoustic features extracted from each word in the target sentence for Experiments 1–3.
Stepwise discriminant analyses demonstrated that the measures in bold provided the
best discrimination among conditions and were used in all reported analyses

Acoustic feature	Units	Description
Duration	ms	Word duration excluding any silence before or after the word
Silence	ms	Duration of silence following the word, not due to stop closure
Duration+silence	**ms**	**The sum of the duration of the word and any following silence**
Mean F0	**Hz**	**Mean F0 of the entire word**
Maximum F0	**Hz**	**Maximum F0 value across the entire word**
F0 peak location	0–1	The proportion of the way through the word where the maximum F0 occurs
Minimum F0	Hz	Minimum F0 across the entire word
F0 valley location	0–1	The proportion of the way through the word where the minimum F0 occurs
Initial F0	Hz	Mean F0 of the initial 5% of the word
Early F0	Hz	Mean F0 value of 5% of the word centered at the point 25% of the way through the word
Center F0	Hz	Mean F0 value of 5% of the word centered on the midpoint of the word
Late F0	Hz	Mean F0 value of 5% of the word centered on a point 75% of the way through the word
Final F0	Hz	Mean F0 of the last 5% of the word
First quarter F0	Hz	The difference between initial F0 and early F0
Second quarter F0	Hz	The difference between early F0 and center F0
Third quarter F0	Hz	The difference between center F0 and late F0
Fourth quarter F0	Hz	The difference between late F0 and final F0
Mean intensity	dB	Mean intensity of the word
Maximum intensity	**dB**	**Maximum dB level in the word**
Minimum intensity	dB	Minimum dB level in the word
Intensity peak location	0–1	The proportion of the way through the word where the maximum intensity occurs
Intensity valley location	0–1	The proportion of the way through the word where the minimum intensity occurs
Maximum amplitude	Pascal	Maximum amplitude across the word
Energy	$(Pascal)^2 \times$ Duration	

analysis on the data from each experiment separately. For each analysis we
entered all 24 acoustic features across each of the three sentence positions
(subject, verb, and object) as possible predictors of the seven experimental
conditions, resulting in 72 predictors. Across the three analyses, the acoustic
features which consistently resulted in the best discrimination of conditions
were (1) duration + silence; (2) mean F0; (3) maximum F0; and (4) maximum

intensity on subject, verb, and object. The fact that these 12 features (four acoustic features across three sentence positions) consistently discriminated among conditions across three independent sets of productions (from different speakers and across somewhat different sets of materials) serves as evidence that these features underlie the differentiation of information structure across speakers and materials. Therefore, we use only these 12 features in the LDA reported for the individual experiments in the paper.

Computing residual values

Because of differences among individuals, including age, gender, speech rate, and level of engagement with the task, speakers produce very different versions of the same sentence even within the same experimental condition, thus adding variance to the acoustic features of interest. Similarly, there is likely to be variability associated with different items due to lexical and world knowledge factors. Researchers have previously dealt with the issue of acoustic variability between speakers by normalising pitch and/or duration by speaker (e.g., Shriberg, Stolcke, Hakkani-Tur, & Tur, 2000; Shriberg et al., 1998; Wightman et al., 1992). In order to remove speaker- and item-related variance in the current studies, we computed linear regression models in which speaker (e.g., $n=18$ in Experiment 1) and item (e.g., $n=28$ in Experiment 1) predicted each of the 12 acoustic features identified in the stepwise discriminant analyses described in the previous section. From each of these models, we calculated the predicted value of each acoustic feature for a specific item from a specific speaker. We then subtracted this predicted value from every production. The differences among the resulting residual values should reflect differences in the acoustic features due only to the experimental manipulations. All subsequently reported analyses were performed on these residual values.

Focus location

The extent to which a discriminant function analysis can separate data points into two or more groups is calculated with a statistical test, Wilks's Lambda.[6]

To determine how well the acoustic features could differentiate focus location in speakers' productions, we computed a model where the 12 acoustic predictors were used to discriminate among three focus locations:

---

[6] Wilks's lambda is a measure of the distance between groups on means of the independent variables, and is computed for each function. It ranges in size from 0 to 1; lower values indicate a larger separation between groups. The extent to which the model can effectively discriminate a new set of data is simulated by a leave-one-out classification, in which the acoustic data from each production are iteratively removed from the dataset, the model is computed, and the left-out case is classified by the resultant functions.

subject, verb, or object. In this analysis, we are averaging across the contrastive and noncontrastive condition for each sentence position.

The overall Wilks's Lambda of the model was significant, $\Lambda = .46$, $\chi^2(24) = 271$, $p < .001$, indicating better-than-chance differentiation of subject focus from verb and object focus. In addition, the residual Wilks's Lambda was significant, $\Lambda = .84$, $\chi^2(24) = 62.65$, $p < .001$, indicating that the acoustic predictors could also differentiate verb focus from object focus (see Figure 2). Leave-one-out classification correctly classified 67% of the productions. The model correctly classified subject focus 76% of the time, verb focus 58% of the time, and object focus 66% of the time. Table 3 presents the standardised canonical discriminant function coefficients of the model.[7]
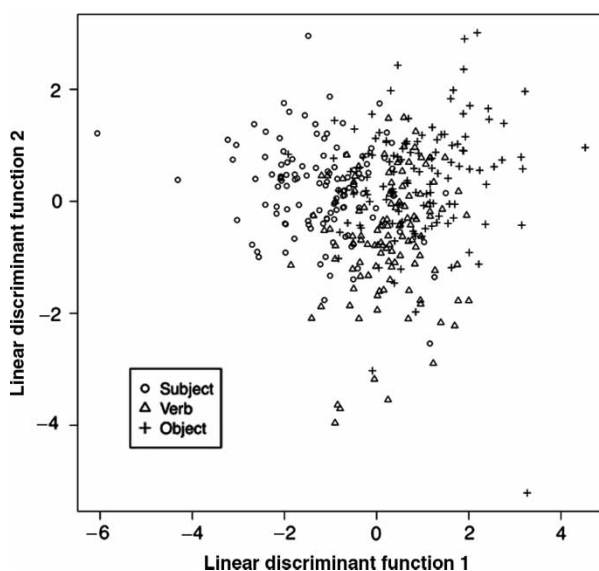

Figure 2. Separation of focus locations on two discriminant functions in Experiment 1. The figure illustrates an effective discrimination among the three groups. Productions of subject focus are clustered in the upper left quadrant; productions of verb focus are clustered in the lower half of the plot; productions of object focus are clustered in the upper right quadrant.

---

[7] The coefficients in Table 3 indicate which acoustic features best discriminate focus location, such that larger absolute values indicate a greater contribution of that feature to discrimination. For example, inspection of the plot in Figure 2 and the coefficients in the focus location columns of Table 3 shows that the acoustic features of *Damon* score around zero, or lower, on the first function (-0.002, 0.001, -0.01, and -0.06) and around zero on the second function (-0.003, 0.021, -0.016, and -0.101). *Fried* shows a different pattern; specifically, the acoustic features of *fried* have coefficients around zero for the first function, and negative coefficients for the second function. Finally, *omelet* shows a third pattern: its acoustic correlates are centered around zero for the first function, but are high for the second function.

TABLE 3
Standardised canonical coefficients of the discriminant functions computed for Experiment 1

|		Focus location		Focus type			
		Function 1	*Function 2*	*Subject focus*	*Verb focus*	*Object focus*	*Focus breadth*
Omelet	Duration+silence	−0.001	0.004	0.008	0.003	0.004	0.003
	Mean F0	−0.006	0.011	0.011	−0.014	−0.019	0.000
	Maximum F0	0.002	0.001	−0.002	0.002	0.006	0.003
	Maximum intensity	−0.037	0.181	−0.137	−0.026	0.189	0.199
Fried	Duration+silence	0.007	−0.001	0.007	0.002	−0.001	0.005
	Mean F0	0.024	−0.003	0.000	−0.040	−0.013	−0.025
	Maximum F0	0.002	−0.002	0.004	−0.007	0.013	0.003
	Maximum intensity	0.094	−0.010	−0.076	0.131	−0.043	0.011
Damon	Duration+silence	−0.002	−0.003	0.005	−0.002	0.005	0.003
	Mean F0	0.001	0.021	−0.016	−0.007	−0.014	0.007
	Maximum F0	−0.010	−0.016	−0.012	0.020	−0.011	−0.005
	Maximum intensity	−0.060	−0.101	0.087	0.056	−0.225	−0.123
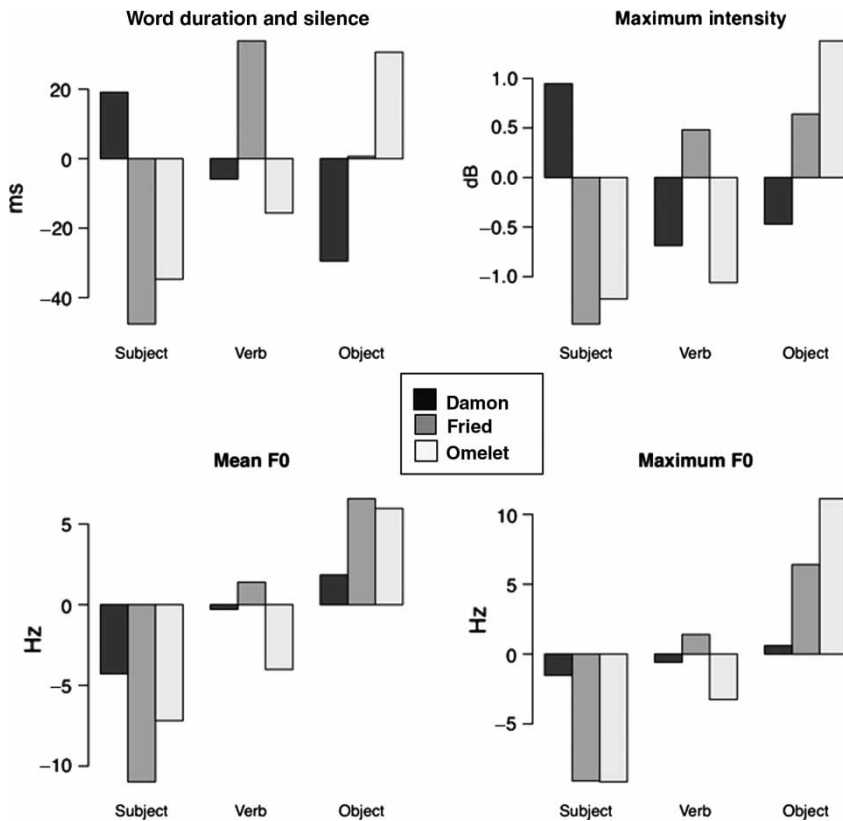
Figure 3. Values of the four discriminating acoustic features of productions of subject, verb, and object focus in Experiment 1.

Figure 3 graphically presents the mean values of the four features, demonstrating that across all three focus locations the intended focus location is produced with the highest maximum intensity, the longest duration and silence, and the highest relative F0.

Focus type

To determine how well the acoustic features could differentiate the type of focus (i.e., noncontrastive vs. contrastive) in speakers' productions, we computed three models in which the 12 acoustic predictors were used to discriminate between two focus groups. The three models investigated differences between noncontrastive and contrastive focus at the three focus locations: subject, verb, and object.

Focus type—subject position. The overall Wilks's Lambda was not significant, $\Lambda = .898$, $\chi^2(12) = 11.95$, $p = .45$, indicating that the acoustic features could not discriminate between noncontrastive and contrastive focus. Because the overall model is not significant, we do not present the scores of the specific acoustic features or the classification statistics here or in the analyses below.

Focus type—verb position. The overall Wilks's Lambda was not significant, $\Lambda = .851$, $\chi^2(12) = 17.92$, $p = .12$, indicating that the acoustic features could not discriminate between noncontrastive and contrastive focus.

Focus type—object position.   The overall Wilks's Lambda was significant, $\Lambda = .82$, $\chi^2(12) = 22.63$, $p < .05$, indicating that the acoustic features could discriminate between noncontrastive and contrastive focus above chance level. Leave-one-out classification correctly classified 59% of the productions. The model correctly classified noncontrastive focus 59% of the time and contrastive focus 59% of the time.

The coefficients in the object focus column of Table 3 indicate that intensity and mean F0 contribute most to classification. Figure 4 graphically presents the mean values of the four features, demonstrating that contrastive focus is produced with a higher maximum intensity, a longer duration and silence, and higher maximum F0. Noncontrastive focus is produced with a higher mean F0.

Focus breadth

To determine how well the acoustic features could differentiate focus breadth, we computed a model in which the 12 acoustic predictors were used to discriminate between productions where the entire event was focused and productions where the object was noncontrastively or contrastively focused.

The overall Wilks's Lambda was significant, $\Lambda = .75$, $\chi^2(12) = 47.83$, $p < .001$, indicating that the acoustic features could successfully discriminate between conditions where the entire event is focused and conditions where the object is narrowly focused. Leave-one-out classification correctly classified 72% of the productions. The model correctly classified wide focus 67% of the time and narrow focus 74% of the time.

The standardised canonical discriminant function coefficients in the focus breadth column of Table 3 indicate that maximum intensity contributes most to focus breadth classification. Figure 5 graphically presents the mean values of the four features, demonstrating that wide focus is produced with a more uniform duration + silence and maximum F0 across the sentence than object
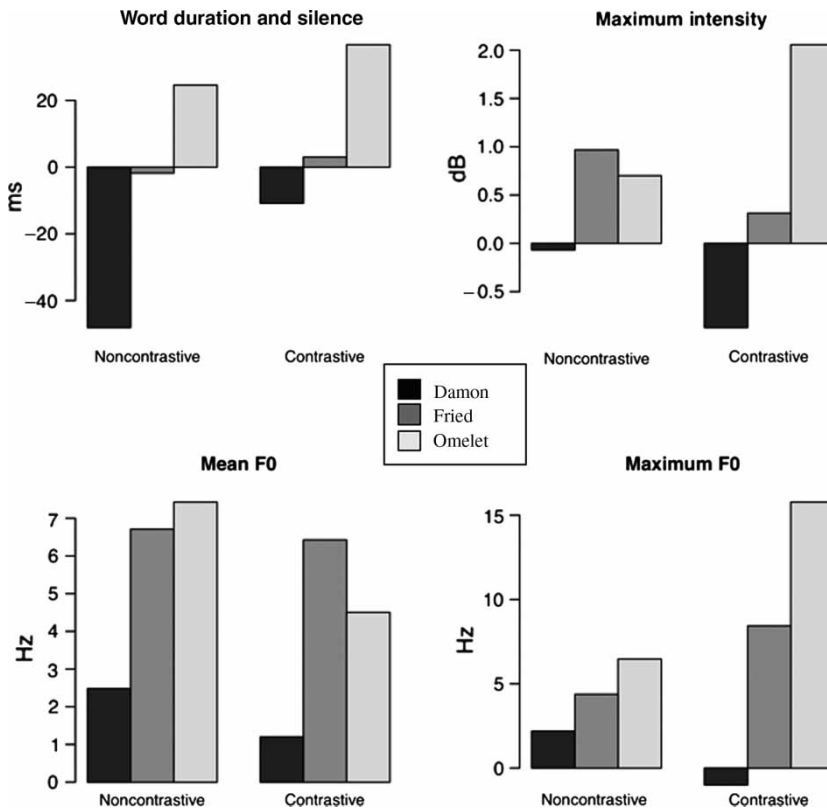
Figure 4. Values of the four discriminating acoustic features of productions of noncontrastive and contrastive object focus when *omelet* is focused in Experiment 1.

focus. Wide focus is also produced with a more uniform, greater intensity than object focus.

Discussion

Focus location

The results demonstrate that speakers consistently provide acoustic cues which disambiguate focus location. Specifically, speakers indicated focus with increased duration, higher intensity, higher mean F0, and higher maximum F0. Furthermore, these results are consistent with the pattern reported in Eady and Cooper (1986), such that the word preceding a focused word is less prominent (produced with shorter duration, lower intensity, and lower F0) than the focused word, and the word following the focused word is less prominent than the word preceding the focused word. Previous studies
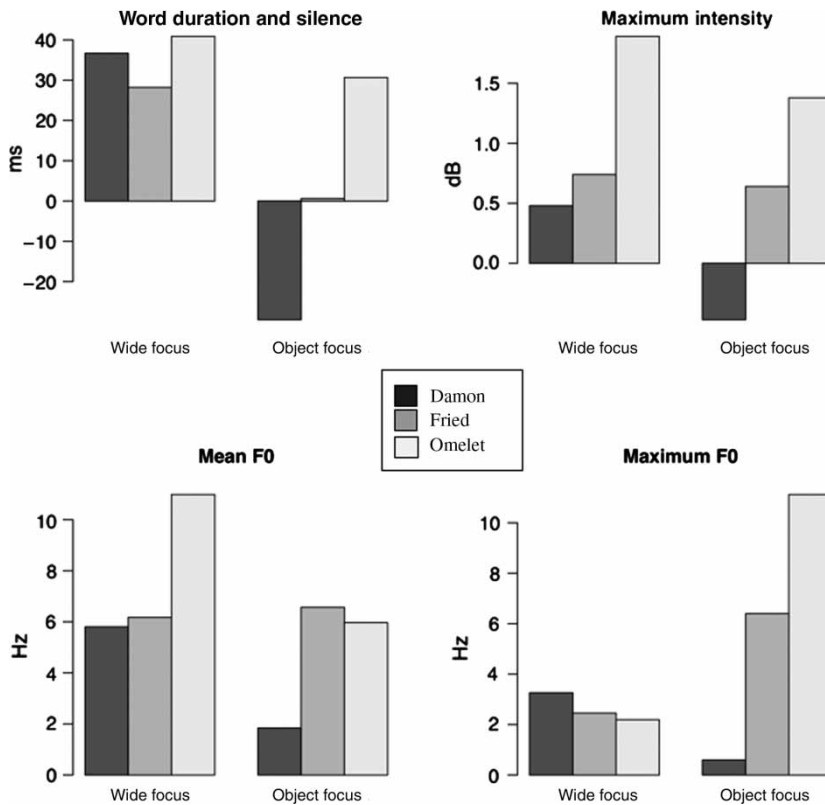
Figure 5. Values of the four discriminating acoustic features of productions of wide and narrow object focus in Experiment 1.

(Eady & Cooper, 1986; Rump & Collier, 1996) have reported this reduction in acoustic prominence following focused elements as being mainly indicated by lower F0 on the postfocal words, though in our data we also find evidence of this reduction in measures of duration and intensity.

Focus type

The results demonstrate that in semi-naturalistic productions speakers do not systematically differentiate between different focus types (focused elements which have explicit contrast sets in the discourse and those which do not). Specifically, at two out of three sentence positions, a discriminant function analysis could not successfully classify speakers' productions of contrastively vs. noncontrastively focused elements. The observation that speakers could not successfully discriminate noncontrastive and contrastive focus in subject or verb positions could be a result of

insufficient experimental power, a limitation of the current experiment which will be addressed in Experiments 2 and 3.

Focus breadth

The results demonstrate that speakers do systematically mark focus breadth prosodically. Narrow object focus is produced with the highest maximum F0, longest duration, and maximum intensity of the object noun, relative to the other words in the sentence. For wide focus, the acoustic features are more similar across the sentence; only intensity and mean F0 are higher on the object than on the other words in the sentence. These differences are subtle, but sufficient for the model to successfully discriminate the productions.

The fact that the model failed to systematically classify productions by focus type (with the exception of the object position), while achieving high accuracy in focus location and focus breadth indicates that speakers were not marking focus type with prosody in Experiment 1. However, the method used to elicit productions did not require that subjects be aware of the information structure ambiguity of the materials. Evidence from other production studies suggests that speakers may not prosodically disambiguate ambiguous productions if they are not aware of the ambiguity. Albritton, McKoon, and Ratcliff (1996), for example, demonstrated that speakers did not disambiguate syntactically ambiguous constructions like "Dave and Pat or Bob" unless they were aware of the ambiguity (see also Snedeker & Trueswell, 2003, but cf. Kraljic & Brennan, 2005; Schafer, Speer, Warren, & White, 2000, for evidence that speakers do disambiguate syntactically ambiguous structures even in the absence of ambiguity awareness). Experiment 2 was designed to be a stronger test of speakers' ability to differentiate focus location, focus type, and focus breadth. We used materials similar to those in Experiment 1, with two important methodological modifications. First, instead of producing the answers to questions with no feedback, the speaker's task now involved trying to enable the answerer to choose the question that s/he was answering from a set of possible questions. Moreover, we introduced feedback so that the speaker would always know whether his/her partner had chosen the correct answer. Second, we changed the design from a between- to a within-subjects manipulation. This ensured that speakers were aware of the manipulation, as they were producing the same answer seven times over the course of the experiment with explicit instructions to differentiate their answers for their partner.

In addition to making the speaker's task explicit, the new design also allowed us to analyse the subset of the productions for which the listeners could successfully identify the question-type and which therefore contain

sufficient information for differentiating utterances along the three relevant dimensions of information structure.

EXPERIMENT 2

Method

Participants

Seventeen pairs of participants were recorded for this experiment. Subjects were MIT students or members of the surrounding community. All reported being native speakers of American English. None had participated in Experiment 1. Participants were paid for their participation.

Materials

The materials had the same structure as those from Experiment 1, though the critical words differed. Specifically, a larger set of names and a wider variety of temporal modifiers were used, and some verbs and objects differed from Experiment 1. Unlike Experiment 1, each subject pair was presented with all seven versions of each of 14 items, according to a full within-subjects within-items design. All materials can be found in Appendix 2.

Procedure

Two participants sat at computers in the same room such that neither could see the other's screen. One participant was the speaker and the other was the listener. Speakers were told that they would be producing answers to questions out loud for their partners (the listeners), and that the listeners would be required to choose which question the speaker was answering from a set of seven choices.

At the beginning of each trial, the speaker was presented with a question on the computer screen to read silently. After pressing a button, the answer to the question appeared below the question, accompanied by a reminder to the speaker that s/he would only be producing the answer aloud, and not the question. Following this, the speaker had one more chance to read the question and answer, and then he/she was instructed to press a key to begin recording (after being told by the listener that he/she is ready), to produce the answer, and then to press another key to stop recording.

The listener sat at another computer, and pressed a key to see the seven questions that s/he would have to choose his/her answer from. When s/he felt familiar with the questions, s/he told the speaker s/he was ready. After the speaker produced a sentence out loud for the listener, the listener chose the

question s/he thought the speaker was answering. If the listener answered incorrectly, his/her computer produced a buzzer sound, like the sound when a contestant makes an incorrect answer on a game show. This cue was included to ensure that speakers knew when their productions did not contain enough information for the listener to choose the correct answer.[8]

Results—production

Two speaker–listener pairs were excluded as the listener did not achieve comprehension accuracy greater than 20%. One further pair was excluded as one member was not a native speaker of American English. Finally, another pair of subjects was excluded because they did not take the task seriously, and produced unnaturally emphatic contrastive accents, often shouting the target word, and laughing while doing so. These exclusions left a total of 13 pairs of participants whose responses were analysed.

Sixty-seven of the 1274 trials (5%) were excluded because (1) the speaker failed to produce the correct words; (2) the speaker was disfluent; or (3) the production was poorly recorded. Analyses were performed on all trials, and on the subset of trials for which the listener correctly identified the question. The results were very similar in the two analyses. For brevity of presentation, we present results from analyses conducted on the correct trials ($n = 660$, 55%). The productions from Experiment 2 were analysed using the acoustic features chosen in the feature-selection procedure described in Experiment 1. All analyses were performed on the residual values of these features, after removing speaker and item variance with the method described in Experiment 1.

Focus location

The overall Wilks's Lambda was significant, $\Lambda = .085$, $\chi^2(24) = 1335$, $p < .001$, indicating that the acoustic features could differentiate subject focus from verb and object focus. In addition, the residual Wilks's Lambda was significant, $\Lambda = .306$, $\chi^2(11) = 641$, $p < .001$, indicating that the acoustic features could also discriminate verb focus from object focus (see Figure 6).

Leave-one-out classification correctly classified 93% of the productions. The model correctly classified subject focus 94% of the time, verb focus 90% of the time, and object focus 95% of the time.

The standardised canonical coefficients in the first two columns of Table 4 indicate that the acoustic features contributing most to the discrimination of focus location are once again mean F0 and maximum intensity, though the

---

[8] In early pilots in which there was no feedback for incorrect responses, we observed that listeners were at chance in choosing the correct question.
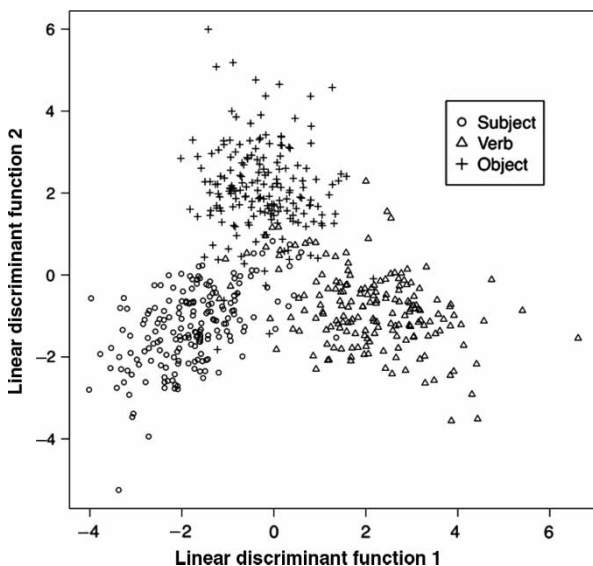
Figure 6.   Separation of focus locations on two discriminant functions for Experiment 2. The figure illustrates an effective discrimination among the three groups. Productions of subject focus are clustered in the lower left quadrant of the plot; productions of verb focus are clustered in the lower right quadrant; productions of object focus are clustered in the lower half.

other two features are also contributing. In fact, inspection of the acoustic feature means in Figure 7 demonstrates that the highest value of every acoustic feature is associated with the intended focused item, with the exception of mean F0 when the subject is focused.

Focus type

Focus type—subject position.   The overall Wilks's Lambda was significant, $\Lambda = .633$, $\chi^2(12) = 81.41$, $p < .001$, indicating that the acoustic features could discriminate between noncontrastive and contrastive focus better than chance. Leave-one-out classification correctly classified 75% of the productions. The model correctly classified noncontrastive focus 78% of the time and contrastive focus 71% of the time.

The standardised canonical discriminant function coefficients in Table 4 indicate that maximum intensity at all three locations (i.e., large intensity differences between the subject and verb and the subject and object) contributes most to classification. Figure 8 graphically presents the mean values of the four features, demonstrating that, in addition to intensity differences, contrastive focus is produced with longer duration and silence, as well as lower mean and maximum F0.

TABLE 4
Standardised canonical coefficients of all discriminant functions computed for Experiment 2

|		Focus location		Focus type			
		Function 1	Function 2	Subject focus	Verb focus	Object focus	Focus breadth
Omelet	Duration+silence	−0.001	0.004	0.004	0.006	0.003	0.003
	Mean F0	−0.006	0.011	−0.003	0.005	−0.023	0.000
	Maximum F0	0.002	0.001	0.004	−0.009	−0.003	0.003
	Maximum intensity	−0.025	0.183	−0.052	−0.171	0.012	0.199
Fried	Duration+silence	0.007	−0.002	0.006	0.002	−0.007	0.005
	Mean F0	0.024	−0.005	0.001	−0.022	0.006	−0.025
	Maximum F0	0.001	−0.002	−0.007	0.001	0.003	0.003
	Maximum intensity	0.093	−0.016	−0.105	0.063	−0.084	0.011
Damon	Duration+silence	−0.002	−0.002	0.002	0.005	0.009	0.003
	Mean F0	0.003	0.021	−0.010	0.004	−0.009	0.007
	Maximum F0	−0.011	−0.015	−0.014	−0.012	−0.006	−0.005
	Maximum intensity	−0.067	−0.097	0.094	−0.014	0.010	−0.123
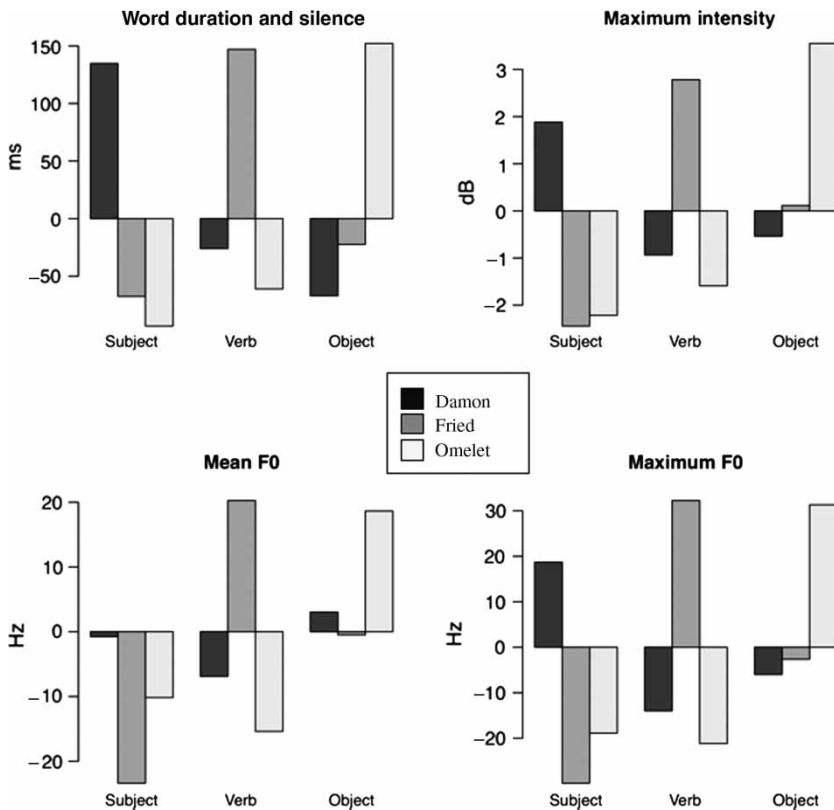
Figure 7. Values of the four discriminating acoustic features of productions of subject, verb, and object focus in Experiment 2.

Focus type—verb position. The overall Wilks's Lambda was significant, $\Lambda = .654$, $\chi^2(12) = 72.27$, $p < .001$, indicating that the acoustic features could discriminate between noncontrastive and contrastive focus better than chance. Leave-one-out classification correctly classified 72% of the productions. The model correctly classified noncontrastive focus 70% of the time and contrastive focus 75% of the time.

The standardised canonical discriminant function coefficients in Table 4 indicate that, once again, maximum intensity contributes most to classification. Figure 9 graphically presents the mean values of the four features, demonstrating that contrastive focus is produced with a higher maximum intensity, and a longer duration and silence, than noncontrastive focus. Once again, noncontrastive focus is produced with higher mean and maximum F0 than contrastive focus.
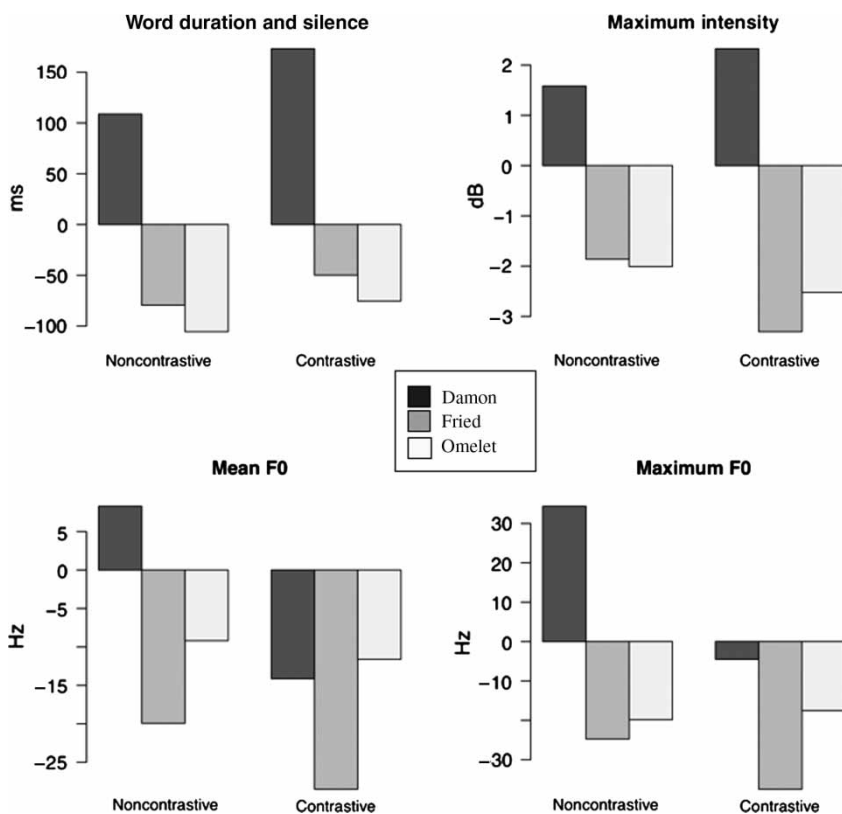
Figure 8. Values of the four discriminating acoustic features of productions of noncontrastive and contrastive object focus when *Damon* is focused in Experiment 2.

Focus type—object position. The overall Wilks's Lambda was significant, $\Lambda = .793$, $\chi^2(12) = 41.3$, $p < .001$, indicating that the acoustic features could discriminate between noncontrastive and contrastive focus better than chance. Leave-one-out classification correctly classified 67% of the productions. The model correctly classified noncontrastive focus 69% of the time and contrastive focus 66% of the time.

The standardised canonical discriminant function coefficients in Table 4 indicate that contrastive focus is most strongly associated with lower mean F0. Figure 10 graphically presents the mean values of the four features, demonstrating that contrastive focus is produced with a lower mean and maximum F0 than noncontrastive focus.
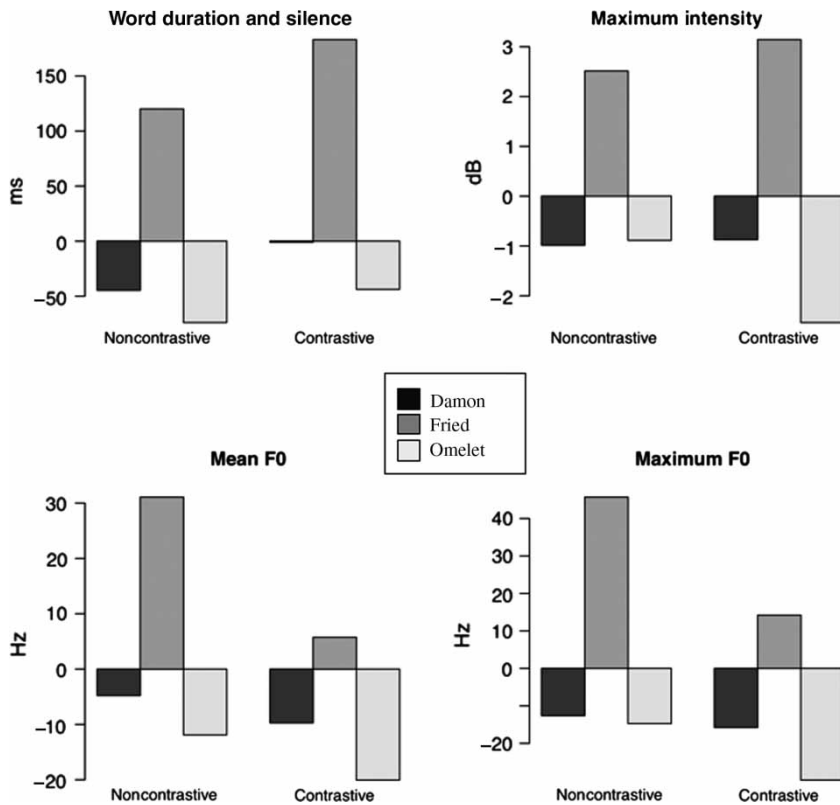
Figure 9.   Values of the four discriminating acoustic features of productions of noncontrastive and contrastive object focus when *fried* is focused in Experiment 2.

Focus breadth

The overall Wilks's Lambda was significant, $\Lambda = .59$, $\chi^2(12) = 148$, $p < .001$, indicating that the acoustic features could differentiate between wide focus and narrow object focus. Leave-one-out classification correctly classified 84% of productions; wide focus was correctly classified 77% of the time, and object focus was correctly classified 88% of the time.

The standard canonical coefficients in the "Focus breadth" column of Table 4 indicate that the maximum intensity of each of the target words contributes most strongly to the discrimination of focus breadth. Although intensity is contributing most strongly to classification, inspection of the acoustic means in Figure 11 indicates that wide focus is marked by lesser prominence on the object, reflected in shorter duration, lower F0, and lower
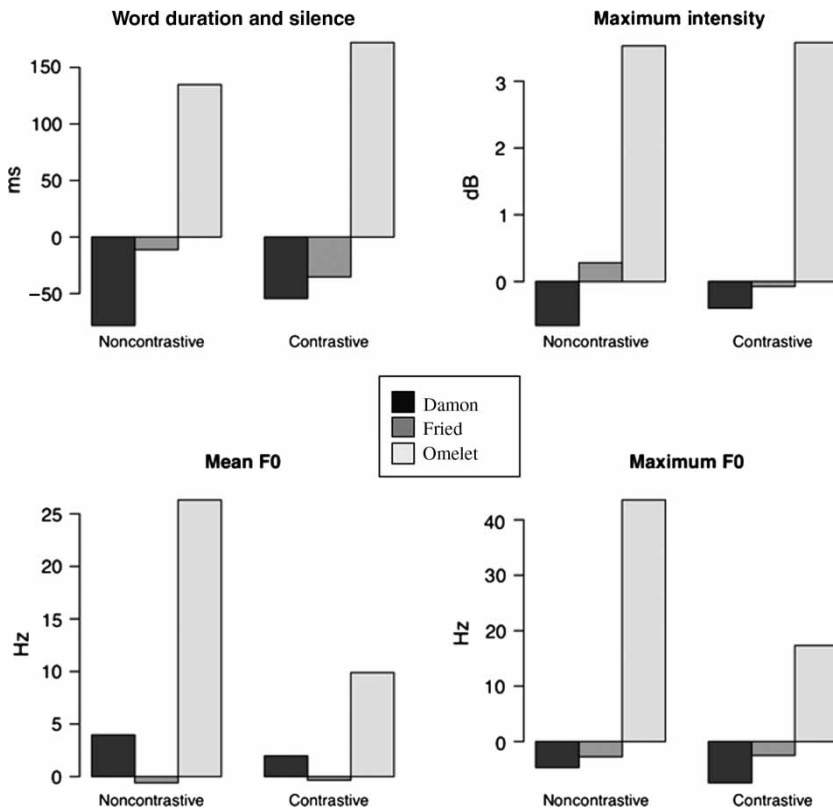
Figure 10. Values of the four discriminating acoustic features of productions of noncontrastive and contrastive object focus when *omelet* is focused in Experiment 2.

intensity; conversely, narrow object focus is marked by greater prominence on the object, reflected in longer duration, higher F0, and higher intensity.

Results—perception

Listeners' choices of question sorted by the intended question are plotted in Figure 12. Listeners' overall accuracy was 55%. To determine whether listeners were able to determine the speaker's intended sentence meaning, we assessed, for focus location and focus type, whether each subject's proportion of correct responses exceeded chance; wide focus productions were excluded from the analysis, so that chance performance for focus location was 0.33, and chance performance for focus type was 0.5. Results demonstrated that listeners were able to successfully identify focus location: all 13 subjects'
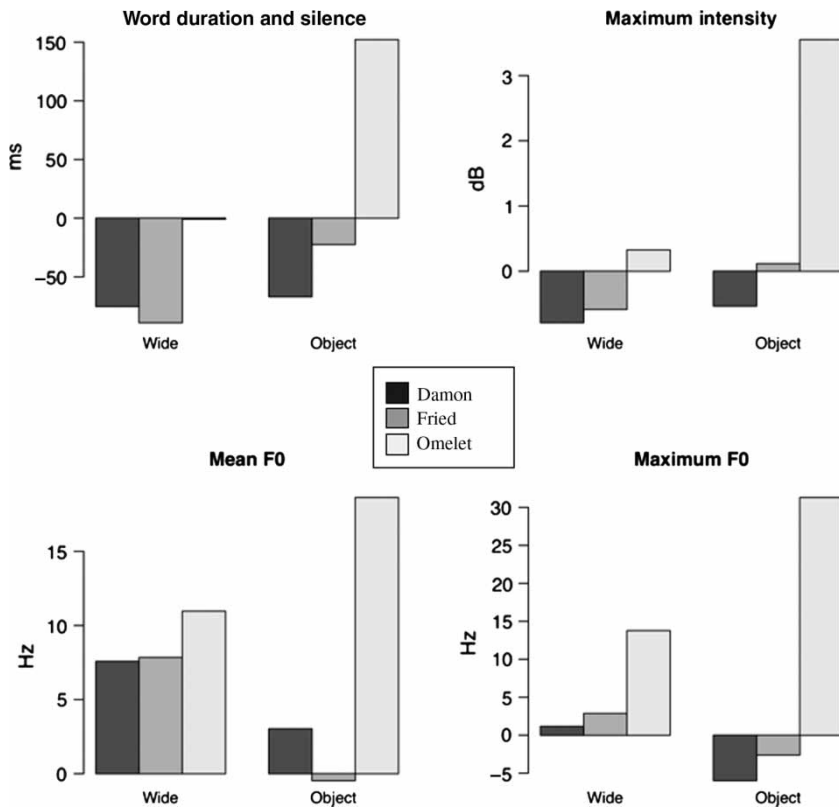
Figure 11.   Values of the four discriminating acoustic features of productions of wide and narrow object focus in Experiment 2.

performance significantly exceeded chance performance, $p = .05$, two-tailed. However, listeners were unable to successfully identify focus type: only three of 13 subjects performed at above-chance levels (based on the binomial distribution), $p = .05$, two-tailed. To investigate focus breadth, we assessed, for wide focus and narrow object focus separately, whether each subject's proportion of correct responses exceeded chance. For these analyses, we excluded subject and verb focus productions, so that chance performance was 0.33 for wide focus, and 0.67 for narrow object focus. Results demonstrated that listeners were moderately successful at identifying focus breadth: six out of 13 subjects identified wide focus at rates above chance, and nine out of 13 subjects identified narrow object focus at levels above chance $p = .05$, two-tailed.
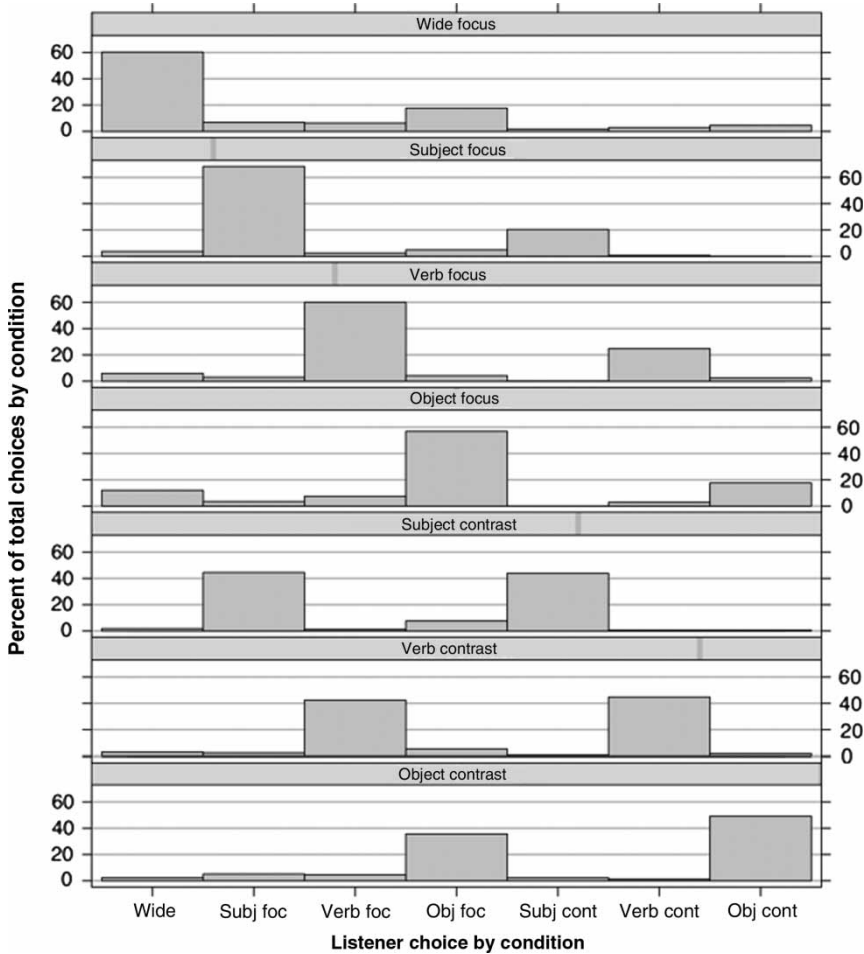
Figure 12.   Percentage of Listeners' condition choice by intended sentence type for Experiment 2.

Discussion

The production results replicated the two main findings from Experiment 1, and provided evidence for acoustic discrimination of focus type across sentence positions as well. First, these results demonstrated that focused elements have longer durations than given elements, incur larger F0 excursions, are more likely to be followed by silence, and are produced with greater intensity. Second, speakers consistently differentiate between wide and narrow focus by producing the object in the latter case with higher F0, longer duration, and greater intensity. Specifically, although object focus

was indicated by increased duration, higher intensity, and higher F0 on the object than on the subject or the verb, wide focus was indicated by comparatively greater duration, higher intensity, and higher F0 on the subject and the verb, and shorter duration, lower intensity, and lower F0 on the object. These results are consistent with those obtained by Baumann et al. (2006), who demonstrated that narrow focus on an element was indicated with longer duration and a higher F0 peak than wide focus on an event encompassing that element.

Most importantly, although speakers in Experiment 1 did not differentiate conditions with and without an explicit contrast set for the focused element (except for the object position), these conditions were differentiated by speakers in Experiment 2, at every syntactic position. There are two possible interpretations of this difference. First, in Experiment 1, speakers produced only four trials in each of the seven conditions, whereas speakers in Experiment 2 (and Experiment 3, reported below), produced 14 trials in each condition, resulting in greater power in the latter two experiments. The fact that, in Experiment 2, speakers successfully discriminated contrastive and noncontrastive focus in all three positions, suggests that the lack of such an effect in Experiment 1 could be due to a lack of power.

As mentioned above, the difference in the findings between Experiments 1 and 2 is also consistent with results from Albritton et al. (1996) and Snedeker and Trueswell (2003) who demonstrated that speakers do not disambiguate syntactically ambiguous sentences with prosody unless they are aware of the ambiguity. The current results demonstrate a similar effect for acoustic prominence, such that speakers do not differentiate two kinds of acoustically prominent elements (contrastively vs. noncontrastively focused elements) unless they are aware of the information structure ambiguity in the structures they are producing.

The discriminant analyses indicated that contrastively focused words were produced with longer durations and higher intensity than noncontrastively focused words, but that noncontrastively focused words were produced with higher F0 than contrastively focused words.[9] This latter finding is surprising when compared to some previous studies. For example, Ladd and Morton (1997) found that higher F0 and larger F0 range is perceived as more "emphatic" or "contrastive" by listeners. Similarly, Ito and Speer (2008) demonstrated that contrastively focused words were produced with higher F0 than noncontrastive ones. Given the unexpected results, we inspected individual pitch tracks to more closely observe the F0 patterns across the

---

[9] Importantly, the F0 results are not artefacts of the residualisation procedure employed to remove variance from the acoustic features due to speaker and item. The same numerical pattern of F0 values is observed whether residualisation is employed or not, though only the residualised acoustic features successfully discriminate focus type.
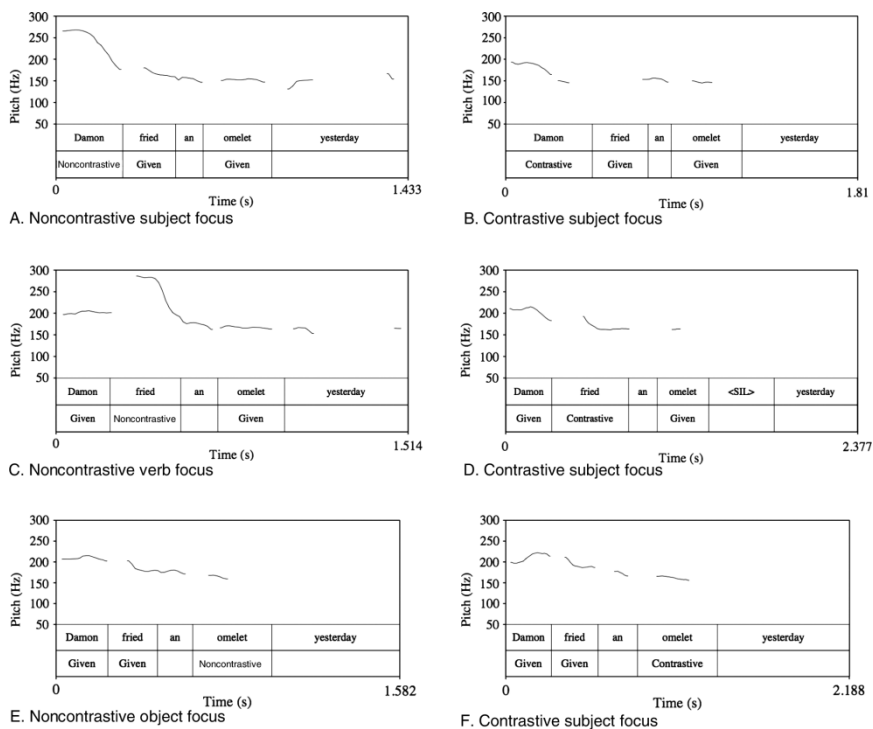
Figure 13. Pitch tracks for noncontrastive and contrastive subject focus, noncontrastive and contrastive verb focus, and noncontrastive and contrastive object focus, respectively, from a typical speaker from Experiment 2.

entire utterances. The pitch tracks presented in Figure 13 were generated from the productions of a typical speaker, and they exemplify the higher F0 observed for noncontrastive focus than contrastive focus in the subject position (A vs. B) and verb position (C vs. D). Contrastive focus on the object is realised with the same F0 as noncontrastive focus on the object (E vs. F).

Note that our finding that noncontrastive focus is realised with higher F0 than contrastive focus is still consistent with the claim that contrastive focus is more prominent than noncontrastive focus. As the graphs in Figures 8–10, and the pitch tracks in Figure 13 indicate, although contrastive elements were consistently produced with lower pitch, they were also consistently produced with longer durations and greater intensity than noncontrastive elements. As reviewed in the introduction, there is evidence that intensity and duration can convey prominence more effectively than higher pitch (Beckman, 1986; Fry, 1955; Kochanski et al., 2005; Lieberman, 1960; Turk &

Sawusch, 1996). Our data are therefore consistent with prior claims that contrastive focus is produced with greater prominence than noncontrastive focus.

As discussed in the introduction, the production elicitation and analysis methods used in the current experiment are more robust than methods used in many previous studies, including those whose results are inconsistent with the current findings. In particular, the current results are based on productions from naïve subjects in a communication task, and the analyses were performed on data with speaker and item variability removed. The current results are therefore more likely to reflect the underlying generalisations about the relationship between acoustics and meaning.

The perception results only partially mirrored the production results. Consistent with the production results, listeners were highly successful in discriminating among the three focus locations. In contrast to the production results, however, listeners were only moderately successful in identifying focus type (noncontrastive vs. contrastive) from the speakers' productions. In fact, listeners most often confused noncontrastive focus with contrastive focus (see Figure 12). These results suggest that, even though speakers may be consistently signalling focus type with their prosody, listeners are not able to exploit those cues for comprehension.

With regard to focus breadth, the perception results are incompatible with a strong version of the focus projection hypothesis (Selkirk, 1995). According to this hypothesis, an acoustic prominence on the object NP can be interpreted as marking the entire clause as focused. Listeners are therefore predicted to treat a production with an acoustically prominent object NP as ambiguous between the narrow object focus reading and the wide focus reading. However, as can be seen in Figure 12, listeners correctly identified narrow object noncontrastive focus 57% of the time, interpreting it as wide focus only 13% of the time, and correctly identified narrow object contrastive focus 49% of the time, interpreting it as wide focus only 6% of the time. These results are not consistent with Gussenhoven's (1983) finding that listeners cannot reliably distinguish between narrow objects focus and wide focus.

Experiments 1 and 2 provide evidence that speakers systematically indicate focus location and focus breadth using a set of four acoustic features. These experiments further suggest that speakers can, but don't always, indicate focus type. In particular, the results suggest that speakers prosodically differentiate contrastive from noncontrastive focus only when they are aware of the meaning ambiguity and/or when the task involves conveying a particular meaning to a listener.

To further investigate the speakers' ability to prosodically differentiate contrastive from noncontrastive focus, we conducted an additional experiment. Acoustic analyses in Experiments 1 and 2 were limited to three

words (i.e., subject, verb, and object) in the sentence. However, in natural productions, speakers' utterances are often prefaced by attribution expressions (e.g., *I think* or *I heard*), or expressions of emotional attitudes towards the described events (e.g., *Unfortunately* or *Luckily*). It is therefore possible that contrastive information might be partially conveyed by prosodically manipulating these kinds of expressions. We explored this possibility in Experiment 3, in which we had speakers produce target SVO constructions with a preamble. Experiment 3 was also intended to serve as a replication of the results of Experiment 2; in particular, the somewhat unexpected finding that noncontrastive focus is produced with higher F0 than contrastive focus.

EXPERIMENT 3

Method

Participants

Fourteen pairs of participants (speakers and listeners) were recorded for this experiment. Subjects were MIT students or members of the surrounding community. All reported being native speakers of American English. None had participated in Experiments 1 or 2. Participants were paid for their participation.

Materials

The materials for Experiment 3 were identical to those from Experiment 1 described above with the exception that an attribution expression (I heard that) was appended to the beginning of each target sentence.

Procedure

The procedure for Experiment 3 was identical to that for Experiment 2.

Results—production

Four speaker–listener pairs were excluded as the listener did not achieve comprehension accuracy greater than 20%. These exclusions left a total of 10 pairs of participants whose responses were analysed. Eighty-one of the 980 recorded trials (8%) were excluded because (1) the speaker failed to produce the correct words; (2) the speaker was disfluent; or (3) the production was poorly recorded. Analyses were performed on all trials, and on the subset of trials for which the listener correctly identified the question the speaker produced the sentence in response to. As in Experiment 2, the results were very similar for the two analyses. For brevity of presentation, we present results from analyses conducted on the correct trials ($n = 632$, 70%).

Focus location

In order to investigate the contribution of the prosody of *I heard that* to the differentiation of the focus type in Experiment 2, we performed a stepwise discriminant function analysis which included as predictors measures of the four acoustic features we had selected initially (duration + silence, mean F0; maximum F0, and maximum intensity) (1) for the subject (Damon), verb (fried), and object (omelet) and (2) for each of the first three words of the sentence (*I*, *heard*, *that*). Of the 24 predictors included in the stepwise discriminant function analysis, the features which resulted in the best discrimination of focus type were (1) the duration + silence of *I*; (2) the maximum F0 of *I*; and (3) the maximum intensity of *I*. Based on these results, we conducted an additional analysis in which we included a subset of 16 predictors: the duration + silence, mean F0, maximum F0, and maximum intensity of the subject, verb, object, and *I*.

As in Experiments 1 and 2, we conducted a discriminant analysis to determine whether the measures of (1) duration + silence; (2) maximum F0; (3) mean F0; and (4) maximum intensity of the four critical words in the sentence could predict focus location.

The overall Wilks's Lambda was significant, $\Lambda = .058$, $\chi^2(32) = 1467.09$, $p < .001$, indicating that the acoustic features could differentiate subject focus from verb and object focus. In addition, the residual Wilks's Lambda was significant, $\Lambda = .275$, $\chi^2(15) = 664.75$, $p < .001$, indicating that the acoustic features could also discriminate verb focus from object focus (Figure 14). Leave-one-out classification procedure correctly classified 97% of the productions. At individual focus locations, the model correctly classified subject focus 96% of the time, verb focus 97% of the time, and object focus 97% of the time (Figure 15).

Focus type

Focus type—subject position. The overall Wilks's Lambda was significant, $\Lambda = .39$, $\chi^2(16) = 157.44$, $p < .001$, indicating that the acoustic features could successfully discriminate between noncontrastive and contrastive focus. Leave-one-out classification correctly classified 85% of the productions. The model correctly classified noncontrastive focus 85% of the time and contrastive focus 85% of the time.

The standardised canonical discriminant function coefficients in Table 5 indicate that maximum intensity overall, and specifically, maximum intensity on "I", is contributing most to classification. Figure 16 graphically presents the mean values of the four features, demonstrating that, in addition to intensity differences, contrastive focus is produced with longer duration and silence, and with lower mean and maximum F0.
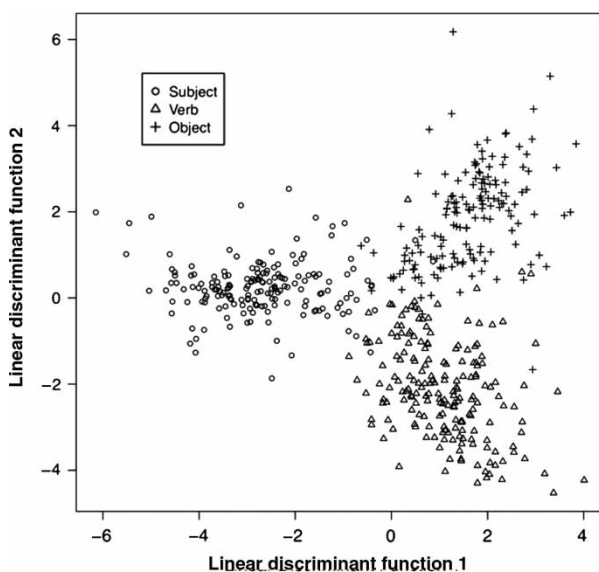
Figure 14. Separation of focus locations on two discriminant functions for Experiment 3. The figure illustrates an effective discrimination among the three groups. Productions of subject focus are clustered in the left half of the plot; productions of verb focus are clustered in the lower right quadrant; productions of object focus are clustered in the upper right quadrant.

Focus type—verb position. The overall Wilks's Lambda was significant, $\Lambda = .46$, $\chi^2(16) = 139.28$, $p < .001$, indicating that the acoustic features could discriminate between noncontrastive and contrastive focus. Leave-one-out classification correctly classified 80% of the productions. The model correctly classified noncontrastive focus 86% of the time and contrastive focus 74% of the time.

The standardised canonical discriminant function coefficients in Table 5 indicate that, intensity on *I* is contributing the most to classification. Figure 17 graphically presents the mean values of the four features, demonstrating that contrastive focus is produced with a higher maximum intensity, and a longer duration and silence, than noncontrastive focus. As in Experiment 2, noncontrastive focus is produced with higher mean and maximum F0 than contrastive focus.

Focus type—object position. The overall Wilks's Lambda was significant, $\Lambda = .40$, $\chi^2(16) = 133.37$, $p < .001$, indicating that the acoustic features could discriminate between noncontrastive and contrastive focus better than chance. Leave-one-out classification correctly classified 83% of the productions. The model correctly classified noncontrastive focus 89% of the time and contrastive focus 76% of the time.
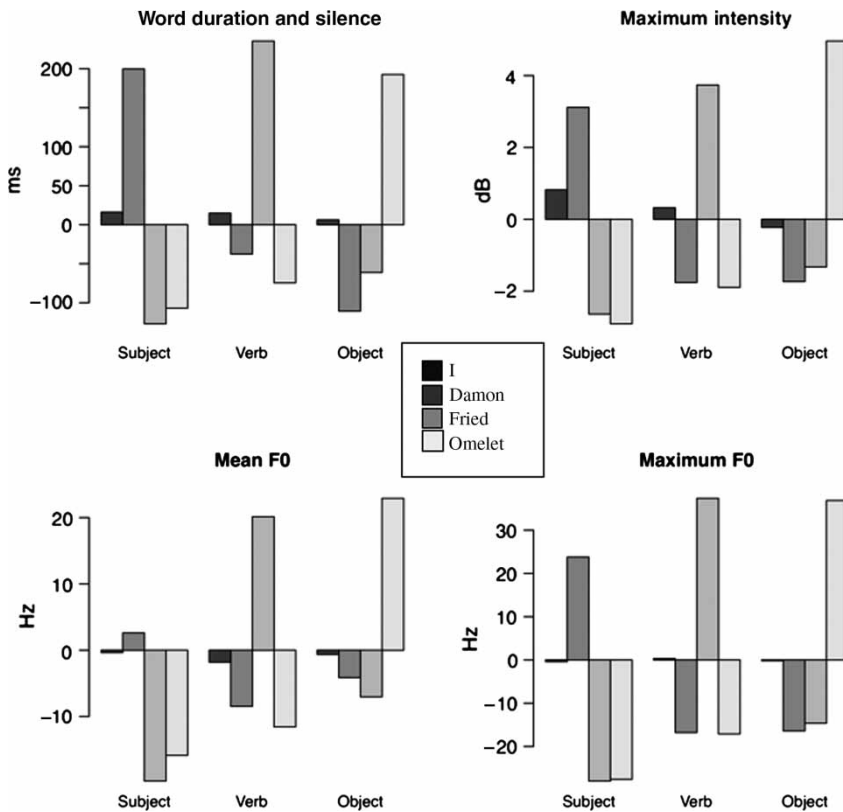
Figure 15. Values of the four discriminating acoustic features of productions of subject, verb, and object focus in Experiment 3.

The standardised canonical discriminant function coefficients in Table 5 indicate that intensity and mean F0 on *I* are contributing the most to accurate classification. Figure 18 graphically presents the mean values of the four features, demonstrating that contrastive focus is produced with a higher mean and maximum F0 than noncontrastive focus.

Focus breadth

The overall Wilks's Lambda was significant, $\Lambda = .48$, $\chi^2(16) = 148$, $p < .001$, indicating that the acoustic features could differentiate between wide focus and narrow object focus. Leave-one-out classification correctly classified 87% of productions; wide focus was correctly classified 79% of the time, and object focus was correctly classified 92% of the time.

TABLE 5
Standardised canonical coefficients of all discriminant functions computed for Experiment 3

		Focus location		Focus type			
		Function 1	Function 2	Subject focus	Verb focus	Object focus	Focus breadth
Omelet	Duration+silence	0.002	0.003	0.000	0.000	0.003	0.002
	Mean F0	0.005	0.012	−0.013	−0.009	0.005	−0.010
	Maximum F0	0.003	0.000	0.005	0.006	−0.003	0.003
	Maximum intensity	0.069	0.106	−0.037	−0.011	0.007	0.151
Fried	Duration+silence	0.001	−0.003	0.000	0.002	0.001	0.005
	Mean F0	0.025	−0.021	−0.001	−0.002	−0.001	−0.006
	Maximum F0	−0.005	−0.002	−0.001	−0.005	0.000	−0.003
	Maximum intensity	0.091	−0.077	−0.086	−0.015	0.027	−0.048
Damon	Duration+silence	−0.003	0.000	0.002	0.000	0.000	0.002
	Mean F0	0.011	0.011	−0.011	−0.020	0.019	−0.003
	Maximum F0	−0.014	−0.003	−0.001	0.007	−0.014	−0.008
	Maximum intensity	−0.147	0.011	0.159	−0.006	−0.064	−0.123
I	Duration+silence	0.000	0.000	0.004	0.005	0.005	−0.001
	Mean F0	−0.005	0.000	−0.013	−0.008	−0.003	−0.009
	Maximum F0	0.004	−0.002	0.017	0.010	0.014	0.005
	Maximum intensity	−0.021	−0.017	0.142	0.133	0.126	0.014

Figure 16. Values of the four discriminating acoustic features of productions of noncontrastive and contrastive object focus when *Damon* is focused in Experiment 3.

The standard canonical coefficients in the "Focus Breadth" column of Table 5 indicate that the maximum intensity of each of the target words contributes most strongly to the discrimination of focus breadth. Specifically, greater intensity on the object is a strong predictor of object focus; less intensity on the subject and the verb is a strong predictor of wide focus. Although intensity is contributing most strongly to classification, inspection of the acoustic means in Figure 19 indicates that wide focus is indicated by lesser prominence on the object, reflected in shorter duration, lower F0, and lower intensity; conversely, narrow object focus is indicated by greater prominence on the object, reflected in longer duration, higher F0, and higher intensity.

Results—perception

Listeners' overall accuracy percentage by condition is plotted in Figure 20. Listeners' overall accuracy was 70%. As described in Experiment 2, we
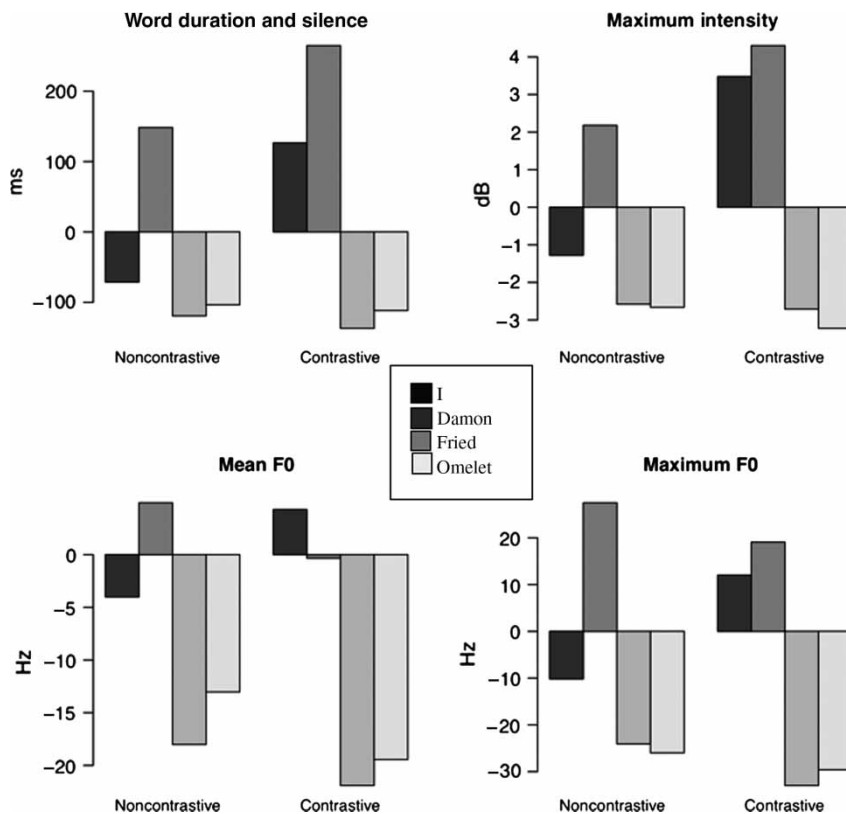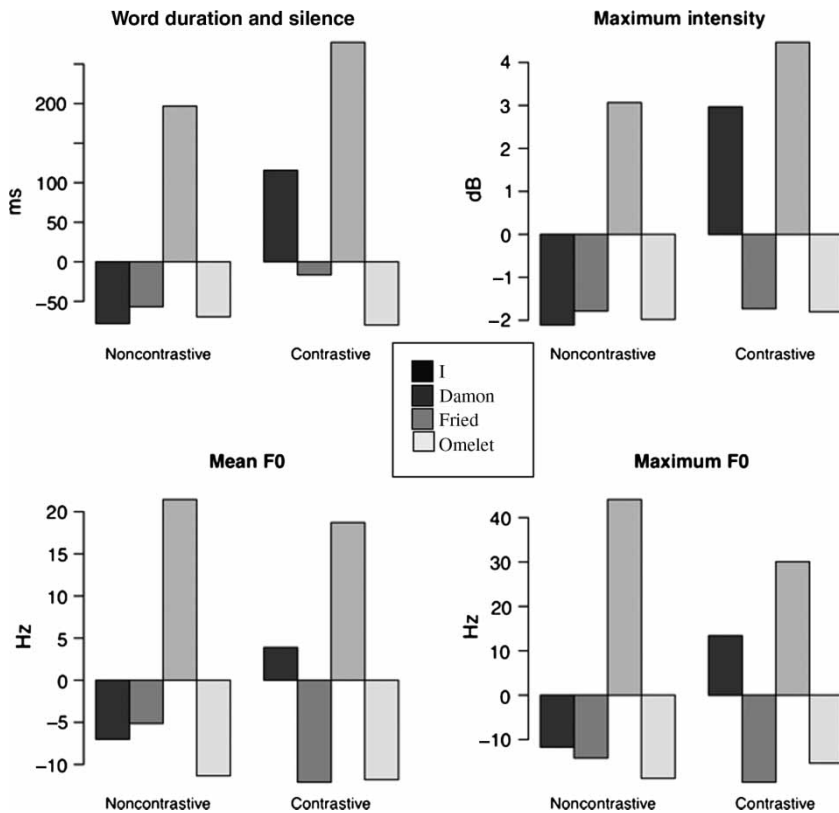
Figure 17.   Values of the four discriminating acoustic features of productions of noncontrastive and contrastive object focus when *fried* is focused in Experiment 3.

compared each subject's responses to chance performance. Results demonstrated that listeners were able to successfully identify focus location, as all 10 subjects' performance significantly exceeded chance performance, $p = .05$, two-tailed. Listeners were moderately successful at discriminating focus type, as six of 10 subjects' performance exceeded chance levels, $p = .05$, two-tailed. Listeners successfully identified focus breadth as eight out of 10 subjects identified wide focus at rates above chance, and eight out of 10 subjects identified narrow object focus at levels above chance $p = .05$, two-tailed.

Discussion

Experiment 3 was conducted in order to (1) investigate whether speakers could differentiate focus type with prosody if the sentences contained an attribution expression that could convey contrastive information, in addition
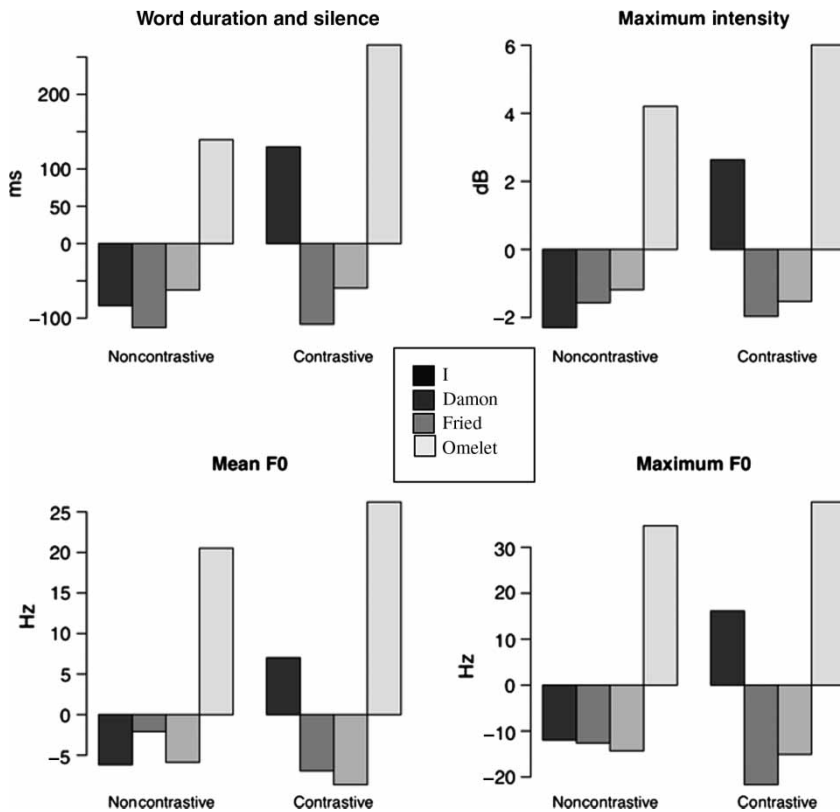
Figure 18. Values of the four discriminating acoustic features of productions of noncontrastive and contrastive object focus when *omelet* is focused in Experiment 3.

to the elements that describe the target event, and (2) replicate the results of Experiment 2.

With regard to the second goal, the production results of Experiment 3 successfully replicated the findings from Experiments 1 and 2. As in Experiments 1 and 2, speakers systematically differentiated focus location and focus breadth with a combination of duration, intensity, and F0 cues. Furthermore, as in Experiment 2, noncontrastive focus was produced with higher F0 than contrastive focus (though only when the subject or verb was focused), and contrastive focus was always produced with greater duration and intensity. As discussed above, these F0 results contrast with prior findings (Bartels & Kingston, 1994; Couper-Kuhlen, 1984; Ladd & Morton, 1997; Ito & Speer, 2008), but can be interpreted in light of more recent evidence that higher intensity is a stronger cue to greater prominence than higher pitch (Kochanski et al., 2005).
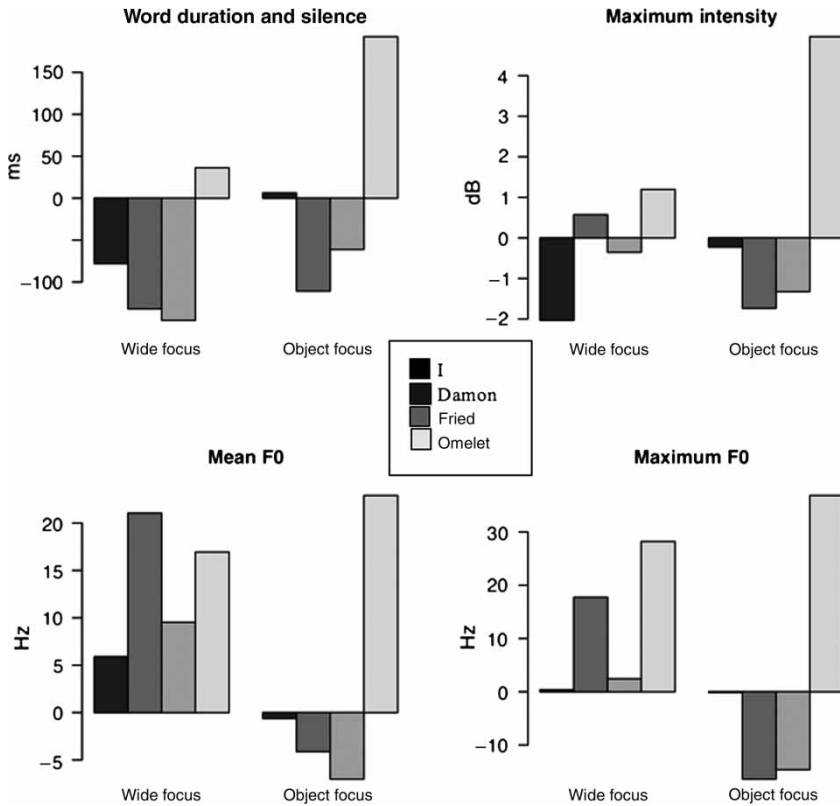
Figure 19. Values of the four discriminating acoustic features of productions of wide and narrow object focus in Experiment 3.

In addition, results from Experiment 3 demonstrated that the strongest cues to discrimination of focus type were the acoustics of *I* (from the attribution expression *I heard that*). Specifically, in contrastive focus conditions, the word *I* was produced with longer duration, higher intensity, and higher mean F0 and maximum F0. Indeed, discrimination of focus type in Experiment 3 was far better than in Experiment 2. It therefore appears that speakers can manipulate prosody on sentence elements outside of the target clause (e.g., in attribution expressions) to convey contrastiveness.

The perception results demonstrated that listeners could accurately determine focus location, similar to the results of Experiment 2. Furthermore, listeners were more accurate in determining focus type than listeners in Experiment 2. This increase in accuracy was likely due to speakers' tendency to prosodically mark *I* in the contrastive conditions.
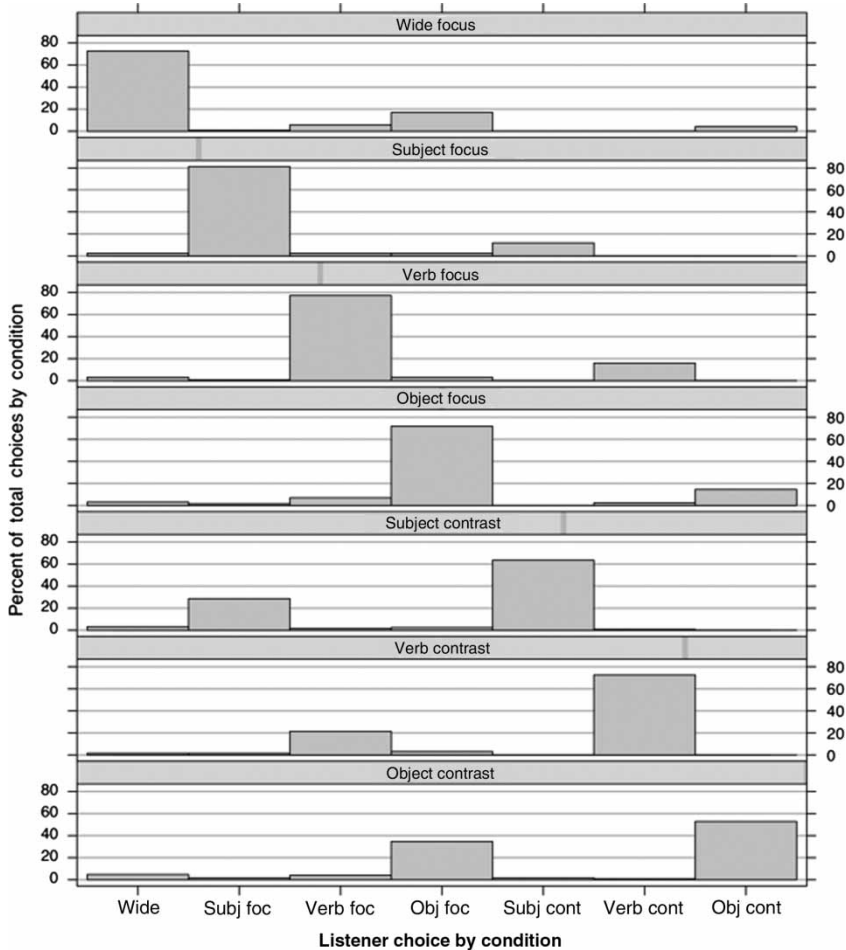
Figure 20. Percentage of listeners' condition choice by intended sentence type for Experiment 3.

GENERAL DISCUSSION

The three experiments reported in the current paper explored the ways in which focus location, focus type, and focus breadth are conveyed with prosody. In each experiment, naïve speakers and listeners engaged in tasks in which the information status of sentence elements in SVO sentences was manipulated via preceding questions. The prosody of the target sentences was analysed using a series of classification models to select a subset from the set of acoustic features that would best be able to discriminate among focus locations and between focus types. In addition, in Experiments 2 and 3,

the production results were complemented by the perception results that demonstrated listeners' ability to use the prosodic cues in the speakers' utterances to arrive at the intended meaning.

At the beginning of the paper, we posed three questions about the relationship between acoustics and information structure: (1) do speakers mark information structure prosodically, and, to the extent they do; (2) what are the acoustic features associated with different aspects of information structure; and (3) how well can listeners retrieve this information from the signal? We are now in a position to answer these questions.

First, we have demonstrated that speakers systematically provide prosodic cues to the location of focused material. Across all three experiments, speakers provided cues to focus location whether or not the task explicitly demanded it, across subject, verb and object positions. In addition, across all three experiments, speakers systematically provided cues to focus breadth, such that wide focus was prosodically differentiated from narrow object focus. Finally, we found that speakers can, but don't always, prosodically differentiate contrastive and noncontrastive focus. Specifically, speakers did not prosodically differentiate focus type in Experiment 1, but they did so in Experiment 2 and, even more strongly, in Experiment 3. As discussed above, the fact that speakers did not differentiate focus type in Experiment 1, where they were plausibly not aware of the meaning ambiguity, but did differentiate between contrastive and noncontrastive focus in Experiments 2 and 3, where the task made the meanings more salient, is consistent with results from the literature on intonational boundary production demonstrating that speakers only produce disambiguating boundaries when they are aware of the syntactic ambiguity (Albritton et al., 1996; Snedeker & Trueswell, 2003; cf. Kraljic & Brennan, 2005; Schafer et al., 2000). Furthermore, the results from Experiment 3, where the critical words were preceded by the attribution expression *I heard that*, demonstrated even stronger differentiation of focus type than in Experiment 2, suggesting that speakers are able to convey contrastiveness using words outside of the clause containing the contrastively focused element.

To answer the question of which acoustic features are associated with different meaning categories of information structure, we conducted a series of discriminant function analyses with the goal of objectively identifying which of 24 measures of duration, intensity, and F0 allowed for the best discrimination of conditions. Across all experiments, and across different sentence positions, the best differentiation among conditions was achieved using the following four features: word duration, maximum word intensity, mean F0, and maximum F0. These results are consistent with many previous studies in the literature, implicating these features in conveying aspects of information structure. An important contribution of the current studies is

that these results were obtained using a quantitative analysis across many naïve speakers and items, and are therefore more likely to be generalisable.

These data also demonstrate how exactly these four features are used in conveying different aspects of information structure. With regard to focus location, focused material is produced with longer duration, higher F0, and greater intensity than given material. With regard to focus type, noncontrastive focus is realised with higher mean and maximum F0 on the focused word than contrastive focus, whereas contrastive focus is realised with greater intensity on the focused word than noncontrastive focus. Finally, with regard to focus breadth, narrow focus on the object is indicated by higher F0 and longer duration on the object, compared to wide focus, and wide focus is conveyed by higher intensity and F0, and longer duration on prefocal words.

To answer the question of how well listeners can retrieve prosodic information from the signal, we included a perception task in Experiments 2 and 3. When the relevant acoustic cues were present in the input (as demonstrated by successful classification by the models), listeners were also able to classify the utterances, although not quite as successfully as the models. Furthermore, the fact that the model always achieved high classification accuracy suggests that the utterances contained enough acoustic information to make these discriminations, and that we did not leave any particularly informative acoustic features out of the analyses.

Implications for theories of the mapping of acoustics to meaning

While our production and perception results are compatible with a direct relationship between acoustics and meaning, they are also consistent with the existence of mediating phonological categories, as in the intonational phonology framework. For example, a standard assumption within intonational phonology is that there is a phonological category "accent" mediating acoustics and semantic focus, such that a focused element is accented, and an unfocused element is unaccented (e.g., Brown, 1983). Our production and perception results are compatible with this assumption. First, if speakers are signalling focus location by means of placing acoustic features corresponding to a + accent category on focused elements, then we would expect to see strong acoustic differences between focused and given elements, as we have observed. Moreover, if listeners perceive accents categorically, then we would predict successful discrimination of productions on the basis of focus location, as we have observed. Second, when the object is focused, it will be accented, resulting in higher acoustic measures on the object compared to other positions, as we have observed. Furthermore, in the wide focus

condition, the subject, verb, and object—all of which are focused—would all receive accents, and would therefore be more acoustically similar to one another than they are in the wide focus condition. This difference in accent placement would lead to successful discrimination between wide and narrow focus by listeners, as we have observed. Finally, there has been much debate in the intonational phonology literature about whether there is a phonological category $+/-$ contrastive. The results of our experiments are perhaps best explained without such a category. In particular, if speakers accent focused elements without differentiating between contrastive and noncontrastive focus, then we would expect similar acoustic results between productions which differ only on focus type, which would lead to poor discrimination by the model. Moreover, listeners would not be successful in discriminating focus type, as we have observed. Our experimental results are thus compatible with an intonational phonological approach which includes an accent category mediating acoustics and meaning, but no category for contrastiveness. Importantly, although our results do not support a categorical difference between noncontrastive and contrastive focus, they do not exclude the possibility that speakers can mark these distinctions with relative differences in prominence (Calhoun, 2006).

Implications for semantic theories of information structure

The current results are relevant to two open questions in the semantics of information structure: (1) whether contrastive and noncontrastive focus constitute two distinct categories and (2) whether focus on the object of a verb can project to the entire verb phrase.

As described in the introduction, Rooth (1992) proposed an account of focus which makes no distinction between noncontrastive focus and contrastive focus. (6) shows the F-marking (focus-marking) that Rooth's account would assign to the conditions in Experiments 1 and 2. Importantly, words and phrases which evoke alternatives, either explicit or implicit, are considered focused (i.e., F-marked).

(6) a. Subject, subject contrast: *Damon*$_F$ fried an omelet last night.
     b. Verb, verb contrast: Damon *fried*$_F$ an omelet last night.
     c. Object, object contrast: Damon fried *an omelet*$_F$ last night.
     d. Wide: [Damon *fried an omelet*]$_F$ last night.

Our results provide tentative support for Rooth's proposal that F-marked constituents do not differ substantively as a function of whether the alternatives they evoke are explicit (our contrastive condition) or implicit

(our noncontrastive condition). Although speakers differentiated these two conditions acoustically, they only did so when the contrast between the conditions was made salient (Experiments 2 and 3). Moreover, even when speakers did mark this distinction, listeners were unable to consistently use this information to recover the intended meaning (Experiment 2). These results suggest that there are no consistent semantic differences between foci with explicit alternatives in the discourse and those with implicit alternatives.

The second semantic issue that these results bear upon is whether narrow focus on the object can project to the entire verb phrase. According to the theory of focus projection proposed in Selkirk (1984, 1995), an acoustic prominence on the direct object (*omelet*) can project focus to the entire verb phrase (*fried an omelet*) and then up to the entire clause/sentence. Gussenhoven (1983, 1999) makes a similar claim. Both Selkirk's and Gussenhoven's accounts therefore predict that a verb phrase with a prominence on the object would be ambiguous between a narrow object focus interpretation and a wide focus interpretation. Neither the production nor the perception results were consistent with this prediction. In production, speakers distinguished between narrow object focus and wide focus, and in perception, listeners were able to distinguish these two conditions.

One aspect of the production results (the acoustic realisation of the subject) for the narrow object focus and wide focus conditions is, however, predicted by both Selkirk and Gussenhoven's accounts. In particular, in the wide focus condition, the subject constitutes new information while in the narrow object focus condition the subject is given. Selkirk and Gussenhoven both predict that the subject would be more acoustically prominent in the wide focus condition than in the narrow object focus condition. This is exactly what we observed (especially in Experiments 1 and 3). Nevertheless, as discussed above, speakers also systematically disambiguated wide focus from narrow object focus across all three experiments with their realisation of the object and the verb. Specifically, wide focus was produced with stable or increasing duration, intensity, and F0 across the subject, verb, and object; narrow object focus, on the other hand, was characterised by shorter duration and lower intensity and F0 on the subject and verb, followed by a steep increase in each of these values on the object.

Similar to our production findings, Gussenhoven (1983) found that, at least in some productions, wide focus differed from narrow object focus in that the verb was more prominent under wide focus. Listeners, however, were unable to use this acoustic information to distinguish wide focus from narrow object focus. Gussenhoven took this result as evidence that the two conditions are not reliably distinguished (consistent with his theory). Our results did not replicate this production/perception asymmetry: listeners are

able to successfully classify productions with a single prominence on *omelet* as indicating narrow object focus and did not confuse these productions with those from the wide focus condition.

Methodological contributions

A further contribution of the current research to investigations of prosody and information structure is methodological. With regard to the methods used to elicit productions, we used multiple, untrained speakers to ensure that our results are generaliseable to all speakers and are not due to speakers' prior beliefs about what pattern of acoustic prominence signals a particular meaning (see Gibson & Fedorenko, 2010, for similar arguments with respect to linguistic judgements). Furthermore, unlike most previous work in which productions were selected for analysis based on perceptual differentiability or on ratings of the appropriateness of prosodic contours, we elicited and selected for analysis productions using a meaning task. Thus our analyses were based on the communicative function of language. Finally, we did not exclude speakers based on our perceptions of their productions; speakers were excluded for failure to provide information to their listeners.

The analyses used here also constitute an improvement over previous analyses. First, using discriminant modelling, we were able to simultaneously investigate the contribution of multiple sentence elements to acoustic differentiation of conditions. Second, we demonstrated that residualisation is a useful method for controlling for variability among speakers and lexical items. For example, preliminary analyses performed on the productions from Experiment 2 without first computing residual values of the acoustic features revealed a 13% average increase in values of Wilks' Lambda (where lower values indicate better discrimination) and a 7% average decrease in classification accuracy. Third, the discriminant modelling proved successful in objectively determining which acoustic features were the biggest contributors to differences among conditions. The success of the analyses used in the current studies is encouraging for future investigations of prosodic phenomena previously considered too variable for study in a laboratory setting with naïve speakers.

One question that arises from the current set of studies is, to what extent the current results can be generalised to all speakers and all sentences. In production studies, there is always a trade-off between (1) having enough control over what participants are producing to ensure sufficient data for analysis, and (2) ensuring that the speech is as natural as possible. In Experiment 1, we attempted to elicit natural productions, but failed to find systematic differences between focus types. In making the speakers' task—to

help their listeners choose the correct question-type—explicit, we may have also encouraged speakers to produce these sentences with somewhat exaggerated prosody. Further experiments and corpus investigations will be necessary to determine whether speakers normally produce contrastive meanings in this way.

In conclusion, the current studies used rigorous scientific methods to explore several important questions about the acoustic correlates of information structure. By providing some initial answers to these questions, along with some implications for semantic theory, and by offering a novel, objective way to approach these and other questions, these studies open the door to future investigations of the relationship between acoustics and meaning.

REFERENCES

Albritton, D., McKoon, G., & Ratcliff, R. (1996). Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 714–735.

Bartels, C., & Kingston, J. (1994). Salient pitch cues in the perception of contrastive focus. In P. Boach & R. Van der Sandt (Eds.), *Focus & natural language processing* [IBM Working Papers on Logic and Linguistics, v. 6], (pp. 1–10), Heidelberg, Germany.

Baumann, S., Grice, M., & Steindamm, S. (2006). Prosodic marking of focus domains-categorical or gradient? In *Proceedings of speech prosody* (pp. 301–304). Dresden, Germany.

Beckman, M. (1986). *Stress and non-stress accent* [Netherlands Phonetic Archives Series No. 7]. Dordrecht, The Netherlands: Foris.

Beckman, M., & Ayers Elam, G. (1997). *Guidelines for ToBI labeling, version 3*. Ohio State University.

Birch, S., & Clifton, C. (1995). Focus, accent, and argument structure: Effects on language comprehension. *Language and Speech*, *38*(4), 365–391.

Birner, B. (1994). Information status and word order: An analysis of english inversion. *Language*, *70*(2), 233–259.

Boersma, P., & Weenink, D. (2006). *Praat: Doing phonetics by computer (Version 4.3.10)* [Computer program]. Retrieved from http://www.praat.org/

Bolinger, D. (1961). Contrastive accent and contrastive stress. *Language*, *37*, 83–96.

Breen, M., Dilley, L., Gibson, E., Bolivar, M., & Kraemer, J. (2006, March). *Advances in prosodic annotation: A test of inter-coder reliability for the RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices) transcription systems*. Poster presented at the 19th CUNY Conference on Human Sentence Processing, New York, NY.

Brown, G. (1983). Prosodic structures and the given/new distinction. In D. R. Ladd & A. Cutler (Eds.), *Prosody: Models and measurements* (pp. 67–77). Berlin, Germany: Springer.

Calhoun, S. (2005, January). It's the difference that matters: An argument for contextually-grounded acoustic intonational phonology. In *Linguistics Society of America Annual Meeting*. Oakland, CA.

Calhoun, S. (2006). *Information structure and the prosodic structure of English: A probabilistic relationship* (PhD thesis, University of Edinburgh, UK).

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, *topics and points of view*. In C. N. Li (Ed.), *Subject and topic* (pp. 27–55). New York: Academic Press.

Clark, E. V., & Clark, H. H. (1978). Universals, relativity, and language processing. In J. H. Greenberg (Ed.), *Universals of human language* (Vol. I, pp. 225–277). Stanford, CA: Stanford University Press.

Cooper, W., Eady, S., & Mueller, P. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of Acoustical Society of America*, *77*(6), 2142–2156.

Couper-Kuhlen, E. (1984). A new look at contrastive intonation. In R. Watts & U. Weidman (Eds.), *Modes of interpretation: Essays presented to Ernst Leisi* (pp. 137–158). Tübingen, Germany: Gunter Narr Verlag.

Cutler, A. (1977). The context-independence of "Intonational Meaning". *Chicago Linguistic Society*, (CLS 13), 104–115.

Dilley, L. C. (2005). *The phonetics and phonology of tonal systems*. Unpublished Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, MA.

Eady, S. J., & Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America*, *80*, 402–415.

Féry, C., & Krifka, M. (2008). Information structure: Notional distinctions, ways of expression. In P. van Sterkenburg (Ed.), *Unity and diversity of languages* (pp. 123–136). Amsterdam: John Benjamins.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, *27*, 765–768.

Gibson, E., & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences, 14,* 233–234.

Gussenhoven, C. (1983). Testing the reality of focus domains. *Language and Speech*, *26*, 61–80.

Gussenhoven, C. (1999). On the limits of focus projection in English. In P. Bosch & R. van der Sandt (Eds.), *Focus: Linguistic, cognitive, and computational perspectives* (pp. 43–55). Cambridge, UK: Cambridge University Press.

Gussenhoven, C., Repp, B. H., Rietveld, A., Rump, W. H., & Terken, J. (1997). The perceptual prominence of fundamental frequency peaks. *Journal of the Acoustical Society of America*, *102*, 3009–3022.

Halliday, M. (1967). *Intonation and grammar in British English*. The Hague: Mouton.

Hawkins, S., & Warren, P. (1991). Factors affecting the given-new distinction in speech. In *Proceedings of the 12th International Congress of Phonetic Sciences* (pp. 66–69). Aix en Provence, France.

Ito, K., & Speer, S. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*, 541–573.

Ito, K., Speer, S. R., & Beckman, M. E. (2004). Informational status and pitch accent distribution in spontaneous dialogues in English. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 279–282). Nara, Japan.

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. Cambridge, MA: MIT Press.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

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

Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication*, *34*, 391–405.

Kraljic, T., & Brennan, S. E. (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology*, *50*, 194–231.

Ladd, D. R. (1996). *Intonational phonology*. [Cambridge Studies in Linguistics 79]. Cambridge, UK: Cambridge University Press.

Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: Continuous or categorical? *Journal of Phonetics, 25*, 313–342.

Lambrecht, K. (2001). A framework for the analysis of cleft constructions. *Linguistics, 39*, 463–516.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *The Journal of the Acoustical Society of America, 32*(4), 451–454.

Molnar, V. (2002). Information structure in a cross-linguistic perspective. In H. Hasselgård, S. Johansson, B. Behrens, & C. Fabricius-Hansen (Eds.), *Language and computers* (Vol. 39, pp. 147–161). Amsterdam, The Netherlands: Rodopi.

Paul, H. (1880). *Prinzipien der sprachgeschichte* [Principles of the history of language]. Tübingen, Germany: Max Niemeyer.

Pell, M. D. (2001) Influence of emotion and focus location on prosody in matched statements and questions. *Journal of the Acoustical Society of America, 109*(4), 1668–1680.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation.* Unpublished dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 271–311). Cambridge, MA: MIT Press.

Pierrehumbert, J., & Steele, S. (1989). Categories of tonal alignment in English. *Phonetica, 46*, 181–196.

Pitrelli, J., Beckman, M., & Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 123–126). Yokohama, Japan.

Rietveld, A. C. M., & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics, 13*, 299–308.

Rochemont, M. S. (1986). *Focus in generative grammar.* Amsterdam, Philadelphia: John Benjamins.

Rooth, M. (1985). *Association with focus.* Doctoral thesis, University of Massachusetts Amherst, MA.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics, 1*, 75–116.

Rump, H. H., & Collier, R. (1996). Focus conditions and the prominence of pitch-accented syllables. *Language and Speech, 39*, 1–17.

Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research, 29*, 169–182.

Schwarzchild, R. (1999). GIVENness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics, 7*, 141–177.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure.* Cambridge, MA: MIT Press.

Selkirk, E. (1995). Sentence prosody: Intonation, stress, and phrasing. In J. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 550–569). Oxford: Blackwell.

Shriberg, E., Stolcke, A., Hakkani-Tur, D., & Tur, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication, 32*, 127–154.

Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., ... Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech, 41*(3–4), 439–487.

Silverman, K. E. A., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C. W. S., Price, P., et al. (1992). ToBI: A standard scheme for labeling prosody. In *Proceedings of the 2nd International Conference on Spoken Language Processing* (pp. 867–879). Banff, Canada.

Sluijter, A., & van Heuven, V. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America, 100*, 2471–2485.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential contest. *Journal of Memory and Language, 48*, 103–130.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, *25*, 701–721.

Syrdal, A., & McGory, J. (2000). Inter-transcriber reliability of ToBI prosodic labeling. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 235–238). Beijing, China.

Terken, J. (1991). Fundamental frequency and perceived prominence accented syllables. *Journal of the Acoustical Society of America*, *89*, 1768–1776.

't Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation*. Cambridge, UK: Cambridge University Press.

Turk, A., & Sawusch, J. (1996). The processing of duration and intensity cues to prominence. *Journal of the Acoustical Society of America*, *99*, 3782–3790.

Welby, P. (2003). Effects of pitch accent position, type, and status on focus projection. *Language and Speech*, *46*, 53–81.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, *91*(3), 1707–1717.

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, *33*, 159–197.

Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 2729–2732). Nara, Japan.

APPENDIX 1

Experiment 1 items

Full items are recoverable as follows: Question A is always "What happened last night?" Questions B–D are wh-questions about the subject, verb, and object, respectively. Questions E–G are questions which introduce the explicit alternative subject, verb, or object, indicated in parentheses.

1. Question A: What happened last night?
   Question B: Who fed a bunny last night?
   Question C: What did Damon do to a bunny last night?
   Question D: What did Damon feed last night?
   Question E: Did Jenny feed a bunny last night?
   Question F: Did Damon pet a bunny last night?
   Question G: Did Damon feed a baby last night?
   Response: Damon fed a bunny last night.
2. Damon (Lauren) caught (pet) a bunny (a squirrel) last night.
3. Damon (Molly) burned (break) a candle (a log) last night.
4. Darren (Lauren) cleaned (eat) a carrot (a chicken) last night.
5. Darren (Molly) peeled (eat) a carrot (a potato) last night.
6. Darren (Nora) found (buy) a diamond (a ring) last night.
7. Darren (Jenny) sold (lose) a diamond (a sapphire) last night.
8. Jenny (Damon) found (lose) a dollar (a quarter) last night.
9. Jenny (Darren) sewed (rip) a dolly (a blanket) last night.
10. Jenny (Logan) read (open) an email (a letter) last night.
11. Jenny (Nolan) smelled (plant) a flower (a skunk) last night.
12. Lauren (Darren) burned (write) a letter (a magazine) last night.

13.   Lauren (Logan) mailed (open) a letter (a package) last night.
14.   Lauren (Nolan) read (write) a novel (a newspaper) last night.
15.   Lauren (Damon) fried (bake) an omelet (a chicken) last night.
16.   Logan (Molly) peeled (chop) an onion (an apple) last night.
17.   Logan (Nora) fried (chop) an onion (a potato) last night.
18.   Logan (Jenny) cleaned (buy) a pillow (a rug) last night.
19.   Molly (Logan) dried (wash) a platter (a bowl) last night.
20.   Molly (Nolan) sold (find) a platter (a vase) last night.
21.   Molly (Damon) poured (drink) a smoothie (a cocktail) last night.
22.   Nolan (Nora) pulled (push) a stroller (a sled) last night.
23.   Nolan (Jenny) bought (sell) a stroller (a wheelbarrow) last night.
24.   Nolan (Lauren) sewed (knit) a sweater (a quilt) last night.
25.   Nora (Nolan) killed (trap) a termite (a cockroach) last night.
26.   Nora (Damon) changed (wash) a toddler (a baby) last night.
27.   Nora (Darren) fed (dress) a toddler (a bunny) last night.
28.   Nora (Logan) pulled (push) a wagon (a wheelbarrow) last night.

APPENDIX 2

Items used for Experiments 2–3

Full items are recoverable as follows: Question A always asks "What happened _____?" where the blank corresponds to the temporal adverb. Questions B–D are wh-questions about the subject, verb, and object, respectively. Questions E–G are questions which introduce the explicit alternative subject, verb, or object, indicated in parentheses.

1.   a.   Context: What happened yesterday?
     b.   Context: Who fried an omelet yesterday?
     c.   Context: What did Damon do to an omelet yesterday?
     d.   Context: What did Damon fry yesterday?
     e.   Context: Did Harry fry an omelet yesterday?
     f.   Context: Did Damon bake an omelet yesterday?
     g.   Context: Did Damon fry a chicken yesterday?
          Target: No, Damon fried an omelet yesterday.
2.   (I heard that) (No,) Megan (Jodi) sold (lose) her diamond (her sapphire) yesterday.
3.   (I heard that) (No,) Mother (Daddy) dried (wash) a platter (a bowl) last night.
4.   (I heard that) (No,) Norman (Kelly) read (write) an email (a letter) last night.
5.   (I heard that) (No,) Lauren (Judy) poured (drink) a smoothie (a cocktail) this morning.
6.   (I heard that) (No,) Nora (Jenny) sewed (rip) her dolly (her blanket) this morning.
7.   (I heard that) (No,) Molly (Sarah) trimmed (wax) her eyebrows (her hair) on Tuesday.
8.   (I heard that) (No,) Nolan (Steven) burned (break) a candle (a log) on Tuesday.
9.   (I heard that) (No,) Logan (Billy) killed (trap) a termite (a cockroach) last week.
10. (I heard that) (No,) Radar (Fido) caught (lick) a bunny (a squirrel) last week.
11. (I heard that) (No,) Darren (Maggie) pulled (push) a stroller (a sled) on Sunday.
12. (I heard that) (No,) Brandon (Tommy) peeled (eat) a carrot (a potato) on Sunday.
13. (I heard that) (No,) Maren (Debbie) cleaned (buy) a pillow (a rug) on Friday.
14. (I heard that) (No,) Lindon (Kelly) fooled (fight) a bully (a teacher) on Friday.