The Semantics of Prosody: Acoustic and Perceptual Evidence of Prosodic Correlates to Word Meaning

Lynne C. Nygaard, Debora S. Herold, Laura L. Namya

Department of Psychology, Emory University
Department of Psychology, Indiana University-Purdue University Indianapolis

Received 14 October 2007; received in revised form 3 May 2008; accepted 4 July 2008

Abstract
This investigation examined whether speakers produce reliable prosodic correlates to meaning across semantic domains and whether listeners use these cues to derive word meaning from novel words. Speakers were asked to produce phrases in infant-directed speech in which novel words were used to convey one of two meanings from a set of antonym pairs (e.g., big/small). Acoustic analyses revealed that some acoustic features were correlated with overall valence of the meaning. However, each word meaning also displayed a unique acoustic signature, and semantically related meanings elicited similar acoustic profiles. In two perceptual tests, listeners either attempted to identify the novel words with a matching meaning dimension (picture pair) or with mismatched meaning dimensions. Listeners inferred the meaning of the novel words significantly more often when prosody matched the word meaning choices than when prosody mismatched. These findings suggest that speech contains reliable prosodic markers to word meaning and that listeners use these prosodic cues to differentiate meanings. That prosody is semantic suggests a reconceptualization of traditional distinctions between linguistic and nonlinguistic properties of spoken language.

Keywords: Prosody; Semantics; Acoustic analysis of speech; Word learning; Word meaning; Spoken language processing

1. Introduction

Spoken language is a rich communicative signal that provides information to listeners both through the linguistic content of speech—the syllables, words, or phrases of a talker’s utterance—and through a host of extra-linguistic elements. One particularly important
property that serves a variety of communicative functions is the prosody of speech. Prosody, or the melody of speech, consists of the intonation, rhythm, and relative loudness and timing of components of an utterance and is instantiated primarily in the acoustic correlates of fundamental frequency, amplitude, and relative duration.

Research into prosodic characteristics of speech has focused almost exclusively on two aspects of its contribution to communicative events. The first concerns the processing and disambiguation of the structure of spoken language and the second concerns the role of prosody in indicating the emotional or attitudinal state of the speaker. These two functions of prosodic contours are crucially important for spoken language processing. Prosodic cues such as intonation contour and stress patterns provide cues that help listeners parse the speech stream at syllable, word, and phrase levels (Cutler, 1997; Cutler & Norris, 1988; Shattuck-Hufnagel & Turk, 1996). Even very young infants capitalize on prosodic cues to identify word and phrase boundaries (Christophe, Gout, Peperkamp, & Morgan, 2003; Jusczyk, et al., 1992), and both children and adults use acoustic properties such as syllable duration and intonation contour to constrain their interpretation of syntactic and discourse structure (Beach, Katz, & Skowronski, 1996; Pierrehumbert & Hirschberg, 1990; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Speer, Crowder, & Thomas, 1993). Prosody also conveys information about a communicative partner’s attitude, intentions, and emotion (Fernald, 1989; Ofuka, McKeown, Waterman, & Roach, 2000; Scherer, 1994; Spence & Moore, 2003). Listeners utilize variations in prosody as cues to a speaker’s emphasis and conversational focus (Cohen, Douaire, & Elsabbagh, 2001; Geluykens & Swerts, 1994; Nakajima & Allen, 1993; Schaffer, 1984), and they also base their judgments of a speaker’s emotional or attitudinal state on changes in the acoustic speech signal (Bachorowski, 1999; Scherer, 1994; Scherer, Banse, Wallbott, & Goldbeck, 1991).

The functional significance of prosody has long been assumed to reside either in its support of the formal linguistic structuring of language or, alternatively, in the information it provides about the state or intentions of the speaker. These traditional characterizations assume that prosodic contours carry no direct meaning into the linguistic event. On this view, prosody overlays the formal structural properties (i.e., words and sentences) that have been traditionally assumed to convey meaning. Representational systems governing linguistic structure are divorced from on-line prosodic instantiation. Indeed, this traditional view posits that it is the arbitrary and discrete nature of linguistic structure that affords language its communicative power (de Saussure, 1966[1915]; Hockett, 1960). Thus, formal theories of language processing have assumed that prosody makes no direct contribution to the processing of meaning. Rather, prosody is conceptualized as constituting an additional stream or type of information within the communicative event. The present investigation questions this functional division, assessing the possibility that prosodic structure provides information about the meaning of words.

The assumption that prosody serves a separate function from the propositional content of language has remained largely unchallenged. However, recent studies have suggested that the distinction between prosody and meaning may be less clear than previously believed. A handful of studies have examined the extent to which emotional tone of voice may contribute to and interact with the semantic content of emotion words (Ishii, Reyes, &
Kitayama, 2003; Nygaard & Lunders, 2002; Schirmer & Kotz, 2003; Wurm, Vakoch, Strasser, Calin-Jageman, & Ross, 2001. For example, Nygaard and Lunders (2002) found that emotional prosody influenced listeners’ interpretation of spoken words. Listeners heard emotional homophones such as bridal/bridle or die/dye that had one emotional meaning (positive or negative) and one neutral meaning. Listeners were more likely to report the emotional meaning when the homophones were spoken in a tone of voice congruent with that meaning. This suggests that, at least within the context of emotional tone-of-voice, prosody contributes to the linguistic interpretation of an utterance.

The possibility that prosody may inform word meaning outside the realm of emotional meaning has received even less attention (but see Lehiste, 1973; Wales & Toner, 1979). One exception is a recent study by Shintel, Nusbaum, and Okrent (2006). Shintel et al. found that speakers describing a visual display of a dot moving upward used a higher fundamental frequency (F₀)—or pitch—than when describing a visual display of a dot moving downward. Similarly, speakers described the trajectory (left or right) of a faster-moving dot (e.g., “It is going right” or “It is going left”) using a faster speaking rate than when describing the trajectory of a slow-moving dot. What’s more, in a forced-choice perceptual task, listeners could reliably guess which dot display the speaker was describing (e.g., slow vs. fast) from their utterances alone. Kunihira (1971) also investigated the role of expressive intonation in determining word meaning by asking monolingual native speakers of English to choose the correct meaning for Japanese words. Participants were presented with Japanese antonym pairs (e.g., hot/cold) in three conditions—in a printed form using English spelling conventions, spoken in a monotone voice, and spoken with expressive prosody. Listeners were significantly better at choosing the meaning corresponding to the Japanese words when words were produced with expressive prosody suggesting a nonarbitrary link between prosodic contour and word meaning. These studies suggest that speakers appear to produce prosodic cues to word meaning and that prosody may contribute to word interpretation for listeners. However, the basic theoretical and empirical assumption prevails that prosody is an independent overlay to the linguistic content of speech.

The present investigation examines whether prosody (1) is instantiated via specific acoustics-to-meaning correspondences and (2) can disambiguate the meaning of novel words. Three adult native speakers of English were asked to produce novel words in infant-directed speech (IDS) that were paired with specific dimensional adjectives. IDS is a style of speech characterized by relatively exaggerated intonation and prosodic contours (similar to “happy” speech; Singh, Morgan, & Best, 2002) that is often used in speech to infants and young children (Fernald, 1989; Spence & Moore, 2003). Our rationale for using this speaking style was to provide a naturalistic context in which adult speakers would be engaging in a naming routine with nonwords and to elicit a style of speech likely to exhibit prosodic variation that reliably relates to meaning (see also Fernald, 1989). To that end, speakers were asked to produce utterances such as “Can you get the blicket one?” meaning tall or short, or—in a baseline control condition—with no specific meaning assigned. Acoustic analyses of these utterances were conducted to determine whether speakers produced reliable acoustic cues to meaning. In two perceptual tests, separate groups of listeners viewed pictures depicting each of the two antonyms (e.g., a tall person and a short person)
and judged which of the two pictures matched the meaning of the novel word embedded in the question phrase. In Experiment 1, prosody matched the semantic dimension depicted in the pictures (e.g., prosody produced in conjunction with Tall and Short word meanings paired with *tall*/*short* pictures). In contrast, in Experiment 2, prosody was mismatched with each alternative semantic dimension depicted in the pictures (e.g., prosody produced in conjunction with Tall and Short word meanings paired with *hot*/*cold* pictures). Responses to trials with meaningful prosody were compared to listeners’ responses to baseline utterances produced when speakers had no information about the novel words’ meanings.

We hypothesized that speakers would produce reliable prosodic cues to word meaning under these circumstances and that listeners would be sensitive to these prosodic correlates to meaning. Certainly, the use of IDS and the limited response set for our listeners increased the likelihood that speakers might produce prosodic correlates to meaning and listeners would be able to utilize those cues. However, because of the paucity of evidence suggesting that prosodic cues are associated with a variety of word meanings sampled across semantic/conceptual domains, our goal was to investigate whether, (1) under constrained circumstances, reliable and consistent prosodic cues to meaning are recruited across speakers, (2) prosodic cues are interpreted in similar ways to derive word meaning across listeners, and (3) prosodic cues are acoustically and perceptually specific to particular semantic domains.

2. Acoustic analyses

We first performed acoustic analyses of the speakers’ utterances to determine whether and how prosodic correlates to meaning were instantiated in the acoustic structure of our speakers’ utterances. One possibility is that prosodic correlates that differentiated meanings within or across dimensions would not be produced. If so, we would expect similar acoustic features of the IDS sentences when produced with meaningful versus neutral prosody. A second possibility is that each of the three speakers would prosodically differentiate meaningful from neutral sentences, but do so idiosyncratically with little consistency across speakers or within meanings. A third possibility is that the speakers would exhibit consistent prosodic characteristics that reflect the emotional valence of each meaning. For example, meanings such as Weak or Yucky may be interpreted as more negative than meanings such as Strong or Yummy. If prosody reflects emotional context—either the speaker’s state or the semantic connotation—then the acoustic features of prosody would be expected to vary on a continuum from negative to positive. Finally, meanings may be instantiated with unique acoustic signatures that reflect semantic features both within and outside the domain of emotion. That is, although prosody may reflect positive or negative aspects of meaning, other aspects of semantic content may also be reflected in particular acoustic features that do not co-vary reliably with emotional valence and are consistently recruited across speakers.

To evaluate these possibilities, we conducted acoustic analyses of the prosodic characteristics of novel utterances used to refer to specific meanings (i.e., tall or short) or produced in a neutral baseline condition. Ratings of emotional valence were also collected for
each meaning to determine whether acoustic features of meaning-related prosody would be associated with relative positive or negative semantic tone. We hypothesized that acoustic characteristics would change systematically as a function of word meaning, and that these changes would be stable across speakers.

2.1. Materials

The six-dimensional adjective pairs studied were happy/sad, hot/cold, big/small, yummy/yucky, tall/short, and strong/weak. Twenty-four clip art pictures (two depicting each adjective) were selected both to reinforce novel word meaning during the production task as well as to serve as visual stimuli in the subsequent perceptual experiment.

Three female speakers were recorded producing novel words embedded in the query phrase, “Can you get the (novel word) one?” Novel words were blicket, seebow, daxen, foppick, tillen, and riffel. Speakers were instructed to employ IDS in order to make the novel word learning task more plausible. All three speakers had worked extensively with young children, and one was also the parent of a preschooler. In order to elicit prosodically meaningful utterances using novel words, our speakers were, by necessity, not naïve with respect to the goals of the experiment. Speakers were aware that they would be producing novel words with different meanings and that we were interested in how they produced the novel words. However, speakers were given no explicit guidance regarding what particular types of prosodic cues might be employed to correspond with any particular meaning. Speakers first recorded each of the six novel words embedded in the question phrase using neutral prosody. Speakers were instructed to use IDS but were not told the meaning of the novel words. Speakers then recorded each utterance assuming that the novel word had a particular meaning. For example, the speakers were shown both big pictures and both small pictures and were told to use the novel word (e.g., “seebow”) to mean big. They then recorded the same sentence using the novel word to mean small. The assignment of novel words to each dimension was random, and two different words were recorded for each dimension. This yielded a total of four prosodically meaningful sentences per dimension (e.g., blicket and seebow were both recorded meaning hot and were also both recorded meaning cold) for each of the six dimensions. A total of 24 prosodically meaningful sentences and six prosodically neutral sentences (one for each novel word) were recorded for each speaker.

Utterances were digitally recorded in a sound-attenuated room using a SHURE microphone (SHURE Incorporate, Niles, IL) onto a SONY Digital Audio Tape-corder TCD-D7 (Sony Manufacturing Systems Corp., Kuki-shi, Saitama, Japan). Each stimulus item was re-digitized on a PowerMac G3 (Apple, Inc., Cupertino, CA) at 44.1 kHz sampling rate and edited into separate files for acoustic analysis and subsequent presentation to listeners.

2.2. Procedure

For each utterance described above, fundamental frequency ($F_0$), $F_0$ variation, overall amplitude, and duration were measured using SoundScope, a speech analysis software
package.\(^1\) \(F_0\) indicates the number of cycles per second in a periodic sound. \(F_0\) variation was measured using the standard deviation in \(F_0\) values across the utterance.\(^2\) Overall amplitude reflects amount of energy, and duration is simply the length of the utterance. Given that we would not necessarily expect prosodic correlates to be confined to word boundaries in this task, we performed analyses both on the entire utterance and separately on the novel word edited from the phrase. Because speakers were preparing a naming phrase introducing a novel term, prosodic cues to meaning could easily have been produced to a greater or lesser extent in anticipation of novel word production over the course of the sentence.

2.2.1. Valence rating task

We collected ratings of the degree of positivity and negativity of each of the 12 meanings. Ten participants (two male, eight female) were instructed to “rate each of the following words on how much you associate it with being positive and how much you associate it with being negative.” Using a Likert-type scale with 1 meaning not at all positive/negative and 7 meaning extremely positive/negative, each participant rated each of the 12 word meanings (happy, sad, hot, cold, big, small, strong, weak, tall, short, yummy, yucky), along with eight additional filler words, on separate scales of positivity and negativity.\(^3\)

2.3. Results

To evaluate the extent to which acoustic parameters varied uniquely for each meaning, mean values for each meaning were calculated across speakers for each acoustic dimension. \(t\)-tests were conducted comparing mean values for \(F_0\), \(F_0\) variation, amplitude, and duration for each meaning in the antonym pair (e.g., Tall compared to Short). To control for the distinct phonemic content of individual sentences within each speaker, we conducted analyses by item rather than by speaker repeated for each antonym pair. All reported effects were significant at the \(p < .05\) level unless otherwise noted.

Tables 1 and 2 provide mean values of each acoustic measure as well as a summary of the significant differences for novel words and sentence-length utterances (that included the novel words) respectively as indexed by \(t\)-tests between meanings within each antonym pair, for each acoustic dimension. Overall, analyses conducted at the word level were similar to those for the sentence-length utterances. For both word and sentence analyses, each domain appears to have a unique acoustic profile with the exception of similarities between the Big/Small and Tall/Short dimensions and between the Happy/Sad and Yummy/Yucky dimensions.

To evaluate the extent to which prosodic correlates to meaning were predicted by valence, correlations were conducted between both the positive and negative valence ratings and each of the four acoustic measures: \(F_0\), \(F_0\) variation, mean amplitude, and mean duration for both words and sentences. These analyses revealed that for sentences, \(F_0\) was significantly correlated with both positive valence ratings \((r = .595, p < .05)\) and negative valence ratings \((r = −.613, p < .05)\). This relation was marginal for the word analysis for
both positive \((r = .54, p < .08)\) and negative \((r = -.506, p = .092)\) valence ratings. For words but not sentences, \(F_0\) variation was significantly correlated with both positive (words, \(r = .704, p < .01\); sentences, \(r = .462, p < .13\)) and negative (words, \(r = -.707, p < .01\); sentences, \(r = -.437, p = .156\)) valence ratings. For both words and sentences, amplitude was also significantly correlated with both positive (words, \(r = .708, p < .01\); sentences, \(r = .723, p < .01\)) and negative (words, \(r = -.615, p < .05\); sentences, \(r = -.655, p < .05\)) valence ratings. Duration was not significantly correlated with valence ratings for either words or sentences.

### 3. Experiment 1

Given that there are reliable and consistent prosodic correlates to meaning, we next investigated whether listeners were able to infer novel word meaning from this prosodic information alone.
3.1. Method

3.1.1. Participants

Twenty adults participated in the study and received $5 for participation. Participants reported no history of speech or hearing disorders and all were native speakers of American English.

3.1.2. Procedure

Auditory stimuli were presented over Beyerdynamic DT100 headphones (Beyerdynamic, Inc., Heilbronn, Germany) and stimulus presentation was controlled on a Macintosh computer using PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). On each trial, two pictures depicting the two meanings within a dimension were presented side by side on the screen at the start of each trial (e.g., one picture representing big and one representing small). The auditory stimuli were presented 700 ms after the start of the trial. The participant heard a voice asking, for example, ‘‘Can you get the blickey one?’’ The participant was

Table 2
Acoustic measurements of sentence-length utterances including novel words produced with meaningful and neutral prosody

<table>
<thead>
<tr>
<th></th>
<th>( F_0 ) (Hz)</th>
<th>( F_0 ) SD (Hz)</th>
<th>Amplitude (rms)</th>
<th>Duration (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy/Sad dimension</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↓</td>
</tr>
<tr>
<td>Happy</td>
<td>294.04</td>
<td>76.90</td>
<td>0.81</td>
<td>1556.20</td>
</tr>
<tr>
<td>Sad</td>
<td>221.11</td>
<td>43.79</td>
<td>0.43</td>
<td>1905.46</td>
</tr>
<tr>
<td>Neutral</td>
<td>265.53</td>
<td>60.33</td>
<td>0.58</td>
<td>1600.22</td>
</tr>
<tr>
<td>Hot/Cold dimension</td>
<td>↔</td>
<td>↓</td>
<td>↓</td>
<td>↔</td>
</tr>
<tr>
<td>Hot</td>
<td>234.08</td>
<td>51.47</td>
<td>0.55</td>
<td>1887.57</td>
</tr>
<tr>
<td>Cold</td>
<td>254.51</td>
<td>72.40</td>
<td>0.82</td>
<td>1874.02</td>
</tr>
<tr>
<td>Neutral</td>
<td>271.20</td>
<td>79.56</td>
<td>0.63</td>
<td>1583.77</td>
</tr>
<tr>
<td>Big/Small dimension</td>
<td>↓</td>
<td>↔</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>Big</td>
<td>246.32</td>
<td>69.91</td>
<td>0.81</td>
<td>2224.26</td>
</tr>
<tr>
<td>Small</td>
<td>293.96</td>
<td>73.18</td>
<td>0.52</td>
<td>1597.80</td>
</tr>
<tr>
<td>Neutral</td>
<td>267.67</td>
<td>60.70</td>
<td>0.60</td>
<td>1639.40</td>
</tr>
<tr>
<td>Tall/Short dimension</td>
<td>↔</td>
<td>↔</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>Tall</td>
<td>275.20</td>
<td>87.33</td>
<td>0.95</td>
<td>2170.08</td>
</tr>
<tr>
<td>Short</td>
<td>310.13</td>
<td>82.83</td>
<td>0.65</td>
<td>1599.25</td>
</tr>
<tr>
<td>Neutral</td>
<td>275.50</td>
<td>72.07</td>
<td>0.60</td>
<td>1574.58</td>
</tr>
<tr>
<td>Yummy/Yucky dimension</td>
<td>↑</td>
<td>↑</td>
<td>↔</td>
<td>↔</td>
</tr>
<tr>
<td>Yummy</td>
<td>290.96</td>
<td>64.68</td>
<td>0.78</td>
<td>1815.00</td>
</tr>
<tr>
<td>Yucky</td>
<td>228.58</td>
<td>48.74</td>
<td>0.62</td>
<td>1771.47</td>
</tr>
<tr>
<td>Neutral</td>
<td>278.58</td>
<td>75.78</td>
<td>0.58</td>
<td>1511.69</td>
</tr>
<tr>
<td>Strong/Weak dimension</td>
<td>↔</td>
<td>↔</td>
<td>↑</td>
<td>↔</td>
</tr>
<tr>
<td>Strong</td>
<td>266.27</td>
<td>47.56</td>
<td>0.84</td>
<td>1845.00</td>
</tr>
<tr>
<td>Weak</td>
<td>256.81</td>
<td>51.90</td>
<td>0.44</td>
<td>1706.64</td>
</tr>
<tr>
<td>Neutral</td>
<td>280.74</td>
<td>73.66</td>
<td>0.53</td>
<td>1463.32</td>
</tr>
</tbody>
</table>

NB: ↑ = First dimension > Second dimension \( (p < .05) \); ↓ = First dimension < Second dimension \( (p < .05) \); ↔ = No difference; ↑↓ = Marginal difference.
then asked to select one of two pictures on the screen by pressing one of two designated keys on a keyboard, corresponding to the left and right pictures.

Each of the 24 prosodically meaningful utterances per speaker was presented twice, once with each of the two picture pairs for each dimension resulting in 48 meaning trials per speaker. The participants also received 24 control trials per speaker with neutral prosody. Thus, participants completed 72 trials per speaker resulting in 216 total trials. Trials were blocked by speaker. The order of presentation of the 72 trials within speaker block was randomized for each participant. Order of presentation of the three speaker blocks was counterbalanced across participants.

3.2. Results and discussion

The percent of responses on which the participant chose the picture corresponding to one of the dimension endpoints was calculated for each experimental condition. That is, we calculated the percentage of happy, hot, yummy, strong, tall, big responses for each dimension. To assess the overall performance collapsed across dimension, we averaged the number of “positive” responses for each condition. “Positive” responses were defined as responses in which the listener chose the dimensional adjective that was rated as more positive in the valence rating task than its dimensional adjective pair. Fig. 1 depicts percent responses collapsed across semantic dimension for meaning and neutral trials (see the Appendix for performance on each dimension separately). In order to evaluate whether meaningful prosody increased the selection of the picture corresponding to the target meaning, we conducted an overall two-way repeated-measures analysis of variance (ANOVA) collapsed across dimension with Prosody (Positive, Neutral, Negative) and Speaker (Speakers 1, 2, and 3) as factors.

![Fig. 1. Percent novel word identification for each speaker as a function of prosody condition collapsed across semantic dimension.](image-url)
The analysis yielded a main effect of Prosody, \( F(2,38) = 65.28, p < .001, \text{partial } \eta^2 = .775 \), indicating that the proportion of “positive” responses changed as a function of prosody condition. This main effect was mediated by a Speaker by Prosody interaction, \( F(4,76) = 12.18, p < .001, \text{partial } \eta^2 = .91 \), suggesting that the pattern of performance across prosody conditions changed as a function of speaker. Follow-up analyses of simple main effects revealed a significant main effect of prosody for each speaker (Speaker 1, \( F(2,38) = 61.78, p < .001, \text{partial } \eta^2 = .765 \); Speaker 2, \( F(2,38) = 53.03, p < .001, \text{partial } \eta^2 = .736 \); Speaker 3, \( F(2,38) = 34.89, p < .001, \text{partial } \eta^2 = .647 \)). Pairwise means comparisons for each speaker revealed significant differences between neutral and negative and between positive and negative prosody conditions for all three speakers (\( p < .001 \)).

Performance in the positive versus neutral prosody condition was significantly different for Speakers 2 and 3 (\( p < .001 \)), but not for Speaker 1 (\( p < .21 \)). Thus, the interaction appeared to be driven by Speaker 1’s failure to differentiate prosodic characteristics of positive and neutral utterances. Nevertheless, listeners derived different word meanings as a function of prosody condition for each of the three speakers.

These results suggest that listeners derived the meaning of the novel words from prosody alone. Listeners were significantly more likely to choose the picture corresponding to the word meaning when speakers used meaningful as opposed to neutral prosody. In addition, the use of prosody to indicate word meaning was remarkably consistent across both speakers and listeners. Given that the acoustic correlates underlying prosodic variation were found to be specific to particular meanings, listeners’ ability to derive particular word meanings from the suite of acoustic prosodic cues may be specific to a semantic domain. Although the acoustic analyses and perceptual test results are consistent with this interpretation, an alternative explanation is that listeners were simply sensitive to prosodic cues to valence and based their perceptual judgments of novel word meaning on the correspondence between the valence of the prosody and the general valence of the particular word meaning. Indeed, given the correlation between particular acoustic dimensions and ratings of word and picture valence, we would expect listeners to use the valence of prosodic cues, perhaps in addition to cues specific to particular semantic dimensions, to make their judgments of word meaning. However, it is possible that listeners’ systematic responding was entirely due to valence.

4. Experiment 2

This experiment was designed to determine whether listeners’ judgments of word meaning were based on prosodic correlates to word meaning that are unique to particular semantic domains or on the general correspondence between prosodic and word meaning valence. We presented listeners with the same task as in the first experiment with the exception that word meanings were mismatched with picture response choices. For example, *hot/cold* pictures were paired with novel words recorded to mean Happy/Sad. Across listeners, each set of pictures was presented with each mismatching semantic dimension. If listeners were relying on prosodic correlates to valence in order to interpret the novel words, then performance...
with mismatching dimensions should be similar to that with matching dimensions. However, if listeners are relying on prosodic correlates that are concept- or semantic domain-specific, then performance with mismatching dimensions should be poorer than performance with matching dimensions.

4.1. Method

4.1.1. Participants

Twenty-five adults participated in the study and received $10 for participation. Participants reported no history of speech or hearing disorders and all were native speakers of American English.

4.1.2. Procedure

This experiment replicated the procedure from Experiment 1 but instead of presenting picture pairs with words recorded to convey one of the two meanings depicted in that picture pair, we matched picture pairs with words recorded to convey one of the two meanings depicted in the other picture pairs. Thus, picture pairs and word meanings were completely mismatched. Because we had five mismatched dimensions for each picture pair, we presented five groups of listeners (five listeners per group) with mismatched picture-word dimensions. Across groups, each picture set (e.g., hot/cold) was paired with words recorded to convey each of the mismatching meanings (happy, sad, yummy, yucky, strong, weak, tall, short, big, small). For example, one group of five listeners heard the picture pairs depicting Happy/Sad presented on separate trials with words recorded to mean Tall, Short, and with neutral prosody. A second group heard the Happy/Sad pictures paired with words recorded to mean Strong, Weak, and with neutral prosody. Across the groups, the Happy/Sad pictures were paired with each of the five mismatched word meaning dimensions, completing trials for each of the two meanings and the neutral prosody recorded for that meaning dimension. All other aspects of the procedure and design were identical to that of Experiment 1 with the exception that stimulus presentation was controlled on a Dell Optiplex (Dell, Inc., Round Rock, TX) using EPrime software (Psychological Software Tools, Pittsburgh, PA).

4.2. Results and discussion

Fig. 2 depicts percent “positive” choices collapsed across picture pairs for each word meaning condition. As for Experiment 1, performance is collapsed across semantic dimension such that dimension endpoints that had been rated to be relatively more positive were averaged (happy, yummy, strong, big, tall, hot) and endpoints judged to be relatively more negative were averaged (sad, yucky, weak, small, short, cold). To assess the extent to which listeners distinguished among the word meaning conditions despite mismatched dimensions, a two-way repeated-measures ANOVA, conducted with Prosody (positive, neutral, negative) and Speaker (Speakers 1, 2, and 3), revealed main effects of speaker, $F(2,48) = 4.15, p < .05, \text{partial } \eta^2 = .148$, and prosody, $F(2,48) = 19.67, p < .001, \text{partial } \eta^2 = .450$. These main effects were mediated by a significant interaction between prosody and speaker,
Follow-up analyses of simple main effects revealed a significant main effect of prosody for each speaker (Speaker 1, \(F(2,48) = 16.40, p < .001, \text{partial } \eta^2 = .406\); Speaker 2, \(F(2,48) = 28.52, p < .001, \text{partial } \eta^2 = .543\); Speaker 3, \(F(2,48) = 12.68, p < .001, \text{partial } \eta^2 = .346\). Pairwise means comparisons for each speaker revealed significant differences between positive and negative and between positive and neutral prosody conditions for all three speakers \((p < .001)\). Performance in the neutral versus negative prosody condition was not significantly different for any speaker \((p > .1)\). These results suggest that listeners were somewhat sensitive to the relative valence of the prosodic and word meanings, using this cue to infer word meaning in a manner that cut across particular semantic domains. In order to evaluate whether listeners were using valence information exclusively or were also using prosodic information specific to particular semantic domains in addition to valence information, performance in Experiment 2 was compared to that in Experiment 1. Fig. 3 shows percent positive choices collapsed across speaker and dimension for each word meaning condition. On the left are data from Experiment 1 showing performance when picture pair and meaning condition matched. On the right is performance when picture pair and meaning condition mismatched. Although cross-experiment comparisons must be interpreted with some caution, the identical materials and similar subject populations make the statistical comparison less of a concern. To that end, a two-way repeated-measures ANOVA was conducted with Prosody (positive, neutral, negative) as a within subjects factor and Experiment (match vs. mismatch) as a between-subjects factor. The analysis revealed an expected main effect of prosody, \(F(2,86) = 81.407, p < .001, \text{partial } \eta^2 = .654\). However, this main effect was mediated by an interaction between prosody and experiment, \(F(2,86) = 17.062, p < .001, \text{partial } \eta^2 = .284\), indicating that performance across prosody conditions changed significantly depending on whether picture pairs
and word meaning matched or mismatched. Follow-up means comparisons revealed significant differences between match and mismatch conditions for each prosody condition \( (p < .001) \). For the positive prosody condition, percent positive meaning choices declined, moving toward chance, when meaning and prosody did not correspond. Although not necessarily expected, percent positive meaning choices declined for the neutral prosody condition as well, perhaps reflecting a change in the overall response contingencies between the two experiments. However, for the negative prosody condition, percent positive meaning choices increased, again indicating performance moved toward chance, when meaning and prosody did not correspond. These findings suggest that listeners were relying on aspects of prosodic structure that were concept- or semantic domain-specific. If listeners had been relying on prosodic structure that was general to all semantic domains, then performance should have been unaffected by the mismatched picture-word meaning manipulation. That performance differed significantly across tasks strongly suggests that listeners were deriving word meaning from prosodic cues that specifically reflected particular conceptual domains.

5. General discussion

Acoustic analyses revealed that individual word meaning introduced in IDS elicit unique acoustic signatures for individual semantic domains and that these prosodic cues to meaning are consistent across speakers. Speakers’ productions of each meaning varied acoustically in unique ways from both its antonym pair and neutral speech. Although semantically related domains shared similar acoustic profiles, the profiles were not identical. For example, words meaning Big and Tall were produced with greater overall amplitude and longer duration

![Fig. 3. Percent positive identification collapsed across speaker as a function of prosody condition in Experiment 1 (when picture choices matched semantic dimension) and in Experiment 2 (when picture choices mismatched semantic dimension).](image-url)
than their antonyms (Small and Short), but differed in relative fundamental frequency, suggesting that acoustic instantiation of prosodic correlates to meaning reflect subtle differences in conceptual domains. Speakers also appeared to produce prosodic features that reflected the emotional valence of a meaning. Positive and negative valence ratings were correlated with $F_0$ (for sentences), $F_0$ variation (for words), and amplitude (for both words and sentences). Meanings rated as more positive were produced with higher fundamental frequencies, greater variability in fundamental frequency, and higher amplitude. Those rated more negative were produced with lower fundamental frequencies, less $F_0$ variability, and lower amplitude. This finding suggests that some—but not all—of the meaning was recovered via prosodic correlates to emotion. Overall, the acoustic signatures accompanying individual word meanings suggest that speakers convey semantic content prosodically by producing reliable cues to word meaning.

Further, these results reveal that the prosodic information in IDS contributes to novel word interpretation during spoken language comprehension. Adult listeners inferred which of two antonyms was conveyed by a novel word using variations in prosodic contours alone. Remarkably, this ability was evident not only for word meanings related to emotions such as Happy and Sad but also for word meanings related to other semantic domains such as Tall/Short, Big/Small, Hot/Cold, Strong/Weak, and Yummy/Yucky. In addition, this ability to derive word meaning was based on information specific to a particular conceptual or semantic domain. When listeners were given picture choices that did not match the intended word meaning, performance was significantly worse than when picture choices matched. These findings constitute one of the first demonstrations that listeners can use prosody to infer word meaning outside the domain of emotion.

The finding that prosody provides cues to word meaning, at least in IDS, calls into question traditional assumptions about the function of prosody in speech and its relationship to formal linguistic content (de Saussure, 1966[1915]; Vaissière, 2005). Prosody not only provides information about the formal structure of language and a speaker’s intentions and emotions but also contributes to the semantic interpretation of spoken words. In this sense, prosody infiltrates the meaningful aspects of spoken language, serving as an integral part of the communicative signal. That prosody can contribute to, or in the case of these novel words, provide the basis for, linguistic semantic interpretation and understanding challenges the notion of a strict separation between linguistic and non- or extra-linguistic aspects of speech. Prosody does not simply overlay semantic content, but rather contributes to semantic reference directly. The assumption that the fundamental nature of linguistic reference relies on arbitrary, sublexical linguistic units and their hypothesized combinatorial power (Chomsky & Halle, 1968; de Saussure, 1966[1915]; Hockett, 1960) must be expanded to include the continuous phoneme-, word-, and phrase-spanning features of prosody that provide listeners with concomitant cues to meaning. These findings also raise the possibility that prosody may be an important avenue to aid the acquisition process for novel word interpretation in young word learners by providing an additional cue to meaning.

The present findings not only demonstrate that listeners can use prosodic information to infer meaning but also demonstrate the manner by which meaning is conveyed through prosody. Unique acoustic signatures paired with particular semantic domains, similar
acoustic correlates for similar semantic dimensions, and the relationship between valence ratings and certain acoustic dimensions all suggest that prosody varies as a function of semantic features. Of course, the forced-choice task used in this experiment offered a considerably constrained range of possible interpretations. In natural discourse, the range of interpretations would rarely be so stripped down. Likewise, the use of a single type of speaking style (IDS) and the constraints of our procedure for eliciting utterances sample only one of many styles or registers of speech and only one of many contexts common in spoken communication.

Of particular concern is the extent to which our speakers may have used prosody intentionally because of the structure of our task and, consequently, may not use prosodic variation spontaneously. Certainly, our task may have encouraged prosodic disambiguation. However, the constellation of meaning-specific acoustic correlates for any particular meaning would be difficult to intentionally produce and, perhaps more important, the fact that the three speakers converged on the same dimensions of differentiation within each antonym pair that varied systematically from other antonym pairs suggests that the prosodic information conveyed resembled those employed in natural speech. In addition, recent findings by Herold (2006) suggest that unique acoustic signatures for semantic dimensions are also present in spontaneous interactions between mothers and children. Herold recorded mothers using dimensional adjectives with their children in a picture book reading task. Herold’s participants were naive with respect to the goals of the study and were simply asked to read a storybook to their children that included descriptions of the same pictures used in the current task, labeled with their familiar English dimensional adjectives. Herold found that acoustic correlates to meaning were produced spontaneously in her task and were similar to those found in the current study. The current approach adds to Herold’s (2006) and other similar findings (Kunihira, 1971; Shintel et al., 2006) by providing a controlled assessment of the unique contributions of prosody, revealing that prosody does correlate with meaning and that listeners are aware of these prosody-to-meaning correspondences.

Listeners may make use of the relationship between semantic features and prosodic correlates via several routes. First, prosodic contours may be conventionally related to particular meanings within a particular language community. On this view, each linguistic community might use prosody to signal meaning in an idiosyncratic, perhaps arbitrary manner. Listeners within the community may come to recognize and exploit these conventional relations over the course of language development. Second, listeners may make use of iconic or metaphoric relationships between prosody and intended meaning (Armstrong, 1999; Taub, 2001). For example, meanings that include properties that are large may be produced with “large” amounts of sound such as greater amplitude and longer duration. Certainly, an explanation that relies on iconic relationships or physical similarity would be consistent with views suggesting that linguistic reference, and cognition in general, rely on conceptual metaphors (Lakoff & Johnson, 1980; Smith & Sera, 1992). Finally, the characteristics of prosodic correlates to meaning may be the result of perceptual simulation or more generally, embodiment of the contrasting conceptual domains (Barsalou, 1999). Just as gestures or body movements may reflect the content of conceptual domains, so too may the production of prosody reflect the simulation of features of to-be-communicated meanings. This view
suggests that as speakers produce novel words with specific meanings, they simulate or embody the experience of being Big or Cold, for example, and these properties are in turn reflected in the fine acoustic structure of the pronunciation.

Why would prosody contribute to semantic meaning above and beyond purported arbitrary combinatorial reference? One possibility is that prosody augments, disambiguates, or reinforces meaning. For example, evidence suggests that listeners (both adults and children) utilize prosody to derive meaning of novel words (Herold, Namy, Nygaard, & Chicos, 2006; Kunihira, 1971; Shintel et al., 2006; Thorpe & Fernald, 2006). Similarly, when utterances are polysemous and/or have multiple connotative meanings, listeners may use prosodic information to determine the sense intended by the speaker (Capelli, Nakagawa, & Madden, 1990). Within the domain of emotion, Nygaard and Lunders (2002) found that listeners reported meanings of emotional homophones that were congruent with the tone of voice in which they were presented, suggesting that listeners used prosody to resolve lexical ambiguity. The present findings indicate that this type of conceptual, semantic, and lexical disambiguation may occur in semantic domains outside emotion (see also Herold, 2006) suggesting that this function of prosody may be ubiquitous in spoken communication.

The contribution of prosody to linguistic reference calls for a reconceptualization of the role of prosody in communication and suggests that meaning may not be conveyed exclusively through an arbitrary representational system. This conclusion is bolstered by research into signed languages demonstrating that many signs bear a nonarbitrary relationship to their referents (Armstrong, 1999; Taub, 2001). Research into sound symbolism in spoken language reveals that nonarbitrary reference may be common in spoken languages as well (Berlin, 1994; Hinton, Nichols, & Ohala, 1994; Ramachandran & Hubbard, 2001). Taken together, this work suggests that the division between purely linguistic and nonlinguistic properties is difficult to maintain (see Nygaard, 2005). Rather, the key to the computational and communicative power of language may be the continuous, redundant, integrated, and complementary nature of both segmental and suprasegmental properties of speech. The semantic (as well as syntactic and social) contributions of prosody may be an evolutionarily preserved property that is integrally tied to the communicative success of language.

Notes

1. In addition to mean $F_0$, $F_0$ variation, duration, and amplitude, $F_0$ range, shimmer, and jitter were also collected. These acoustic features did not vary reliably across meaning and, as a result, are not reported here.

2. $F_0$ was derived using a peak-picking algorithm. $F_0$ variation was calculated from the standard deviation of frame-by-frame $F_0$ values. Utterances that were longer in duration contributed more values to the statistic than did those that were shorter in duration.

3. Ratings were also collected for the picture stimuli used in Experiments 1 and 2 to ensure that the pictures’ depicted valence corresponded with the 12 meanings. A separate group of 10 participants (two male, eight female) rated each of 24 pictures (two depicting each meaning). Using a Likert-type scale with 1 meaning not at all positive/
negative and 7 meaning extremely positive/negative, each participant rated each of the 24 pictures, along with 16 additional filler pictures, on separate scales of positivity and negativity. Both positive and negative valence ratings for pictures were highly correlated with those for the meanings (positive, $r = .782$, $p < .01$; negative, $r = .780$, $p < .01$), suggesting that the pictures used in the subsequent perceptual tasks depicted meanings that were consistent in valence with the intended meanings of the speakers.

4. In order to further determine the role of meaning-specific prosodic contributions to word interpretation, we reanalyzed the data removing trials on which meaning and prosody with similar acoustic markers were paired. Thus, trials that paired happy/sad and yummy/yucky and tall/short and big/small meanings and prosody were removed. The reanalysis showed that performance moved incrementally in the direction of chance. Percent positive responses decreased for positive prosody (original mismatch, $m = 61.67$; mismatch with trials removed, $m = 60.78$) and increased for negative prosody (original mismatch, $m = 42.22$; mismatch with trials removed, $m = 43.69$). Although trends, these differences are consistent with the argument that when meaning and prosodic correlates did not correspond across semantic domains, listeners had difficulty reliably recovering meaning.

Acknowledgments

 Portions of this research constituted a doctoral dissertation at Emory University by D.S.H. The authors would like to thank Alan Cienki, Philip Wolff, Harold Gouzoules, and two anonymous reviewers for their comments and suggestions.

References


Appendix

Fig. A1. Percent novel word identification for each semantic dimension in Experiment 1 as a function of speaker and prosody condition.