COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 43 (2019) e12799

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12799

Communicative Context Affects Use of Referential Prosody

Christina Y. Tzeng, Laura L. Namy, Lynne C. Nygaard

Department of Psychology, Emory University

Received 4 August 2018; received in revised form 29 July 2019; accepted 2 August 2019

Abstract

The current study assessed the extent to which the use of referential prosody varies with communicative demand. Speaker–listener dyads completed a referential communication task during which speakers attempted to indicate one of two color swatches (one bright, one dark) to listeners. Speakers' bright sentences were reliably higher pitched than dark sentences for ambiguous (e.g., bright red versus dark red) but not unambiguous (e.g., bright red versus dark purple) trials, suggesting that speakers produced meaningful acoustic cues to brightness when the accompanying linguistic content was underspecified (e.g., "Can you get the red one?"). Listening partners reliably chose the correct corresponding swatch for ambiguous trials when lexical information was insufficient to identify the target, suggesting that listeners recruited prosody to resolve lexical ambiguity. Prosody can thus be conceptualized as a type of vocal gesture that can be recruited to resolve referential ambiguity when there is communicative demand to do so.

Keywords: Prosody; Semantics; Referential ambiguity; Vocal gesture; Communicative demand

1. Introduction

Prosody, the timing, rhythm, and intonation of speech, is a rich source of communicative information, providing details about speaker emotional state (e.g., Banse & Scherer, 1996), segmentation (e.g., Cutler & Norris, 1988), syntax (e.g., Snedeker & Trueswell, 2003), and discourse (e.g., Herman, 2000) structure. Traditional characterizations of prosody assume that these cues do not directly convey semantic information about linguistic reference. However, a growing literature suggests not only that speakers use prosody to convey meaning but also that listeners infer referential details (e.g., object size or speed;

Correspondence should be sent to Christina Y. Tzeng, Department of Psychology, Emory University, 36 Eagle Row, Atlanta, GA 30322. E-mail: ctzeng@emory.edu

Laura L. Namy is now at the Society for Research in Child Development. Portions of this work comprised Christina Tzeng's doctoral dissertation.

Perlman, Clark, & Falck, 2015; Shintel, Nusbaum, & Okrent, 2006) from these cues, implying that prosody may augment referential meaning that is conveyed through the accompanying linguistic content.

In addition to conveying information about speaker affective state, prosodic cues are encoded in representations of spoken words and affect lexical processing and speech production (e.g., Hupp & Jungers, 2013; Nygaard, Herold, & Namy, 2009; Perlman et al., 2015; Shintel & Nusbaum, 2008; Tzeng, Duan, Namy, & Nygaard, 2017; Wurm, Vakoch, Strasser, Calin-Jageman, & Ross, 2001). Nygaard et al. (2009), for instance, found that speakers produced reliable prosodic cues to the meaning of novel adjectives (e.g., *daxen*, meaning small). Novel adjectives intended to mean small, for example, were produced with lower amplitude, faster speaking rate, and higher pitch than when the same words were intended to mean big. Although valence accounted for some of the variance in the acoustic features of speakers' utterances, speakers produced unique profiles of pitch, pitch variation, amplitude, and duration for each word meaning that could not be explained by valence alone. Listeners also inferred the meanings of these recorded novel words more accurately when prosody and word meaning matched rather than mismatched, suggesting that language users systematically recruit prosodic cues to convey and infer specific semantic content outside of emotion or valence information. Based on this type of evidence, Shintel et al. (2006) have characterized prosody as an *analog acoustic expression* (Shintel et al., 2006) such that the suprasegmental elements of speech convey information about objects and events that listeners encode along with linguistic input. In the current study, we explore the communicative contexts under which speakers and listeners recruit prosody as an analog acoustic expression, assessing the extent to which such use of prosody varies with communicative demand to resolve referential ambiguity.

The relation between prosody and linguistic reference may be an instantiation of cross-modal associations that affect perceptual processing more generally. Systematic auditory–visual mappings have been found in multiple domains, including between loudness and size (e.g., Smith & Sera, 1992), pitch and brightness (e.g., Marks, 1974; Melara, 1989; Mondloch & Maurer, 2004), pitch and shape (e.g., Marks, 1987), and pitch and visuo-spatial height (e.g., Chiou & Rich, 2012). Evidence suggests that such correspondences are recruited in spoken language to convey visuo-spatial properties of linguistic referents (e.g., Nygaard et al., 2009; Perlman et al., 2015; Shintel et al., 2006; Tzeng et al., 2017). For example, speakers spontaneously modulated their verbal descriptions of vertically moving dots such that descriptions of upward moving dots (e.g., "It is going up.") were higher pitched than those of downward moving dots (Shintel et al., 2006). Such evidence of the use of prosody to convey visual properties of linguistic referents suggests that prosodic cues in spoken language *recruit* cross-modal mappings, potentially extending both the range and efficiency of information conveyed in speech beyond what is afforded by propositional content alone.

The current study assesses the extent to which the use of prosodic cues to reference varies as a function of communicative demand to clarify the accompanying linguistic content. Speakers and listeners routinely modulate their communicative behaviors to resolve potential lexical and syntactic ambiguity in dyadic interactions (Kröger, Kopp, &

Lowit, 2010). To ensure mutual intelligibility, speakers adapt their utterances depending on the listeners' knowledge base (Clark & Krych, 2004) and attentional focus (Brennan, 1995). Often referred to as attempts to achieve common ground (Glucksberg, 1986; Grice, 1989), such adaptations facilitate effective communicative interactions. Evidence suggestive of this possibility can be found in the literature on co-speech gesture. Speakers have been found to gesture more when talking to a listener who can see them than when talking to a listener who cannot (Alibali, Heath, & Myers, 2001). Speakers also gesture at higher rates (Galati & Brennan, 2014; Jacobs & Garnham, 2007) and use larger gestures (Holler & Stevens, 2007) when describing information that is unfamiliar to their listeners or when they are particularly motivated to communicate clearly (Hostetter, Alibali, & Schrager, 2011). Using a story-telling task, Holler and Beattie (2003) found that participants produced more representational gestures, or gestures that resemble the referent in form, accompanying homonyms (e.g., glasses, records) than control words (e.g., food), suggesting that speakers employ gesture as a means by which to resolve ambiguous linguistic content when there is communicative need to do so.

Similar patterns have been found in the use of prosodic contours to highlight new or contrastive information. In visual search tasks (Bögels, Schriefers, Vonk, & Chwilla, 2011; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Weber, Braun, & Crocker, 2006), listeners have demonstrated sensitivity to utterances that are louder in amplitude, longer in duration, and higher in pitch to highlight new information in comparison to information that is already established between speaker and listener. Given these findings, one possibility is that speakers will be more likely to recruit referential prosody when there is increased communicative need to provide referential information that cannot be resolved lexically.

Recent findings provide preliminary support for this possibility. Using a verbal labeling task, Tzeng et al. (2017) found that speakers who were recorded producing novel (e.g., *blicket* for bright red) rather than English (e.g., *red* for bright red) color labels produced utterances that were higher pitched, higher in amplitude, and shorter in duration for labeling brighter versus darker shades of color. Listeners who heard these recorded utterances in a subsequent perception task reliably inferred the intended target color referent from the speakers' novel label utterances (e.g., Can you get the *blicket* one?), with listener accuracy increasing as a function of the speakers' prosodic cues to reference. That the speakers modulated their prosody as a function of referent brightness when using novel but not English labels suggests that the extent to which referential prosody is recruited to disambiguate meaning varies with communicative context.

In this experiment, we investigate directly whether communicative demand modulates the recruitment of referential prosody to resolve linguistic ambiguity in a dyadic communication context. The Tzeng et al. (2017) findings suggest that the likelihood of speakers using referential prosody is greater when the target label is underspecified with respect to its intended referent (e.g., *blicket* for dark red). Here, we systematically manipulated the level of ambiguity between the label and target referent such that for half of the trials, lexical content was insufficient to identify the target. During these trials, speakers might be likely to recruit other means by which to convey necessary disambiguating

information. For the other half of the trials, lexical content *did* provide sufficient disambiguating information and may lessen the likelihood that referential prosody is recruited. Comparing speakers' prosody across these two trial types, as well as the degree to which listeners' successful resolution of lexical ambiguity is related to the speakers' prosody, will assess the hypothesis that communicative demand affects referential prosody use. Whereas speakers in Tzeng et al. (2017) were tasked with trying to indicate the intended color swatch to an imaginary listener, participants in the present study completed a referential communication task in pairs. The use of a dyadic task increased the ecological validity of the communicative demands, increasing the likelihood that speakers would convey disambiguating information to the listener. Given previous evidence that speakers are more likely to recruit prosody and co-speech gesture to disambiguate lexical ambiguity, we predicted that speakers and listeners would be more likely to employ referential prosody when lexical content was insufficient to identify the target.

2. Method

2.1. Participants

Sixty (30 speakers, 25 female; 30 listeners, 25 female) Emory University undergraduates participated for course credit. Participants were between 18.11 and 25.20 years of age (M = 19.72, SD = 1.36) and were native English speakers with no history of speech or hearing disorders.[1] Participants were run in pairs (one speaker, one listener) that were not matched on gender.[2] Data from three speaker–listener pairs were excluded due to the speakers' failure to follow task instructions. Data from an additional three speaker–listener pairs were excluded as outliers due to the listeners' accuracy levels falling at least two standard deviations below the mean accuracy across all pairs, yielding a total of 24 speaker–listener pairs to be included in the reported analyses. The resulting sample size matched those of similar dyadic paradigms (e.g., Holler & Stevens, 2007) and allowed adequate power ($1 - \beta = 0.80$) to detect a medium-sized effect (partial $\eta^2 = 0.09$) at $\alpha = 0.05$.

2.2. Stimuli

Visual stimuli presented to the speakers and listeners were drawn from a subset of six color spectra (red, orange, yellow, green, blue, purple) used by Tzeng et al. (2017) that varied in perceived brightness (amount of white) from left to right. Each shade was created by using red, green, and blue (RGB) coordinates. One bright and one dark swatch (1×1.85 inches) from each of the six color spectra were chosen to be presented in pairs in the current experiment. The swatches were normed to ensure that they were (a) consistently associated with the English color labels and (b) equally different in brightness within bright–dark pairs across colors. A separate group of native English-speaking adults (n = 15) completed a computerized task in which they viewed each of the nine swatches in the six color spectra in random order and for each one, provided a single-word color

label that best represented the presented swatch. A different group of native English-speaking adults (n = 15) completed a yes–no task during which they viewed each of the 54 color swatches along with its corresponding single-word color label (e.g., a dark blue swatch labeled as *blue*) and indicated on a response box whether they thought the color label described the color of the presented swatch. Swatches were selected as potential stimuli for the current experiment if they were consistently labeled as a particular color and associated with the color label by at least 70% of participants.

Results from the norming tasks indicated that the bright and dark swatches of red, orange, and yellow spectra were closer to the midpoint of each spectrum than the bright and dark swatches for green, blue, and purple spectra. RGB coordinates of the red, orange, and yellow spectra were then manually adjusted to maximize the perceptual discriminability between bright and dark swatches of each of the three colors. Separate groups of native English-speakers completed a color labeling task (n = 10) and yes-no color label verification task (n = 9) with the adapted color spectra for red, orange, and yellow along with the original green, blue, and purple spectra. One dark and one bright swatch from each of the six spectra were selected using the same criteria described above. A separate group of native-English speakers (n = 21) then viewed all combinations of pairs of color swatches (one bright, one dark) from five[3] of the six color spectra and indicated on a response box whether the two presented swatches were the same or different brightness. Ten color pairs that were rated as different in brightness by at least 80% of participants were selected as stimuli (ambiguous or unambiguous, as explained below) for the current study. Five color pairs consisted of one bright and one dark swatch from the same color spectra (e.g., bright red and dark red), and five pairs from consisted of swatches different color spectra (e.g., bright red and dark blue; see Appendix for selected color swatches and their RGB coordinates).

2.3. Procedure

Participants completed a referential communication task in pairs and were randomly assigned as either a speaker or a listener (e.g., Holler & Stevens, 2007; Keysar, Barr, Balin, & Brauner, 2000). One speaker and one listener were simultaneously seated in a sound-attenuated room at adjacent Dell Optiplex desktop computers separated by an opaque divider. The task objective for the speaker was to convey to the listener, who could not see the speaker's computer screen, which of two color swatches he or she was indicating. On a given trial, the speaker and listener both viewed two swatches, one dark, one bright, presented side by side on the computer screen. The two swatches were identical on the two screens. On the speaker's screen only, one of the two swatches was indicated with an arrow and an English color label (that corresponded to the swatch color [e.g., red, Fig. 1]. Each swatch pair remained on the computer screen while the speaker produced the sentence "Can you get the _____ one?," filling the blank with the provided color label (e.g., "Can you get the red one?"). Speakers were not explicitly told to vary their prosody and were instructed to indicate the color swatch as best as they could to the listener using only the target sentence and label.

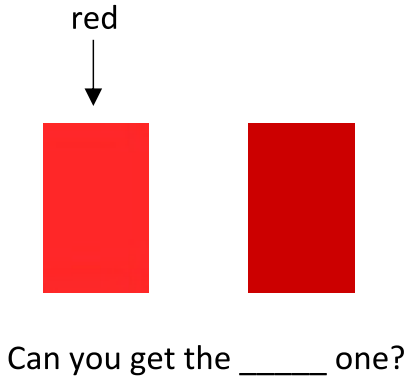

Fig. 1. Speakers' view for a single trial.

After hearing the sentence produced by the speaker, the listener chose which of the two presented swatches corresponded to the one indicated by the speaker by pressing one of two designated keys on a button box corresponding to the left and right swatches on the computer screen. After the listener made his or her response, the experiment was advanced to the next trial by the experimenter.[4]

To assess the extent to which communicative demand affects the recruitment of prosody to resolve referential ambiguity, the experiment consisted of both ambiguous and unambiguous trials. For ambiguous trials, speakers were asked to distinguish between a dark and bright shade of a single color (e.g., bright blue and dark blue, see Fig. 2), whereas for unambiguous trials, speakers were asked to distinguish between a dark and a bright swatch of two different colors (e.g., bright red and dark purple, see Fig. 2). Whether speakers were instructed to label the bright or dark swatch varied across trials. Four differently ordered lists were created, each consisting of three blocks consisting of 20 trials each, with trial type (ambiguous vs. unambiguous), color pairing, and labeled swatch (bright vs. dark) within each pair pseudo-randomized within each block such that participants never saw the same swatches in consecutive trials. To familiarize participants with the task and the distinction between trial types, participants completed two practice trials (one ambiguous, one unambiguous) during which the speaker was told to describe, and the listener to identify, a dark and a bright shade of gray or yellow. Neither speakers nor listeners received corrective feedback during the practice trials or the experimental task. All experiment and consent procedures were approved by the Emory University Institutional Review Board.

Participants sat approximately 18 inches from the computer screen, with the microphone placed on the table 6 inches in front of the speaker. Speakers' utterances were recorded using an audio-technica ATR 20 microphone onto the speaker's computer and segmented by trial using E-prime 2.0 (Schneider et al., 2002). Sentence utterances were re-digitized at a 22.050 kHz sampling rate and amplitude normalized using SoundStudio software. To examine the acoustic features of the individual color labels in addition to

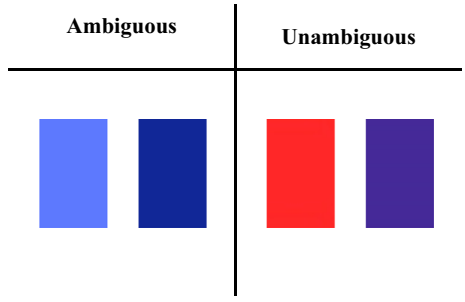

Fig. 2. Examples of ambiguous and unambiguous trials.

the sentence-length utterances, each color label was segmented from the sentence utterance and amplitude normalized.[5]

2.4. Acoustic analyses

Acoustic measures were obtained for both the speakers' sentence-length utterances and separately for the segmented color labels in isolation, as acoustic correlates to brightness might not be confined to speakers' production of the color labels but may extend across the sentence. Given previous evidence for a pitch-brightness correspondence, we were primarily interested in pitch as our target acoustic measure. However, because previous findings have suggested systematic correspondences between other acoustic measures, such as duration, and amplitude, and visual dimensions, including brightness, size, shape, and visuo-spatial location (e.g., Marks, 1974; Mondloch & Maurer, 2004; Nygaard et al., 2009; Shintel et al., 2006), we also examined the extent to which utterance duration and amplitude, individually for both sentences and single-word color labels, varied as a function of brightness. Mean fundamental frequency (F_0), mean amplitude, and duration were measured using an automated script executed in PRAAT.[6] F_0 refers to the number of cycles per second in a periodic sound and corresponds to the perception of pitch. Amplitude reflects the overall energy of the utterance and corresponds to the perception of loudness. Duration is the overall length of the utterance, which, given the fixed sentence lengths across brightness conditions in this experiment, serves as an index of speaking rate.

3. Results

3.1. Speaker performance

Table 1 shows the F_0, amplitude, and duration of sentence-length utterances for bright and dark shades for both ambiguous and unambiguous trials. Because the pattern of statistically reliable main effects and interactions was similar for sentence- and word-level analyses, only sentence-level results are reported. Any cases where the word-level results deviated from those at the sentence-level are noted. Because the acoustic profiles of the

five color labels differed in number of syllables and in phonetic composition, means were collapsed across color. Three repeated-measures ANOVAS (one for each dependent measure) assessed the extent to which speakers' prosody varied as a function of Trial Type (ambiguous, unambiguous), Brightness (bright, dark), and Block (1, 2, 3) as within-subjects variables. Results for F_0, amplitude, and duration are reported separately below.

3.1.1. F_0

Results indicated a significant effect of brightness, $F(1, 23) = 8.13$, $p = .009$, partial $\eta^2 = 0.26$, modified by a significant interaction between Trial Type and Brightness, $F(1, 23) = 11.91$, $p = .002$, partial $\eta^2 = 0.34$ (Fig. 3). Follow-up pairwise comparisons assessing the difference between bright and dark pitch for ambiguous and unambiguous trials separately indicated that bright sentences ($M = 208.78$, $SD = 37.60$) were reliably higher pitched than dark sentences ($M = 197.03$, $SD = 36.45$) for ambiguous, $t(23) = 3.16$, $p = .004$, but not unambiguous trials, $t(23) = -0.54$, $p = .593$. No other main effects or interactions were significant.

3.1.2. Amplitude

Sentences did not differ reliably in amplitude for any of the independent variables of interest, and there were no significant interactions.

3.1.3. Duration

Sentences differed reliably in duration between trial types, $F(1, 23) = 26.87$, $p < .001$, partial $\eta^2 = 0.54$, such that sentences for ambiguous trials ($M = 1{,}562.31$, $SD = 329.78$) were significantly longer than for unambiguous trials ($M = 1{,}249.56$, $SD = 188.65$; $t(23) = 5.21$, $p < .004$). There were no other significant main effects or interactions for the sentence-level analyses. At the word level, the repeated-measures ANOVA yielded main effects of Trial Type, $F(1, 23) = 31.75$, $p < .001$, partial $\eta^2 = 0.59$, and Brightness, $F(1, 23) = 7.94$, $p = .010$, partial $\eta^2 = 0.26$. The interactions between Trial Type and Brightness, $F(1, 23) = 5.01$, $p = .035$, partial $\eta^2 = 0.18$, and between Block and Brightness, $F(2, 46) = 4.87$, $p = .023$, partial $\eta^2 = 0.18$, were also significant. These main effects and interactions were modified by a significant three-way interaction between Trial Type, Brightness, and Block, $F(2, 46) = 5.24$, $p = .009$, partial $\eta^2 = 0.19$ (see Fig. 4). Follow-up Bonferroni-adjusted paired comparisons suggested that for ambiguous trials, labels for dark swatches ($M = 496.50$, $SD = 152.32$) were reliably longer than for bright swatches

Table 1
Acoustic measurements of sentence-length utterances

	Ambiguous		Unambiguous	
	Bright	Dark	Bright	Dark
F_0 (Hz)	208.59	197.19	201.24	201.25
Amplitude (dB)	74.03	74.14	74.39	74.32
Duration (ms)	1,510.75	1,617.35	1,241.51	1,259.34

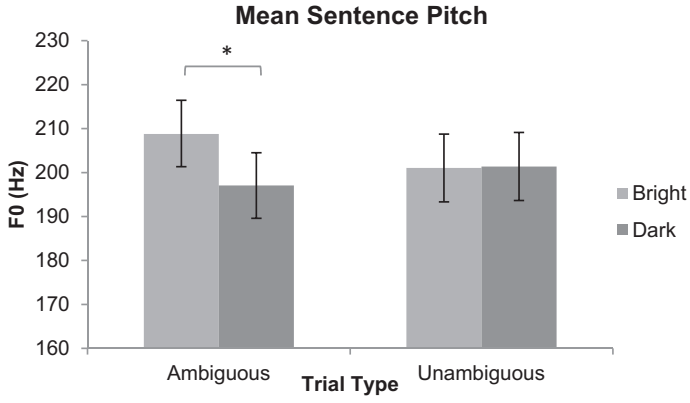

Fig. 3. Mean sentence pitch as a function of trial type and brightness. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

($M = 391.48$, $SD = 106.13$) in Block 2, $t(23) = -3.17$, $p = .024$. All other comparisons between bright and dark were non-significant within each block for ambiguous and unambiguous trials. That labels for dark swatches were significantly longer than those for bright swatches in Block 2 for ambiguous trials suggests that in addition to recruiting pitch to convey brightness information, speakers here also employed duration differences to distinguish between bright and dark. However, because this pattern was not consistent across blocks and appeared only at the level of the individual color word but not the full sentence, duration did not appear to be a robust cue to brightness.

3.2. Listener performance

Listeners' response accuracy was measured by the proportion of times listeners chose the bright swatch after hearing a sentence referring to a bright shade, or a dark swatch after hearing a sentence referring to a dark shade. Listeners reliably chose the correct corresponding color swatch for both unambiguous ($M = 0.998$, $SD = 0.01$; $t(23) = 287.75$, $p < .001$) and ambiguous ($M = 0.83$, $SD = 0.19$; $t(23) = 8.23$, $p < .001$) trials. Above-chance performance for the unambiguous trials is expected and suggests that listeners understood the two-alternative forced-choice task. Above-chance performance for the ambiguous trials suggests that listeners, in the absence of lexical cues to color brightness, inferred brightness information from speakers' prosody in order to choose the correct target referent.

A repeated-measures ANOVA assessed the extent to which listeners' accuracy varied as a function of Trial Type (ambiguous, unambiguous), Brightness (bright, dark), and Block (1, 2, 3) as within-subjects variables. Results yielded significant main effects of Trial Type, indicating higher accuracy for unambiguous versus ambiguous trials, $F(1, 23) = 19.80$, $p < .001$, partial $\eta^2 = 0.46$, and of Block, indicating improvement across blocks, $F(2, 46) = 4.38$, $p = .018$, partial $\eta^2 = 0.16$, as well as a significant interaction between these two variables, $F(2, 46) = 4.44$, $p = .017$, partial $\eta^2 = 0.16$ (Fig. 5). All

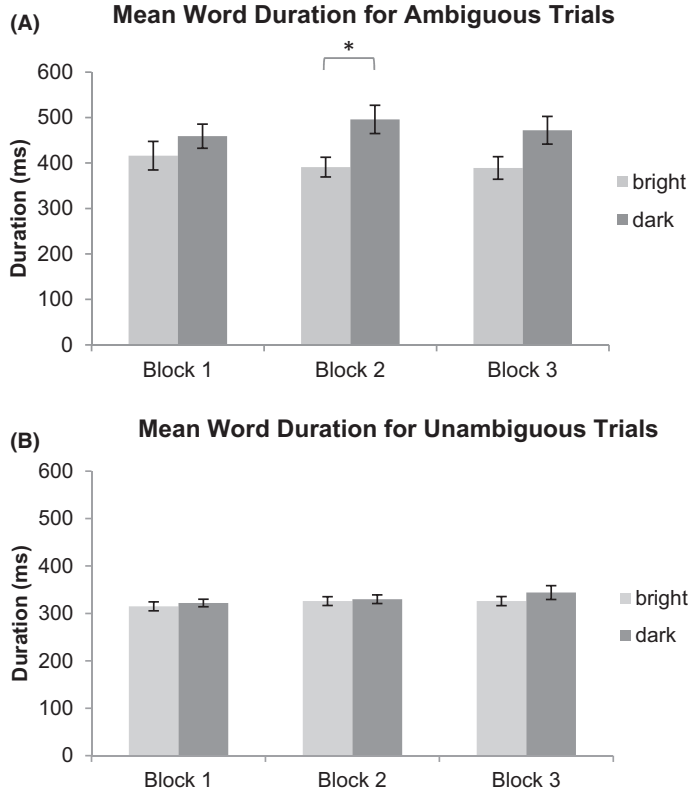

Fig. 4. Mean color word duration for ambiguous (A) and unambiguous trials (B) as a function of brightness and block. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

other main effects and interactions were non-significant. To explore the trial type by block interaction, an ANOVA was conducted to assess the effect of Block separately for each trial type. Results revealed a significant effect of Block for ambiguous, $F(1, 23) = 6.44$, $p = .018$, partial $\eta^2 = 0.22$, but not unambiguous trials, $F(2, 46) = 0.49$, $p = .616$. Follow-up pairwise comparisons across blocks for ambiguous trials indicated a significant increase in listener accuracy between the first and second ($M_1 = 0.76$, $SD_1 = 0.23$; $M_2 = 0.85$, $SD_2 = 0.19$; $t(23) = 2.82$, $p = .010$) and first and third blocks ($M_1 = 0.76$, $SD_1 = 0.23$; $M_3 = 0.86$, $SD_3 = 0.24$; $t(23) = 2.54$, $p = .018$) of the task, suggesting that listeners learned to infer relevant acoustic information from speakers' utterances more systematically across blocks.

3.3. Relation between listener and speaker performance

To assess the extent to which listener performance varied as a function of the robustness of the speakers' prosodic correlates to brightness, response accuracy was separately

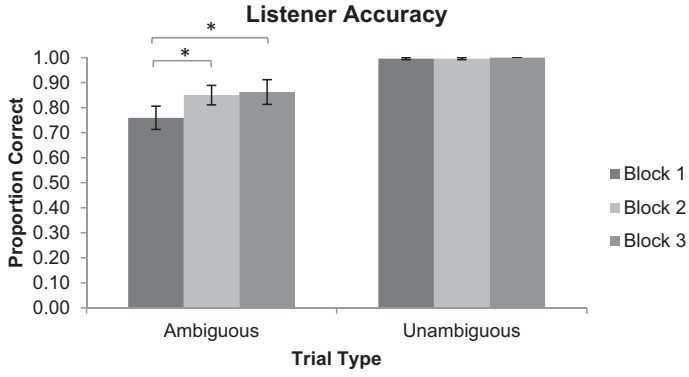

Fig. 5. Listener accuracy as a function of block and trial type. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

regressed on the mean difference between speakers' pitch, duration, and amplitude values for bright versus dark sentences. Three difference scores (one for each acoustic measure) were calculated for each speaker by subtracting the mean sentence pitch, duration, and amplitude of all his or her dark sentences from the mean pitch, duration, and amplitude of all his or her bright sentences. Unlike absolute values, these difference scores (a) account for baseline differences in individual speakers' acoustic characteristics and (b) capture the magnitude of each speaker's prosodic modulations across bright and dark sentences.

3.3.1. F_0

Fig. 6 shows listeners' response accuracy collapsed across bright and dark ambiguous trials as a function of each speaker's pitch difference score. A linear regression equation regressing accuracy on this difference score accounted for a significant portion of variance in accuracy, $R^2 = 0.18$, $F(1, 23) = 4.74$, $p = .040$. Difference scores in pitch reliably predicted accuracy, $\beta = 0.42$, $t(23) = 2.18$, $p = .040$, such that the larger a speaker's difference score, the more accurate listeners' mappings were for both bright and dark sentences. A linear regression equation regressing accuracy on the difference score calculated at the word level did not account for a significant portion of variance in accuracy, $R^2 = 0.12$, $F(1, 23) = 3.11$, $p = .070$, nor did difference scores reliably predict accuracy, $\beta = 0.35$, $t(23) = 1.77$, $p = .091$, suggesting that informative cues to brightness were not localized to the word level.

3.3.2. Amplitude and duration

Linear regression analyses were also conducted to assess the extent to which difference scores in amplitude and duration predicted listeners' response accuracy. Neither difference scores in amplitude nor duration of sentences accounted for a significant portion of variance in accuracy (amplitude, $R^2 = 0.0002$, $F(1, 23) = 0.01$, $p = .945$; duration, $R^2 = 0.005$, $F(1, 12) = 0.12$, $p = .737$), nor did they reliably predict accuracy (amplitude,

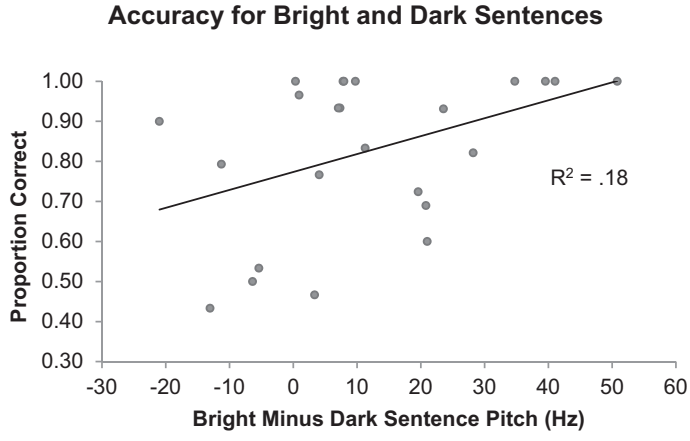

Fig. 6. Each data point represents average listener accuracy across ambiguous trials for each dyad as a function of the speaker's mean difference in pitch between sentences referring to bright versus dark shades. Chance performance is at 0.50.

$\beta = 0.02$, $t(23) = 0.07$, $p = .945$; duration, $\beta = -0.07$, $t(23) = -0.34$, $p = .737$), suggesting that listeners did not infer brightness from the relative difference in amplitude and duration between bright and dark sentences.

4. General discussion

The current results for both speaker performance and listener accuracy suggest a prominent role of communicative demand in the use of prosodic cues to resolve referential ambiguity. Speakers' bright sentences were reliably higher pitched than dark sentences for ambiguous, but not unambiguous trials, suggesting that speakers indeed provided meaningful acoustic cues to brightness level but only when the accompanying linguistic content was underspecified. During unambiguous trials, when lexical content was sufficient to identify the target swatch, speakers did not employ disambiguating prosodic cues, as there was no communicative need to do so. That speakers varied their approach between the two trial types here speaks to the flexibility of the communicative system to provide informative, task-relevant cues to meaning. That listeners reliably chose the correct corresponding swatch for ambiguous trials, when lexical information alone was insufficient to identify the referential target, suggests that listeners recruited acoustic cues in the speakers' utterances to disambiguate between the two choice responses.

The novel contribution of this work lies in the finding that the use of referential prosody varies as a function of communicative need to resolve lexical ambiguity. Beyond what can be inferred through pragmatic inference, language users employ and infer referential detail from prosody when propositional content alone is insufficient for listener

comprehension. The current results extend previous findings suggesting that referential prosody is more likely to be recruited with increased need to resolve underspecified linguistic content. In Tzeng et al. (2017), the recruitment of correspondences between pitch and brightness occurred when participants labeled colors with novel color labels (e.g., "Can you get the blicket one?") but not with English color words (e.g., "Can you get the red one?"), suggesting that speakers are more likely to produce prosodic correlates to brightness when the accompanying linguistic content is relatively underspecified. A potential alternative, but not mutually exclusive, explanation for this pattern of findings is that English color labels were more familiar and thus produced more automatically (Hickok & Poeppel, 2007). In the current study, word familiarity is controlled across the two experimental conditions, thus ensuring that any differences across groups were solely attributed to communicative demand to resolve lexical ambiguity.

That the demand to resolve lexical ambiguity affects referential prosody use is consistent with recent findings suggesting that communicative context modulates speakers' use of prosody to signal new or contrastive information. Buxó-Lugo, Toscano, and Watson (2018), for example, found that relative to participants who completed a color description task performed in isolation, those who completed a dyadic navigation task using the same stimuli produced more discriminable cues (changes in F0 and duration) to highlight relevant contrastive information to the listener. Together with the results of the current study, the prosodic modulations produced in Buxó-Lugo et al. (2018) provide accumulating evidence that prosody use varies as a function of the need to maximize listener comprehension.

The current findings suggest that language users readily engage in prosodic modulation to resolve lexical ambiguity in natural discourse. Alternatively, speaker and listener performance may have been a product of the particular task demands of the experimental situation. Given the dyadic nature of the task, speakers were presumably highly motivated to communicate task-relevant information to the listener. The content of the speakers' utterances was also constrained such that speakers could only use non-lexical cues rather than lexical ones (e.g., "Can you get the *bright* red one?") to disambiguate between brightness levels, thus increasing the likelihood that speakers would recruit alternative acoustic cues to convey intended information. However, speakers were never explicitly told whether or how to modulate their prosody, nor were they told to convey brightness information in any systematic manner. That the pitch–brightness relation in speakers' utterances was consistent with previous demonstrations of the pitch–brightness association using both non-linguistic (Haryu & Kajikawa, 2012; Marks, 1987; Melara, 1989; Mondloch & Maurer, 2004) and linguistic (Tzeng et al., 2017) stimuli suggests that the current results are an instantiation of cross-modal associations being recruited to convey referential information when communicatively relevant, rather than simply an artifact of task demands.

The current task was not interactive in the sense that listeners and speakers did not freely converse with each other during the experimental session. As such, speakers were not adapting their responses in accordance with dynamic levels of listener comprehension, feedback, or performance. Although the paradigm employed here is less interactive and

thus perhaps less naturalistic than paradigms that allow the speaker and listener to adjust their behavior in response to their dyadic partner (e.g., Buxó-Lugo et al., 2018; Clark & Krych, 2004), one advantage of the current procedure is that it isolates the influence of one specific aspect of communicative context: the need to resolve lexical ambiguity. Given evidence that speakers adapt their utterances to maximize listener comprehension (e.g., Brown-Schmidt, 2009; Buxó-Lugo et al., 2018; Clark & Krych, 2004), it is possible that speakers in the current study might have been even more likely to recruit referential prosody if they had knowledge of listeners' trial-to-trial task accuracy.

The present findings suggest that prosody can be recruited as an additional channel of referential detail that accompanies and potentially clarifies the linguistic utterance. Characterized in this way, referential prosody can be conceptualized as analogous to co-speech gesture (Perlman, 2010). Co-speech gestures often convey details that are non-redundant with the accompanying utterance (Broaders & Goldin-Meadow, 2010; Goldin-Meadow & Singer, 2003; Holler & Stevens, 2007; McNeill, 1992; Melinger & Levelt, 2004). Of particular relevance to the current study are findings suggesting that co-speech gesture is especially prominent when lexical content is ambiguous or underspecified and may facilitate the resolution of lexical ambiguity, as in the case of homonyms (e.g., Holler & Beattie, 2003). A related proposal is that speakers recruit gestures when they are particularly motivated to communicate clearly. When conveying task-relevant information, speakers are more likely to produce more and larger gestures for novice versus expert listeners (Campisi & Özyürek, 2013) and more representational gestures when the gestural information is thought to be particularly useful for the listener (Kelly et al., 2011). Speakers are thus aware of the listeners' knowledge base and goals and adjust their gestures accordingly to maximize the likelihood of comprehension. Referential prosody, like co-speech gesture, can thus be conceptualized as a means by which interlocutors can modify their own behavior to better ensure mutual intelligibility (e.g., Kröger et al., 2010; Pezzulo, Donnarumma, & Dindo, 2013).

The current findings join accumulating evidence that prosody not only provides affective, syntactic, and pragmatic information but also can be used to describe physical properties of external linguistic referents, including visuo-spatial height (Shintel et al., 2006), object speed (Shintel & Nusbaum, 2007), and size and strength (Herold et al., 2011). Pitch, in particular, is strongly implicated in many of these auditory–visual mappings. As pitch is a salient feature of an utterance's prosodic contour, it is perhaps unsurprising that relative to duration and amplitude, pitch provided the most reliable disambiguating information about brightness in the current study. A related possibility is that duration and amplitude cues had the potential to disambiguate reference but were employed inconsistently across speakers, with some speakers using these cues more reliably than others. That there was variability in listener accuracy across speakers with similar pitch difference scores (see Fig. 6) is suggestive of this possibility.

Although previous work has found that speakers recruit the pitch-height mapping in their descriptions of vertical movement even in the absence of any demand to resolve ambiguity (Shintel et al., 2006), this finding is not entirely inconsistent with the current results. Unlike the pitch-brightness mapping, the pitch–height mapping is readily observed

in the natural environment such that high-pitched sounds tend to originate from higher elevations (Parise, Knorre, & Ernst, 2014). The sound of rustling leaves on trees, for example, occurs at a higher pitch range than the sound of footsteps on the ground. The mapping between pitch and height is also semantically mediated, as both pitch and height can be described as *high* and *low*. Speakers of languages that do not share descriptors of pitch and height have been shown to exhibit lessened sensitivity to the pitch–height mapping (e.g., Dolscheid, Shayan, Majid, & Casasanto, 2013), suggesting that semantic mediation may augment or attenuate the extent to which listeners are sensitive to specific cross-modal associations.

Manifestation of cross-modal correspondences in prosody is consistent with the view that language is grounded in multi-modal experiences (Barsalou, 1999, 2003; Glenberg & Kaschak, 2002; Zwaan, Madden, Yaxley, & Aveyard, 2004). Integral to grounded theories of language is the assumption that during linguistic processing, language users experience simulations of perceptual experiences associated with external referents (Barsalou, 1999; Matlock, 2004; Šetić & Domijan, 2007; Zwaan, Stanfield, & Yaxley, 2002). This view aligns well with the characterization of prosody as a source of analog acoustic expression (Shintel et al., 2006), as the use of referential prosody to convey perceptual detail may be a natural means of expressing meaning that is grounded in the perceptual system.

Prosodic cues have long been thought to provide many types of information crucial for effective communication. The current work expands traditional characterizations of prosody, as it implies that prosody can be recruited to convey meaningful *referential* information about external linguistic referents. Language users employ and infer referential detail from prosody when propositional content alone is insufficient for listener comprehension. Prosody thus capitalizes on cross-modal associations to extend the range of referential information that can be conveyed in discrete linguistic units. The current findings suggest that prosodic cues can be conceptualized as a type of vocal gesture that is recruited to resolve referential ambiguity when there is communicative demand to do so.

Notes

1. Participants were screened for hearing and speech disorders via self-reported responses on a questionnaire that was completed prior to the start of the experiment. Participants were not screened for color blindness. However, each speaker–listener pair completed two practice trials (one ambiguous, one unambiguous) during which the speaker was told to describe, and the listener to identify, a dark and a bright shade of a gray or yellow. Participants were also told that they would be viewing between one bright swatch and one dark swatch on each trial. These steps ensured that participants were aware of differences in brightness among shades regardless of the specific color instantiated.
2. We did not control for whether speakers and listeners in each pair were familiar with one another. All participants were recruited from the same participant pool.

However, none of the participants showed any explicit indication of familiarity with his or her dyadic partner before or after administration of the experimental protocol when they did have the opportunity to interact, lessening the likelihood that variation in performance in the experimental task was due to potential differences in the nature of the relationship across speaker–listener pairs.

3. Swatches from the yellow spectrum were not included in the brightness rating task or in the experiment, as pilot data suggested that the bright and dark yellow swatches were consistently rated as brighter than the other bright and dark swatches of the other five colors.
4. To align the content of each trial across the listener and speaker's screens, the listener's trials were advanced automatically in E-prime 2.0 while the speaker's trials were presented in Microsoft PowerPoint and manually advanced by the experimenter, who stood out of view, behind each pair of participants.
5. Utterances were amplitude normalized to account for variations in overall amplitude due to differences in distance between the speaker and the microphone. Although participants were reminded to maintain a constant distance between themselves and the microphone, it is likely that their position shifted during the experiment. Such shifts primarily affect overall amplitude but not pitch and duration. Amplitude normalization preserves relative amplitude differences within an utterance by adjusting the amplitude throughout the file by the same amount, insuring that any differences across conditions were due to differences in the vocal effort or amplitude of the utterances and not due to distance from the microphone.
6. Mean F0 measures were obtained in PRAAT using an autocorrelation pitch extraction algorithm (Boersma, 1993), with minimum and maximum pitch settings set at 75 Hz and 500 Hz, respectively. Minimum and maximum amplitude settings were set at 50 dB and 100 dB, respectively.

References

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production. *Journal of Memory and Language*, *44*, 169–188. https://doi.org/10.1006/jmla.2000.2752

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660. https://doi.org/10.1017/S0140525X99532147

Barsalou, L. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes*, *18*(5–6), 513–562.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-tonoise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, *17*, 97–119.

Bögels, S., Schriefers, H., Vonk, W., & Chwilla, D. J. (2011). Pitch accents in context: How listeners process accentuation in referential communication. *Neuropsychologia*, *49*(7), 2022–2036. https://doi.org/10.1016/j.neuropsychologia.2011.03.032

Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, *10*(2), 137–167. https://doi.org/10.1016/j.jml.2003.08.004

Broaders, S. C., & Goldin-Meadow, S. (2010). Truth is at hand: How gesture adds information during investigative interviews. *Psychological Science*, *21*(5), 623–628. https://doi.org/10.1177/0956797610366082

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*, 171–190. https://doi.org/10.1016/j.jml.2009.04.003

Buxó-Lugo, A., Toscano, J. C., & Watson, D. G. (2018). Effects of participant engagement on prosodic prominence. *Discourse Processes*, *55*(3), 305–323. https://doi.org/10.1080/0163853X.2016.1240742

Campisi, E., & Özyürek, A. (2013). Iconicity as a communicative strategy: Recipient design in multimodal demonstrations for adults and children. *Journal of Pragmatics*, *47*(1), 14–27. https://doi.org/10.1016/j.pragma.2012.12.007

Chiou, R., & Rich, A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception*, *41*(3), 339–353. https://doi.org/10.1068/p7161

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81. https://doi.org/10.1016/j.jml.2003.08.004

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113–121. https://doi.org/10.1037/0096-1523.14.1.113

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological Science*, *24*(5), 613–621. https://doi.org/10.1177/0956797612457374

Galati, A., & Brennan, S. E. (2014). Speakers adapt gestures to addressees' knowledge: Implications for models of co-speech gesture. *Language, Cognition and Neuroscience*, *29*(4), 435–451. https://doi.org/10.1080/01690965.2013.796397

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin and Review*, *9*(3), 558–565. https://doi.org/10.3758/BF03196313

Glucksberg, S. (1986). How people use context to resolve ambiguity: Implications for an interactive model of language understanding. *Advances in Psychology*, *39*, 303–325.

Goldin-Meadow, S., & Singer, M. A. (2003). From children's hands to adults' ears: Gesture's role in the learning process. *Developmental Psychology*, *39*(3), 509–520. https://doi.org/10.1037/0012-1649.39.3.509

Grice, P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Haryu, E., & Kajikawa, S. (2012). Are higher-frequency sounds brighter in color and smaller in size? Auditory–visual correspondences in 10-month-old infants. *Infant Behavior and Development*, *35*(4), 727–732. https://doi.org/10.1016/j.infbeh.2012.07.015

Herman, R. (2000). Phonetic markers of global discourse structures in English. *Journal of Phonetics*, *28*, 466–493. https://doi.org/10.1006/jpho.2000.0127

Herold, D. S., Nygaard, L. C., Chicos, K. A., & Namy, L. L. (2011). The developing role of prosody in novel word interpretation. *Journal of Experimental Child Psychology*, *108*(2), 229–241. https://doi.org/10.1016/j.jecp.2010.09.005

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Holler, J., & Beattie, G. (2003). Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener? *Gesture*, *3*(2), 127–154. https://doi.org/10.1075/gest.3.2.02hol

Holler, J., & Stevens, R. (2007). The effect of common ground on how speakers use gesture and speech to represent size information. *Journal of Language and Social Psychology*, *26*, 4–27. https://doi.org/10.1177/0261927X06296428

Hostetter, A. B., Alibali, M. W., & Schrager, S. M. (2011). If you don't already know, I'm certainly not going to show you! Motivation to communicate affects gesture production. In G. Stam & M. Ishino

(Eds.), *Integrating gestures: The interdisciplinary nature of gesture* (pp. 61–74). Philadelphia, PA: John Benjamins.

Hupp, J. M., & Jungers, M. K. (2013). Beyond words: Comprehension and production of pragmatic prosody in adults and children. *Journal of Experimental Child Psychology*, *115*(3), 536–551. https://doi.org/10.1016/j.jecp.2012.12.012

Jacobs, N., & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language*, *56*(2), 291–303. https://doi.org/10.1016/j.jml.2006.07.011

Kelly, S., Byrne, K., & Holler, J. (2011). Raising the ante of communication: Evidence for enhanced gesture use in high stakes situations. *Information*, *2*(4), 579–593. https://doi.org/10.3390/info2040579

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38. https://doi.org/10.1111/1467-9280.00211

Kröger, B. J., Kopp, S., & Lowit, A. (2010). A model for production, perception, and acquisition of actions in face-to-face communication. *Cognitive Processing*, *11*(3), 187–205. https://doi.org/10.1007/s10339-009-0351-2

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch and loudness. *American Journal of Psychology*, *87*, 173–188. https://doi.org/10.2307/1422011

Matlock, T. (2004). Fictive motion as cognitive simulation. *Memory & Cognition*, *32*(8), 1389–1400.

Marks, L. E. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(3), 384–394. https://doi.org/10.1037/0096-1523.13.3.384

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: University of Chicago Press.

Melara, R. D. (1989). Dimensional interaction between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(1), 69–79. https://doi.org/10.1037//0096-1523.15.1.69

Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture*, *4*, 119–141. https://doi.org/10.1075/gest.4.2.02mel

Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch–object correspondences in young children. *Cognitive, Affective, and Behavioral Neuroscience*, *4*(2), 133–136. https://doi.org/10.3758/CABN.4.2.133

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive Science*, *33*(1), 127–146. https://doi.org/10.1111/j.1551-6709.2008.01007.x

Parise, C. V., Knorre, K., & Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences*, *111*, 6104–6108. https://doi.org/10.1073/pnas.1322705111

Perlman, M. (2010). Talking fast: The use of speech rate as iconic gesture. In F. Parrill, V. Tobin, & M. Turner (Eds.), *Meaning, form, and body* (pp. 245–262). Stanford, CA: CSLI Publications.

Perlman, M., Clark, N., & Falck, M. J. (2015). Iconic prosody in story reading. *Cognitive Science*, 1348–1368. https://doi.org/10.1111/cogs.12190

Pezzulo, G., Donnarumma, F., & Dindo, H. (2013). Human sensorimotor communication: A theory of signaling in online social interactions. *PLoS ONE*, *8*(11), e79876. https://doi.org/10.1371/journal.pone.0079876

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools Inc.

Šetić, M., & Domijan, D. (2007). The influence of vertical spatial orientation on property verification. *Language and Cognitive Processes*, *22*(2), 297–312. https://doi.org/10.1080/01690960600732430

Shintel, H., & Nusbaum, H. C. (2007). The sound of motion in spoken language: Visual information conveyed by acoustic properties of speech. *Cognition*, *105*(3), 681–90. https://doi.org/10.1016/j.cognition.2006.11.005

Shintel, H., & Nusbaum, H. C. (2008). Moving to the speed of sound: Context modulation of the effect of acoustic properties of speech. *Cognitive Science*, *32*(6), 1063–1074. https://doi.org/10.1080/03640210801897831

Shintel, H., Nusbaum, H. C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language*, *55*(2), 167–177. https://doi.org/10.1016/j.jml.2006.03.002

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology*, *24*(1), 99–142. https://doi.org/10.1016/0010-0285(92)90004-L

Snedeker, J., & Trueswell, J. C. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*(1), 103–130. https://doi.org/10.1016/S0749-596X(02)00519-3

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tzeng, C. Y., Duan, J., Namy, L. L., & Nygaard, L. C. (2017). Prosody in speech as a source of referential information. *Language, Cognition and Neuroscience*, *33*(4), 512–526.

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referents in time: Eye-tracking evidence for the role of contrastive accents. *Language and Speech*, *49*(3), 367–392. https://doi.org/10.1177/00238309060490030301

Wurm, L. H., Vakoch, D. A., Strasser, M. R., Calin-Jageman, R., & Ross, S. E. (2001). Speech perception and vocal expression of emotion. *Cognition and Emotion*, *15*(6), 831–852. https://doi.org/10.1080/02699930143000086

Zwaan, R. A., Madden, C. J., Yaxley, R. H., & Aveyard, M. E. (2004). Moving words: Dynamic representations in language comprehension. *Cognitive Science*, *28*(4), 611–619. https://doi.org/10.1016/j.cogsci.2004.03.004

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, *13*(2), 168–171.

Appendix 1:

Appendix

Ambiguous		Unambiguous	
R: 255 G: 40 B: 40	R: 204 G: 0 B: 0	R: 102 G: 102 B: 255	R: 0 G: 153 B: 0
R: 255 G: 58 B: 170	R: 228 G: 0 B: 91	R: 255 G: 58 B: 170	R: 0 G: 0 B: 153
R: 102 G: 102 B: 255	R: 0 G: 0 B: 153	R: 178 G: 255 B: 102	R: 255 G: 58 B: 170
R: 102 G: 255 B: 102	R: 0 G: 153 B: 0	R: 255 G: 40 B: 40	R: 76 G: 153 B: 0
R: 178 G: 255 B: 102	R: 76 G: 153 B: 0	R: 102 G: 255 B: 102	R: 204 G: 0 B: 0