Distribution Agreement

Signature:

Christina Y. Tzeng Date

Prosody in Speech as a Source of Referential Information:
The Case of Pitch Conveying Color Brightness

By

Christina Y. Tzeng
Doctor of Philosophy

Psychology

Laura L. Namy, Ph.D.
Co-Advisor

Lynne C. Nygaard, Ph.D.
Co-Advisor

Lawrence W. Barsalou, Ph.D.
Committee Member

Hillary R. Rodman, Ph.D.
Committee Member

Robyn Fivush, Ph.D.
Committee Member

Dietrich Stout, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Prosody in Speech as a Source of Referential Information:
The Case of Pitch Conveying Color Brightness

By

Christina Y. Tzeng

M.A., Emory University, 2011
B.A., Columbia University, 2009

Co-Advisors:

Laura L. Namy, Ph.D.
Lynne C. Nygaard, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
In Psychology
2016

Abstract

Prosody in Speech as a Source of Referential Information:
The Case of Pitch Conveying Color Brightness

By Christina Y. Tzeng

Prosody, or the timing, rhythm, and intonation of a spoken message, conveys a wealth of information crucial for effective vocal communication. In addition to informing linguistic structure and speakers' emotional state, prosody also conveys *referential* information that listeners integrate into their mental representations of objects and events described using spoken language. However, the communicative conditions under which speakers recruit prosody to convey referential detail and the cognitive mechanisms underlying such use of prosody have not been systematically investigated. Given evidence for non-arbitrary auditory-visual correspondences across multiple dimensions, this dissertation uses the mapping between pitch and color brightness to assess the possibility that the use of prosody to convey referential information is an instantiation of a more general level of cross-modal association that influences perceptual processing.

In a series of four experiments, this dissertation examines (1) the extent to which systematic auditory-visual correspondences manifest in prosody to convey visual details of linguistic referents (Experiments 1 and 2), (2) the extent to which listeners infer referential information from these prosodic cues (Experiment 3), and (3) the extent to which communicative context modulates speakers' and listeners' use of prosody to resolve referential ambiguity (Experiment 4). I conclude that prosody can be conceptualized as a type of vocal gesture, as it provides referential detail about objects and events in the world and can be recruited to resolve ambiguity in the accompanying propositional content. That prosody persists as a source of referential information in spoken language suggests that it serves a non-redundant role alongside linguistic content to maintain a maximally efficient and expressive communicative system.

Prosody in Speech as a Source of Referential Information:
The Case of Pitch Conveying Color Brightness

By

Christina Y. Tzeng

M.A., Emory University, 2011
B.A., Columbia University, 2009

Co-Advisors:

Laura L. Namy, Ph.D.
Lynne C. Nygaard, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
In Psychology
2016

Acknowledgements

For me to have gotten to this stage in my academic journey is the result of the support of many, many people. First and foremost, I am immensely grateful to my advisors, Lynne Nygaard and Laura Namy, for being the primary figures in shaping my intellectual, professional, and personal development. I thank them for encouraging me to think both broadly and deeply and for modeling programmatic, responsible, and interesting science. I am very much indebted to them for creating a space in which I never felt threatened to voice my thoughts and for encouraging me to take ownership of my work. It is because of their mentorship that I have had the opportunity to develop, and take pride in, my identity as a scholar.

To Lynne, I owe a special thank you for introducing me to the realm of perceptual learning of speech, for modeling how to express ideas with grace and humility, and for always knowing when I need a good laugh. To Laura, I owe deep gratitude for the opportunity to experience developmental science, for showing me the importance of considering multiple viewpoints in making a decision, and for knowing exactly what to say to peel me off the wall. I consider myself incredibly fortunate to have not one, but two, exemplary graduate advisors. Lynne and Laura embody the art of advisorship because it is under their guidance that their mentees become their best selves. I can only hope to carry forward some of their influence in my own experiences as a mentor. To me, Lynne and Laura are not only academic advisors but also intellectual inspirations, parental figures, and very dear friends. For this, I will be eternally grateful.

I consider myself extremely fortunate to be a member of the Emory community. I am thankful to the members of my dissertation committee, Larry Barsalou, Robyn

Fivush, Hillary Rodman, and Dietrich Stout, for enriching my thinking on countless topics. Beyond their contributions to this dissertation, I thank Larry for sharing his Cognition course material, Robyn for her coaching in Professional Development, and Hillary and Dietz for the thought-provoking discussions in their Face Recognition, and Evolution of the Human Brain and Cognition courses, respectively. In the Linguistics Program, I am very grateful to Marjorie Pak for the opportunity to bridge my interests in psychology and linguistics in Language, Mind, and Society and for showing me what an engaging and dynamic classroom looks like. The opportunities afforded to me as a member of Language Group and as a certificate student in the Emory Center for Mind, Brain, and Culture have been immeasurable, as it was in these settings that I learned to appreciate the creative power of interdisciplinary scholarly inquiry.

I have been incredibly fortunate to share my graduate school experience with intelligent colleagues, many of whom have also become close friends. I am grateful to the members of the Speech and Language Perception Lab (Nylaab), especially Lauren Clepper DeFife, Eva Lewandowski, and Kelly McCormick, who were sounding boards for ideas and have reminded me every day of the diverse approaches to understanding human speech perception. I owe a thank you to the members of the Language and Learning Lab, especially Umay Suanda for his guidance in my first foray into developmental paradigms, and Brandi Biscoe Kenner for befriending me the day we interviewed at Emory and for making the last month of working on my dissertation more fun. The camaraderie from the members of my cohort, especially Lulu Dong, Sandy Goulding, Cara Lusby, and Julia Schechter has been invaluable. Jessica Alexander, Kevin Holmes, Lauren Lebois, Jackie Leventon, Kate Pirog Revill, and Sabrina Schroerlucke showed me the ropes in graduate school, and for their guidance, I am immensely grateful. I am also thankful to the undergraduate researchers with whom I have had the honor of collaborating. A special thank you to Julia Nadel for her genuine curiosity and insight and for the many stimulating conversations we have had about this dissertation research.

Beyond Emory, there are several people whose support and guidance I could not have gone without. At Columbia, I am grateful to Bridgid Finn, Lisa Son, and Lexi Suppes for having introduced me to Psychology. A very special thank you to Lexi for helping me navigate my first introduction to Research Methods, my first attempt at creating an experimental paradigm, and my first experience publishing a peer-reviewed article. A deeply heartfelt thank you to Monnica Chan, Vanessa Chow, Jared DeFife, Melanie Hammet, Chloe Liang, and Amy Tan for sharing laughs and tears and for being sources of mental and emotional strength when I needed it most.

Lastly, I would not be where I am today without the support of my family. A sincerest thank you to my husband, Frank Wang, for never wavering in his belief that everything in the end, really, will be okay. In even the most tiring and frustrating of times, Frank has always been able to make me smile. I thank my brother, Christopher Tzeng, for encouraging me to always question and to speak my mind. I am forever indebted to my father, Wen-Yann Tzeng, and my mother, Hsiu-Fang Chang, who instilled in me a life-long appreciation for learning and intellectual growth. For them to have prioritized my and my brother's education over their own quality of living is a sacrifice that I will never forget. I dedicate this dissertation to my father and to the memory of my mother, with whom I would have loved to share this journey.

Table of Contents

List of Tables

Table

List of Figures

Figure

Introduction

Understanding spoken language requires listeners to process not only *what* a speaker says but also *how* the speaker says it. To achieve this, listeners integrate multiple sources of information, including acoustic, phonological, morphological, syntactic, and semantic cues. One informative element of the speech signal is *prosody,* or the timing, rhythm, and intonation of a linguistic utterance. Often described as the suprasegmental properties of spoken language (Cutler, Dahan, & van Donselaar, 1997; Lehiste, 1970), prosodic cues have been found to convey information about syntactic structure (e.g., Clifton, Carlson, & Frazier, 2002) and speakers' emotional state (e.g., Banse & Scherer, 1996) and are essential for effective communication.

Although traditional characterizations of prosody assume that prosodic cues do not convey semantic information about linguistic reference, 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) from these cues. These findings imply that prosody provides information beyond syntactic structure and speakers' affective state and augments meaning that is conveyed through the accompanying linguistic content. However, both the communicative conditions under which speakers recruit prosody to convey referential detail, as well as the cognitive mechanisms underlying such use of prosody, remain unclear. In this dissertation, I examine these two issues. In particular, I explore the extent to which language users produce and infer referential information from prosodic correlates to perceptual features in a specific visual modality, color brightness. Given evidence for consistent auditory-visual correspondences in multiple dimensions, including loudness and size (Smith & Sera, 1992), pitch and shape (Marks, 1987), and pitch and brightness (Marks, 1974; Melara, 1989; Mondloch & Maurer, 2004), I examine the extent to which the relation between prosody and linguistic reference exemplifies a more general level of cross-modal correspondence that affects perceptual processing.

In the pages that follow, I first describe traditional conceptualizations of the functional significance of prosody in spoken language. Next, I discuss how research has challenged these characterizations, implying a more direct relationship between prosody and referential processing than previously acknowledged. Given evidence for the use of prosody to convey referential information, I then discuss the communicative circumstances under which prosody may be recruited in this manner. Finally, I present four empirical studies assessing the possibility that spoken language capitalizes on general-cross-modal correspondences to express referential information. My ultimate conclusion is that prosody can be considered a type of vocal gesture, as it provides referential detail about objects and events in the world and can be recruited to resolve ambiguity in the accompanying linguistic content.

Functional Significance of Prosody: Traditional Characterizations

Research on prosody in spoken language has focused on two primary topics: the role of prosodic cues in conveying speakers' affective states and the role of prosody in disambiguating linguistic structure.

Prosody and speaker emotion. Vocal expression of emotion is characterized by distinct acoustic properties, such as the level and range of fundamental frequency (F_0, perceived as pitch) and amplitude (perceived as intensity), that correlate with specific affective states (Banse & Scherer, 1996; Laukka, Juslin, & Bresin, 2005; Leinonen, Hiltunen, Linnankoski, & Laakso, 1997; Scherer, 1994). Among the most consistent associations reported are those between arousal and F_0 and amplitude (Mauss & Robinson, 2009). Banse and Scherer (1996) recorded nonsense sentences spoken by actors displaying different vocally expressed emotions. In comparison with neutral speech, sentences spoken with fear, joy, and anger exhibited higher mean $F_{0,}$ F_0 variability, and amplitude, whereas portrayals of sadness exhibited lower mean $F_{0,}$ F_0 variability, and amplitude. Further, expressions of boredom and sadness were spoken with slower speech rates than expressions of happiness and anxiety. Additional findings demonstrate similar acoustic correlates to vocal emotion in natural speech (Bachorowski & Owren, 1995).

Individuals reliably identify speaker emotion from these prosodic cues (Bachorowski, 1999; Banse & Scherer, 1996; Greasley, Sherrard, &Waterman, 2000; Laukka et al., 2005). Listeners in Greasley et al. (2000), for example, heard television and radio recordings displaying unscripted emotions and then labeled the emotions displayed in the recording. Although how reliably listeners labeled the emotions varied depending on which emotion was depicted, listeners' choices reflected overall reliable categorization of emotions. Evidence for cross-cultural consistency in categorizing vocal expressions of emotion (Laukka et al., 2005; Pell, Monetta, Paulmann, & Kotz, 2009; Thompson & Balkwill, 2006) further suggests that listeners associate particular patterns of prosodic cues with distinct emotional states. Children, too, detect the emotional content of speech from prosodic patterns (Fernald, 1992; Morton & Trehub, 2001). Taken together, the above findings suggest not only that distinct prosodic characteristics underlie specific vocal expressions of emotion but also that listeners can accurately infer speaker emotion from prosodic cues.

Prosody as a cue to linguistic structure. Prosody also provides information that aids in speech segmentation, syntactic disambiguation, and lexical access. To understand a linguistic utterance, listeners must first parse the continuous speech signal into discrete units. Using prosodic cues, such as patterns in lexical stress, pitch contour, and pause location, both adults and children can segment linguistic utterances at the syllable, word, phrase, and sentence levels (for reviews, see Cutler et al., 1997; Shattuck-Hufnagel, & Turk, 1996; Wagner & Watson, 2012).

Cutler and Norris (1988), for example, proposed that listeners capitalize on distributional cues in lexical stress patterns to aid word segmentation. Lexical stress is indicated through a combination of higher pitch and amplitude, as well as vowel lengthening (Kempe, Schaeffler, & Thoresen, 2010). Stress languages, such as English, exhibit a speech rhythm that is expressed in the juxtaposition of strong and weak syllables. Given that most multi-syllabic words in English have stress on their first syllables (Cutler & Carter, 1987), listeners can use this pattern of cues to form hypotheses about the location of word boundaries. Cutler and Butterfield (1992) examined patterns in listeners' missegmentations of continuous speech. Native English listeners tended to insert boundaries before strong syllables (e.g., *analogy* perceived as *an allergy*) but delete boundaries before weak ones (e.g., *my gorge is* perceived as *my gorgeous)* – a segmentation pattern consistent with listeners' understanding of the distributional stress cues in English.

Infants, too, are sensitive to the typical prosodic contours of words and use these cues to parse the speech stream (Jusczyk & Aslin, 1995; Jusczyk, Cutler, & Redanz, 1993; Thiessen, Hill, & Saffran, 2005). Juscyzk et al. (1993), for example, found that nine-month-old American infants preferred to listen to lists of strong-weak (e.g., *DOnor*) versus weak-strong (e.g., *conDONE*) English words, implying that young infants are sensitive to the dominant stress patterns in their native language. Moreover, prosodic cues that disambiguate word boundaries are more exaggerated and reliable in child-directed speech (CDS) than in adult-directed speech (ADS) (Fernald & Kuhl, 1987; Kempe et al., 2010; Singh, Morgan, & Best, 2002), implying that caregivers recruit exaggerated prosody in their speech to increase the salience of relevant cues for speech segmentation (Thiessen et al., 2005). Taken together, these findings suggest that listeners detect prosodic contours and use them to segment the speech signal into meaningful units.

Prosodic cues also facilitate the resolution of syntactic ambiguities (Clifton et al., 2002; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Watson & Gibson, 2004). To disambiguate between two candidate interpretations of a syntactically ambiguous sentence, speakers insert pauses and lexical stress at different locations. In the sentence *Touch the cat with the spoon*, for example, the phrase *with the spoon* can be interpreted either as an instrument of the action (e.g., *Touch the cat… WITH the spoon*) or as a descriptor of the direct object (e.g., *Touch… the cat with the spoon*). Indeed, listeners' interpretations of linguistic utterances vary as a function of these prosodic cues to syntactic structure (Clifton et al., 2002; Kempe et al., 2010; Snedeker & Trueswell, 2003; Wagner & Watson, 2012).

The studies reviewed thus far demonstrate that prosodic cues serve crucial functions at multiple stages of spoken language processing. Listeners capitalize on lexical stress, pitch contour, and pause location and duration to infer speaker affective state, segment speech, and disambiguate syntax. These traditional characterizations of prosody, however, imply that the suprasegmental elements of speech are *independent* from meaning. In particular, prosody as conceptualized above does not convey any information regarding objects and events in the world. Although prosody facilitates the processing of spoken language, it does not provide semantic-referential information.

Encoding Prosody in Lexical Representations

An alternative characterization of prosody implies that suprasegmental elements of speech impact *lexical* processing. Prosodic cues have been found to facilitate word recognition, lexical access, and disambiguation of meaning, suggesting that prosody and semantics may be less independent than previously acknowledged. Implied in this characterization of prosody is the assumption that prosodic cues, such as speaking rate, intonation contour, and F_0, are encoded in listeners' representations of spoken utterances (e.g., Nygaard, Herold, & Namy, 2009; Shintel & Nusbaum, 2007). Indeed, prosodic cues have been found to affect listeners' memory for spoken words (Bradlow, Nygaard, & Pisoni, 1999; Church & Schacter, 1994; Nygaard, Burt, & Queen, 2000; Palmeri, Goldinger, & Pisoni, 1993). Nygaard, Burt, and Queen (2000), for example, presented listeners with lists of spoken English words that varied in speech rate, amplitude, and vocal effort. During a recognition memory task, listeners recognized words that were repeated with the same speaking rate and vocal effort more accurately than words that exhibited different prosodic characteristics at study and test.

Memory for the acoustic features of a particular speaker's voice also facilitates word recognition. Often termed *indexical cues*, these features underlie perceptual differences among speakers' voices. Goldinger (1996) examined the extent to which listeners retain indexical cues in linguistic representations of spoken words. During a familiarization phase, listeners heard monosyllables spoken by different speakers. Listeners completed either a recognition memory test or a perceptual identification task in which they reported words presented in noise. Words presented at test were spoken either by the same or a different speaker from that heard during familiarization. Listeners correctly recognized words more accurately at test when they heard words presented in the same rather than a different voice. These and similar findings (Bradlow et al., 1999; Church & Schacter, 1994; Palmeri et al., 1993) suggest that listeners retain highly detailed acoustic information in their memory representations of spoken words. Congruency between the prosodic features of a spoken utterance and the memory representations of similar, previously heard utterances thus facilitates the word recognition process.

Prosodic Cues to Meaning

Although the facilitative effect of prosodic cues on word recognition implies that detailed acoustic information is encoded and retained in linguistic representations, the extent to which this impacts semantic processing remains unclear. Many findings support listeners' recruitment of prosodic cues to segment speech and disambiguate syntactic ambiguity. Despite their role in important stages of speech comprehension, prosodic cues have been widely regarded as separate from the *meaning* of a linguistic utterance. A growing literature suggests, however, that the relation between prosodic cues and propositional content is more substantial than previously acknowledged.

Prosody and semantic disambiguation. Nygaard and Lunders (2002), for example, found that emotional prosody facilitated listeners' disambiguation of homophones. Listeners heard and transcribed emotional homophones that differed in the emotional valence of each meaning (positive/negative vs. neutral, e,g., *die* vs. *dye*). Speakers recorded these words with happy, sad, or neutral tone of voice such that the prosody matched or mismatched word valence. Listeners reported the emotional meaning more often than the neutral one when the prosody was congruent rather than incongruent with word meaning, suggesting that emotional tone of voice biases lexical access. Nygaard and Queen (2008) showed that the integration of linguistic and prosodic cues occurs early in lexical processing. Listeners heard happy, sad, or neutral words spoken with congruent, incongruent, or neutral prosody. In a naming task during which listeners repeated the words they heard, listeners' response latencies differed as a function of congruency between prosody and word valence such that listeners responded more quickly when prosody and semantic content matched than when they mismatched. Taken together, these and similar findings (Schirmer, Kotz, & Frederici, 2002; Wurm, Vakoch, Strasser, Calin-Jageman, & Ross, 2001) suggest not only that prosodic cues are integrated in lexical representations of spoken words but also that these cues influence lexical access and selection.

Beyond emotional meaning. The findings discussed above suggest that listeners integrate prosodic cues to emotion with emotional word meaning. The extent to which prosodic cues convey meaning for semantic domains *beyond* emotion, however, has been less extensively explored. Nygaard et al. (2009) found that speakers produced reliable prosodic cues to the meaning of novel adjectives (e.g., *daxen,* meaning big). For example, speakers produced novel adjectives intended to mean big with greater amplitude, slower speaking rate, and lower pitch than when the same words were intended to mean small. Although valence partially accounted for the variance in the acoustic features of speakers' utterances, speakers' productions also varied systematically across word meanings. Speakers produced utterances with unique profiles of pitch, pitch variation, amplitude, and duration for each word meaning, suggesting that speakers recruit prosodic features in specific ways to convey semantic content.

Both adult and child listeners are sensitive to these differences in prosody and use them to disambiguate word meaning (Herold, Nygaard, Chicos, & Namy, 2011; Kunihira, 1971; Nygaard et al., 2009). Nygaard et al. (2009) demonstrated that adult listeners use prosodic cues to identify the correct meaning of spoken novel adjectives. Choosing between two pictorial representations of an antonym pair (e.g., a big and small flower for *big* and *small*), listeners selected the correct picture more often when the prosodic cues matched one of the pictures (e.g., big-small picture pair upon hearing a novel word meaning big) than when the prosodic contour was drawn from a mismatched semantic dimension (e.g., hot-cold picture pair upon hearing a novel word meaning big). This finding suggests that listeners recruited the unique acoustic profiles associated with each word meaning to infer semantic reference. Importantly, listeners made consistent prosody-meaning mappings for words outside of the emotion domain, suggesting that the semantic information conveyed in prosody extends beyond affective connotation or valence.

Prosody and word learning. Prosodic cues to word meaning also facilitate word learning for both adults and children. In an exposure-test paradigm, listeners in Reinisch et al. (2012) heard sentences containing novel adjectives spoken with prosody intended to convey a particular meaning (e.g., *seebow* spoken with long duration and low pitch to mean big). Picture pairs representing the meaning of the adjective and its antonym accompanied the spoken sentences. At test, listeners heard the words spoken with neutral prosody and selected referents from familiar and unfamiliar picture pairs. Despite the lack of informative prosody to guide word-referent mappings at test, listeners selected the correct pictorial representation for both familiar and unfamiliar picture pairs, demonstrating that prosody guides the learning of word-meaning mappings. Further, listeners' ability map novel words onto *unfamiliar* picture referents suggests that prosody promotes the learning of word-concept mappings rather than specific word-referent pairs.

Children, too, capitalize on prosodic cues to infer word meaning. Herold et al. (2011), for example, presented four- and five-year-old children with picture pairs that varied along a single dimension (e.g., big vs. small flower) and asked them "Can you get the *blicket* one?" in meaningful or neutral prosody. Children in both age groups reliably selected the correct meaning of novel adjectives when speakers produced words in meaningful CDS but responded randomly for neutral trials. Although four-year-olds required additional prompting to attend to prosody, these findings suggest that prosodic cues constrain word meaning for young language learners. That the novel adjectives did not refer to emotion suggests that prosody conveys information beyond word valence. Taken together with evidence that mothers spontaneously employ meaningful prosody when reading storybooks to their children (Herold, Nygaard, & Namy, 2011), these findings suggest that prosodic cues facilitate children's learning of word-meaning correspondences.

Prosody as a source of referential information. One potential contributing factor to the influence of prosody on word disambiguation and learning is that prosodic cues provide information regarding the perceptual features of word referents (Nygaard et al., 2009; Perlman, 2014; Shintel, Nusbaum, & Okrent, 2006; Shintel & Nusbaum, 2007, 2008). Shintel and colleagues, for example, argue that prosody is a form of *analog acoustic expression* such that the suprasegmental elements of speech convey information about objects and events that listeners encode along with linguistic input. Critically, such a characterization of prosody implies that prosody, rather than serving exclusively as a source of affective and syntactic information, is also a source of semantic reference.

Shintel et al. (2006), for example, found that speakers use prosody to convey visuo-spatial properties of word referents. Participants described the direction of an animated dot by saying "It is going up" or "It is going down." Descriptions varied in prosody as a function of direction of movement such that descriptions of upward moving dots were higher pitched than those of downward moving dots. Here, the prosody provided complementary information to the propositional content of the description. In their descriptions of horizontal trajectories of moving dots (e.g., “It is going left.”), speakers modulated their speech rate such that they described fast-moving dots with a faster speaking rate than they did with slow-moving dots. The prosodic variation in these descriptions conveyed information *independent* of the linguistic content (the left or right movement of the dot), suggesting that prosody provides yet another channel of information beyond the propositional content of an utterance. Prosodic differences were salient to listeners, who used them to resolve referential ambiguity. In a two-alternative forced-choice task, listeners reliably chose the correct corresponding dot display based on the speech rate of the descriptions. Together, these results suggest that prosody directly expresses information describing the visuo-spatial properties of external referents.

Neural Mechanisms Underlying Non-Arbitrary Sound-Meaning Mappings

In studies investigating cross-modal correspondences in non-linguistic stimuli, listeners consistently associate certain auditory and visual dimensions, such as loudness and size (e.g., Smith & Sera, 1992), pitch and brightness (e.g., Mondloch & Maurer, 2004), and pitch and shape (e.g., Marks, 1987). Individuals have also mapped judgments of higher pitch to brighter versus darker colors (Marks, Hammeal, & Bornstein, 1987) and higher-pitched pure tones to brighter versus darker shades of gray (Marks, 1974). Given listeners' consistent mappings between these dimensions, acoustic properties in speech may be readily associated with particular visual characteristics of the described object or event during speech comprehension and provide informative referential detail (Perlman, 2014; Shintel et al., 2006; Shintel & Nusbaum, 2007).

Evidence for systematic cross-modal correspondences and for speakers' use of prosody to convey and infer visuo-spatial properties of linguistic referents suggest that prosodic cues in spoken language *recruit* systematic cross-modal mappings. Ramachandran and Hubbard (2001) have suggested that non-arbitrary auditory-visual correspondences may have bootstrapped language evolution. According to this perspective, the neural underpinnings of the acoustic and articulatory properties of linguistic sounds may share cortical connections with other sensory modalities such that spoken language activates neural regions associated with the perceptual properties of the referent. Evidence from studies examining the neural basis for sensitivity to sound symbolism – non-arbitrary mappings between sound and meaning in natural language – provides support for this view (Kovic, Plunkett, & Westermann, 2010; Revill, Namy, DeFife, & Nygaard, 2014). For example, Revill et al. (2014) found that relative to non-sound symbolic foreign words, sound symbolic words elicited increased activation in the left superior parietal cortex, an area associated with multi-sensory integration. Systematic cross-modal correspondences between prosody and perceptual properties of linguistic referents (e.g., pitch and visuo-spatial height) may potentially capitalize on these neural bases.

Evidence for systematic auditory-visual mappings in spoken language also aligns with theories that assume language is grounded in multi-modal experiences (Barsalou, 1999, 2003; Glenberg & Kaschak, 2002; Zwaan, Madden, Yaxley, & Aveyard, 2004). Grounded theories of language claim that linguistic symbols elicit simulations of perceptual experiences associated with external referents (Barsalou, 1999, 2003). Representations of these perceptual experiences are reactivated during linguistic processing and facilitate the comprehension of spoken language (Matlock, 2004; Šetić & Domijan, 2007; Zwaan, Stanfield, & Yaxley, 2002). Prosodic correlates to visual features may be an instantiation of the grounding of language in such multi-modal perceptual experiences.

Effects of Communicative Demand on Prosody Use

To the extent that language users recruit prosody to resolve referential ambiguity, use of prosodic cues to reference may vary 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. 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 (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).

Similar patterns have been found in speaker’s use of prosodic contours to highlight contrastive information in referential communication. In interactive (Speer & Ito, 2011) and visual search (Weber, Braun, & Crocker, 2006) tasks, speakers have been found to use 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.

Sources of variation in referential prosody perception and production. The extent to which language users employ and infer information from referential prosody may also vary as a function of differences in personality traits or personal history. Evidence for this possibility can again be drawn from the co-speech gesture literature such that the extent to which speakers produce gestures has been found to vary as a function of verbal and spatial skills (Hostetter & Alibali, 2007) and personality characteristics, such as extraversion levels (Hostetter & Potthoff, 2012). Given evidence that pitch discrimination can vary as a function of musicality (Tervaniemi, Just, Koelsch, Widmann, & Schröger, 2005), the extent to which language users employ and infer information from referential prosody may also vary as a function of differences in speaker and listener characteristics.

Overview of the Dissertation

The association between pitch and brightness level provides unique insight into the extent to which referential prosody is an instantiation of general, systematic cross-modal mappings. To date, evidence for such cross-modal correspondences manifesting in spoken language has been found only in the pitch-visuo-spatial height domain. As one of the most consistently demonstrated mappings in studies using non-linguistic stimuli, the pitch-brightness association provides a reliable foundation for examining the possibility that spoken language recruits traditionally non-linguistic cross-modal associations. By recruiting these non-linguistic mappings in linguistic communication, speakers can extend both the range and efficiency of conveying information in speech beyond what is afforded by propositional content alone.

The current study assesses the extent to which systematic auditory-visual correspondences will manifest in *linguistic* utterances to convey perceptual details that supplement the accompanying linguistic content. Although previous work has shown that listeners systematically associate particular auditory or acoustic properties with brightness levels, whether speakers spontaneously recruit these mappings in their productions of spoken words has yet to be explored. In four experiments, this dissertation assesses the extent to which systematic auditory-visual correspondences manifest in prosody to convey perceptual details (Experiments 1 and 2), the extent to which listeners infer referential information from these prosodic cues (Experiment 3), and the extent to which communicative context modulates speakers' and listeners' use of prosody to resolve referential ambiguity (Experiment 4). Results from these experiments will provide evidence to more accurately define the role of prosody in spoken language as a source of referential detail.

Experiment 1

The objective of Experiment 1 was to examine the extent to which speakers spontaneously produce prosodic correlates to color brightness. Participants completed a production task designed to elicit verbal labels for different brightness levels of six colors. Given previous demonstrations of non-arbitrary pitch-brightness mappings (e.g., Marks, 1974), it was predicted that when referring to brighter shades of a particular color, speakers would produce higher-pitched labels. When referring to darker shades, participants would produce the same color labels using a lower pitch. This pattern of findings would suggest that prosody may recruit cross-modal mappings to provide referential information in spoken language in a manner that has largely been overlooked.

Method

Participants. Thirty-four (22 female, 12 male) Emory University undergraduates received course credit for their participation. Participants were native English speakers and reported no history of speech or hearing disorders[1].

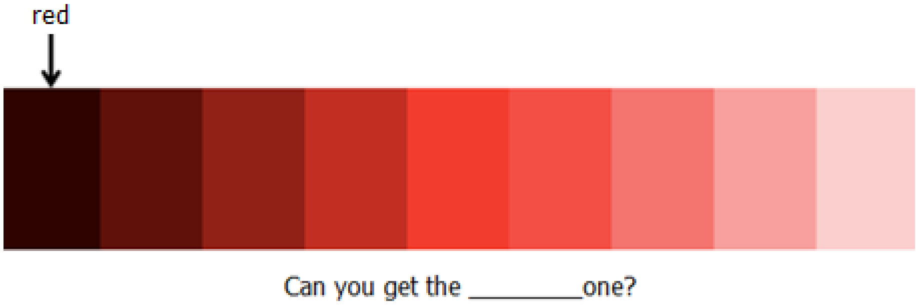
Stimuli. Stimuli consisted of six color spectra (red, orange, yellow, green, blue, purple), each of which included nine shades of a color presented simultaneously in a horizontal orientation such that the shades progressively increased in perceived brightness (amount of white) from left to right. Each shade was created by using red, green, and blue (RGB) coordinates[2] (see Appendix A for spectra and RGB coordinates for each shade). Each color spectrum was 9 x 1.85 inches, with the width and height of each shade equated.

Procedure. In a computerized task, participants viewed a single color spectrum per trial. At the top of each spectrum was an arrow indicating either the brightest, darkest, or the center (intermediate-most) shade and an English color label that corresponded to

[1] Participants were not screened for color blindness in Experiments 1-3. However, before the start of the experiment, participants completed a practice trial during which they were told to label the darkest, intermediate, or brightest shade of a gray color spectrum with either an English color label (Experiment 1) or novel word (Experiment 2). The experimenter also explained that all presented spectra would include shades that progressively increased in brightness from left to right. These steps ensured that participants were aware of differences in brightness among shades.

[2] Levels of blue in the sixth and seventh shades of the yellow spectrum (RGB: 225, 255, 80 and 255, 255, 112) and the fifth and ninth shades of the purple spectrum (127, 0, 225 and 229. 204, 225) were manually adjusted to reduce perceived similarity between consecutive shades.

the spectrum color (e.g., red; see Fig. 1). Each spectrum remained on the computer screen for 5s during which participants produced the sentence “Can you get the ______ one?,” filling the blank with the indicated color label (e.g., “Can you get the red one?”). Participants were not explicitly told to vary their prosody but were instructed to describe the shade of the color as best as they could to an imaginary listener using only the specified color label and target sentence. Implicit in the provided instructions was that participants would not be able to *lexically* disambiguate between levels of brightness. Participants were told that the task objective was to ensure that the imaginary listener had enough information to choose the intended referent. Prior to the start of the task, participants completed a practice trial during which they were told to label the darkest, intermediate, or brightest shade of a gray color spectrum. The experiment consisted of three blocks. In each block of trials, participants saw each spectrum three times, once when asked to label the brightest shade, once the darkest, and once the middle-most shade. Trials within each block were randomized.



Can you get the ________one?

Figure 1. Screenshot of a single trial in Experiment 1

Participants’ utterances were recorded in a sound-attenuated room using an audio-technica ATR 20 microphone onto a Dell computer and segmented by trial using E-prime 2.0 (Schneider, Eschman, & Zuccolotto, 2002). Sentence utterances were re-digitized at a 22.050 kHz sampling rate and amplitude normalized using PRAAT (Boersma & Weenink, 2012). To examine the acoustic features of the individual color labels in addition to the sentence-length utterances, each color label was segmented from the sentence utterance and amplitude normalized[3].

Acoustic analyses. Given that acoustic correlates to brightness might not be confined to speakers' production of the color labels but may extend across the sentence, acoustic measures were obtained for both the sentence-length utterances and separately for the segmented color labels in isolation. Because our recordings of natural speech necessarily vary in acoustic properties in addition to pitch, we also examined the extent to which utterance duration and amplitude varied as a function of brightness. Mean fundamental frequency (F_0), mean amplitude, and duration were measured using PRAAT. 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] Amplitude normalization preserves relative amplitude differences within an utterance by adjusting the amplitude throughout the file by the same amount.

Results and Discussion

Table 1 shows the F_0, amplitude, and duration of sentence-length utterances for bright, dark, and intermediate shades. Because results patterned similarly for sentence- and word-level analyses, only the sentence-level analyses are reported below. Because the acoustic profiles of the six color labels differed in number of syllables and in phonetic composition, means were collapsed across the six colors. Data from one participant was excluded due to failure to use the provided carrier phrase. To assess the extent to which participants' prosody varied as a function of color brightness, three separate (one for each dependent measure) two-way repeated-measures analyses of variance (ANOVA) were conducted with Brightness (Bright, Dark, Intermediate) and Block (1, 2, 3) as within-subjects variables. Results did not indicate a significant effect of brightness for any of the three dependent measures, F_0, *F*(2, 64) = .75, *p* = .479, *partial* η^2 = .02; amplitude, *F* (2, 64) = .038, *p* = .963, *partial* η^2 = .001; duration, *F*(2, 64) = 1.04, *p* = .360, *partial* η^2 = .03. For duration, the main effect of block was significant, *F*(2, 64) = 27.80, *p* < .001, *partial* η^2 = .47, with reliable decreases in sentence duration across Blocks 1 and 2, (M_1 = 1327.94, SD_1 = 190.10; M_2 = 1254.55, SD_2 = 200.14; *t*(32) = -5.31 *p* < .001), Blocks 2 and 3, (M_2 = 1254.55, SD_2 = 200.14; M_3 = 1233.99, SD_3 = 199.95; *t*(32) = -2.63, *p* = .013), and Blocks 1 and 3, (M_1 = 1327.94, SD_1 = 190.10; M_3 = 1233.99, SD_3 = 199.95; *t*(32) = -5.67 *p* < .001), suggesting that speakers were becoming more familiarized with the task across blocks. All other main effects and interactions were not significant.

Table 1

Acoustic measurements of sentence-length utterances in Experiment 1

	Bright	**Intermediate**	**Dark**
F0 (Hz)	165.66	164.96	164.89
Amplitude (dB)	75.23	75.25	75.26
Duration (ms)	1269.15	1268.33	1279.02

Results suggest that participants did not systematically vary their prosody as a function of color brightness. This null result may imply that speakers do not spontaneously employ prosodic cues to signal visual characteristics such as brightness. However, an alternative possibility is that the presence and intensity of prosodic cues to reference vary as a function of the communicative need for referential specificity. Perhaps prosodic cues are more likely to be recruited when linguistic content is underspecified or ambiguous. There is evidence consistent with this notion in the literature on co-speech gesture – speakers tend to produce gestures to clarify linguistic ambiguity (Holler & Beattie, 2003; McNeill, 1992). Using a story-telling task, Holler and Beattie (2003), for example, 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. It may also be the case that for highly conventionalized communication, such as color labeling, speakers have acquired a relatively fixed prosodic contour that is associated with labeling prototypical shades of color (e.g., “red” corresponding to a prototypical shade of red) and that automatically accompanies production of the conventional color label word forms. This perhaps reduces the extent to which prosody is likely to vary as a function of brightness of the shade depicted in this task.

Experiment 2

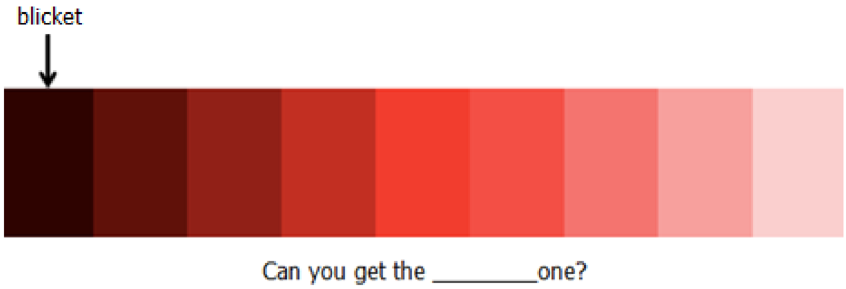
In Experiment 1, speakers were instructed to use familiar color labels to refer to brightness levels. These color labels were conventionalized ones, potentially lessening the likelihood that speakers would use prosody to disambiguate brightness. The objective of Experiment 2 was to assess the possibility that prosodic correlates to referential information about visual characteristics are more likely to occur when (1) they accompany underspecified linguistic content and (2) prosodic conventions associated with particular lexical items are eliminated. In Experiment 2, participants verbally labeled brightness levels using novel words rather than English color labels. It was predicted that speakers would produce labels using higher pitch for brighter relative to darker shades.

Method

Participants. Forty-eight (38 female, 10 male) Emory University undergraduates received course credit for their participation. Participants were native English speakers and reported no history of speech or hearing disorders.

Stimuli. The same color spectra from Experiment 1 were used. However, the color labels elicited from participants were changed to bi-syllabic novel words (*blicket, daxen, foppick, riffel, seebow, tillen*).

Procedure. Participants completed the same production task as in Experiment 1 during which they provided spoken labels for the brightest, darkest, and an intermediate shade of each of the six colors (Fig. 2). Color labels were rotated across conditions such that which novel word was paired with which color varied across participants. All other aspects of the procedure remained the same as in Experiment 1.



Can you get the ________one?

Figure 2. Screenshot of a single trial in Experiment 2

Results and Discussion

Fig. 3a-c show the mean F_0, amplitude, and duration, respectively, of sentence-length responses for bright, dark, and intermediate shades. Because results were similar at the sentence- and word-level, only sentence-level analyses are reported below. Cases where the word-level results deviated from those at the sentence-level are noted when they occurred. To assess the extent to which participants' prosody varied as a function of color brightness, three two-way repeated-measures ANOVAs (one for each acoustic measure) were conducted with Brightness (Bright, Dark, Intermediate) and Block (1, 2, 3) as within-subjects variables. Results indicated a significant effect of brightness for all three dependent measures, as reported below.

Figure 3a

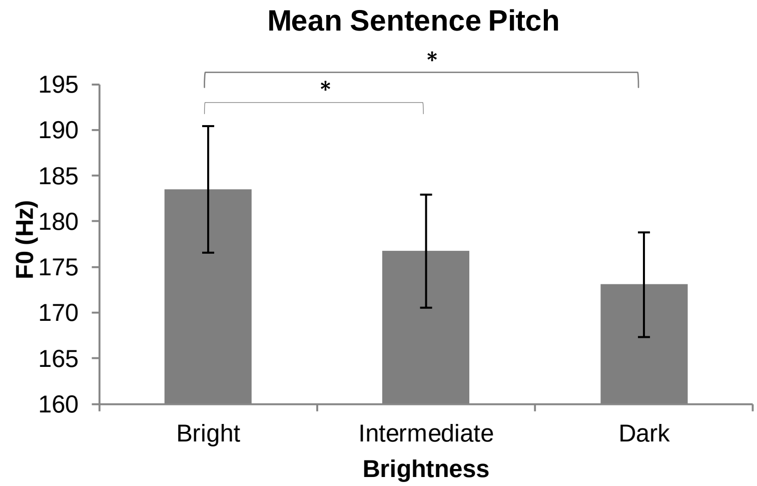


Figure 3b

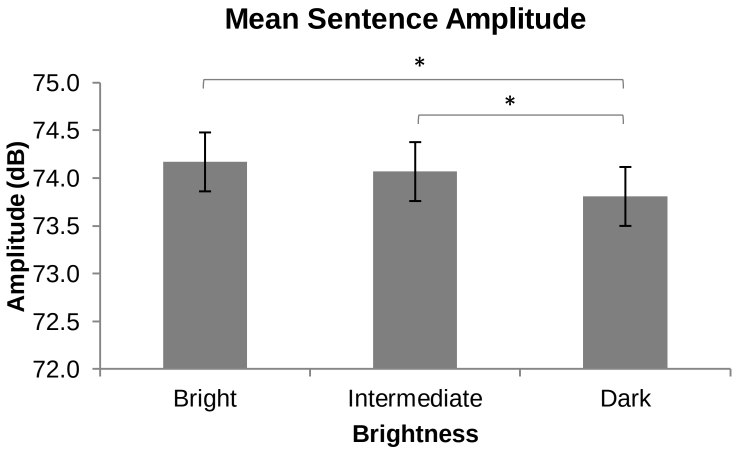
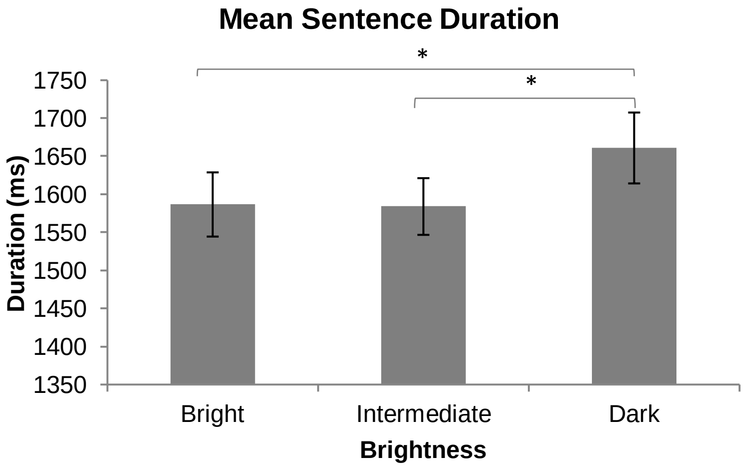


Figure 3c



Figures 3a-c. Mean F_0 (Fig. 3a), amplitude (Fig. 3b), and duration (Fig. 3c) for sentence-length utterances in Experiment 2. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

F_0. Participants' utterances differed reliably in F_0 across brightness levels, *F*(2, 94) = 7.52, *p* = .001, *partial* η^2 = .14. Pairwise comparisons indicated that sentences for bright shades (*M* = 183.58, *SD* = 48.09) were significantly higher-pitched than those for dark shades (*M* = 173.05, *SD* = 39.68; *t*(47) = -2.88, *p* = .006). Sentences for the intermediate shades (*M* = 176.76, *SD* = 42.89) were reliably lower-pitched than those for bright shades (*M* = 183.58, *SD* = 48.09; *t*(47) = 2.96, *p* = .005) but did not differ from labels for dark shades (*M* = 173.05, *SD* = 39.68). A significant main effect of Block observed only for words, *F*(2, 94) = 3.96, *p* = .022, *partial* η^2 = .08, indicated that utterances produced in Block 1(*M* = 169.65, *SD* = 38.97) were reliably lower pitched than those produced in Block 2 (*M* = 172.94, *SD* = 37.93; *t*(47) = -2.05, *p* = .046) and Block 3 (*M* = 174.24, *SD* = 38.94; *t*(47) = -2.41, *p* = .020). There was no significant interaction.

Amplitude. Sentences varied reliably in amplitude across brightness levels, *F*(2, 94) = 6.64, *p* = .002, *partial* η^2 = .12. Dark sentences (*M* = 73.81, *SD* = 2.14) were significantly lower in amplitude than bright (*M* = 74.17, *SD* = 2.14; *t*(47) = -2.80, *p* = .007) and intermediate (*M* = 74.07, *SD* = 2.14; *t*(47) = -2.76, *p* = .008) sentences. Bright and intermediate sentences did not differ reliably. There was no significant effect of block and no significant interaction.

Duration. Sentences differed reliably in duration across brightness levels, *F*(2, 94) = 5.55, *p* = .005, *partial* η^2 = .11. Pairwise comparisons indicated that dark sentences

(M = 1660.97, SD = 322.60) were significantly longer in duration than those for bright (M = 1586.75, SD = 292.03; t(47) = 2.17, p = .035) and intermediate (M = 1584.06, SD = 258.04; t(47) = 3.20, p = .002) shades. Bright and intermediate sentences did not differ reliably. For sentences, there was also a significant main effect of block, F(2, 94) = 25.99, p < .001, *partial* η^2 = .36, such that utterances decreased in duration across blocks as participants became more familiarized with the task (1 vs. 2, M_1 = 1682.57, SD_1 = 308.80, M_2 = 1624.30, SD_2 = 308.22, t(47) = 3.04, p = .004; 2 vs. 3, M_2 = 1624.30, SD_2 = 308.22, M_3 = 1524.44, SD_3 = 237.37, t(47) = 4.81, p < .001; 1 vs. 3, M_1 = 1682.57, SD_1 = 308.80, M_3 = 1524.44, SD_3 = 237.37, t(47) = 6.00, p < .001). There was no significant interaction.

Results suggest that participants varied their prosody across brightness levels such that labels for bright shades were reliably higher-pitched, higher in amplitude, and shorter in duration than those for darker shades. Because participants were not instructed to vary their tone of voice to refer to different brightness levels, these results demonstrate that speakers spontaneously and consistently use prosody to convey perceptual information that is unavailable in the linguistic content of spoken utterances. It is possible that demand characteristics encouraged prosodic modification; however, if so, there would be no reason to expect that participants would employ the same acoustic properties in the speech signal in the same manner. These findings, therefore, suggest that speakers were invoking a shared systematic correspondence between auditory (speech) and visual (brightness) perceptual information.

That speakers' prosody varied as a function of brightness is consistent with previous findings demonstrating the use of prosody to express visuo-spatial properties of external referents (Herold et al., 2011; Nygaard et al., 2009; Perlman, 2014; Shintel et al., 2006; Shintel & Nusbaum, 2007). Evidence from the current work suggests that prosody serves as an additional channel of referential detail that supplements accompanying propositional content. In this sense, prosody can be conceptualized as analogous to co-speech gesture (Perlman, 2010; 2014). Language users produce gestures that convey information that is non-redundant with the accompanying speech (for a review, see Hostetter, 2011) and remember this supplementary information when viewing others' co-speech gestures (Beattie & Shovelton, 1999; Broaders & Goldin-Meadow, 2010). Beattie and Shovelton (1999), for example, found that listeners who viewed videos of speakers' cartoon narrations remembered more details regarding the relative position and size of objects described than listeners who only heard the narrations, suggesting listeners encoded the non-redundant information provided in speakers' gestures. That speakers' prosody varied as a function of brightness in Experiment 2 but not Experiment 1 is consistent with findings in the co-speech gesture literature suggesting that speakers produce extra-linguistic cues to clarify ambiguity in the accompanying linguistic content (Holler & Beattie, 2003; McNeill, 1992). Rather than provide information that is redundant with the accompanying speech, prosody, like gesture, may offer referential information that is supplementary or non-redundant.

Experiment 3

Given evidence for the production of consistent prosodic correlates to brightness in Experiment 2, Experiment 3 examined to what extent listeners would use such cues to infer referential information. Findings from Shintel and Nusbaum (2007) provide one demonstration of listeners' ability to infer referential detail from prosodic cues. Listeners heard sentences describing objects (e.g., "The horse is brown.") spoken at fast or slow speaking rates. After hearing each sentence, listeners viewed a picture of an object and reported whether it had been mentioned in the previous sentence. Listeners recognized the object more quickly when the motion implied by the speaking rate matched the motion implied by the picture (e.g., a horse standing still versus running) than when they mismatched. Prosody thus conveys visuo-spatial information that the listeners integrate into their representations of the referent. In line with these findings, it was predicted that listeners in the current study would map higher-pitched color labels to brighter shades and lower-pitched labels to darker shades.

Method

Participants. Thirty-eight (27 female, 11 male) Emory University undergraduates received course credit for their participation. Participants were native English speakers and reported no history of speech or hearing disorders.

Stimuli. Auditory stimuli consisted of a subset of the sentence-length utterances recorded from the speakers in Experiment 2. Sentences described the darkest and brightest shades for four (red, yellow, blue, green) of the six colors. Four of the six colors were chosen to be included in Experiment 3 in order to limit the length of the experiment. Sentences from one of the 48 speakers were excluded due to the presence of audible background noise, resulting in a total of 376 sentences (47 speakers describing four colors at two brightness levels) to be included as auditory stimuli for Experiment 3. Visual stimuli consisted of the second-brightest and second-darkest shades of the red, yellow, blue, and green color spectra used in Experiments 1 and 2. These particular brightness levels were chosen to minimize the possibility of participants perceiving the brightest and darkest shades of color as similar to white and black, respectively.

Procedure. On each trial, two swatches of the same color (one dark, one bright) were presented side by side on the computer screen. Participants then heard a sentence (e.g., “Can you get the blicket one?”) recorded in Experiment 2. Participants then chose between the two swatches which corresponded to the color swatch referred to in the sentence by pressing one of two designated keys on a button box corresponding to the left and right swatches on the computer screen. Stimulus presentation and data collection were controlled using E-prime 2.0 (Schneider et al., 2002). Auditory stimuli were presented binaurally over Beyerdynamic DT100 headphones at approximately 75 dB sound pressure level (SPL). Order of presentation was randomized with respect to speaker, color, and brightness level.

Results and Discussion

Participants' response accuracy was measured by the proportion of times participants 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. Collapsed across bright and dark trials, participants reliably chose the correct corresponding swatch (*M* = .52, *SD* = .04; *t*(46) = 3.29, *p* = .002). To assess the extent to which performance varied as a function of the robustness of the speakers' prosodic correlates to brightness, response accuracy was regressed separately 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 (1) account for baseline differences in individual speakers' acoustic characteristics and (2) capture each speaker's relative prosodic modulations across bright and dark sentences.

F_0. Fig. 4a shows listeners' response accuracy collapsed across bright and dark 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, $R^2 = .25$, $F(1, 46) = 14.73$, $p < .001$. Difference scores in pitch reliably predicted accuracy, $\beta = .50$, $t(46) = 3.84$, $p = <.001$, such that the larger a speaker’s difference score, the more accurate listeners’ mappings were for both bright and dark sentences. That is, the higher pitched speakers’ bright sentences were relative to their dark sentences, the more likely listeners were to choose the correct corresponding swatch.

Amplitude. Fig. 4b shows listeners' response accuracy collapsed across bright and dark trials as a function of each speaker's amplitude difference score. A linear regression equation regressing listener accuracy on this difference score accounted for a significant portion of variance in accuracy, $R^2 = .31$, $F(1, 46) = 20.10$, $p < .001$. Difference scores in pitch reliably predicted accuracy, $\beta = .56$, $t(46) = 4.48$, $p = <.001$, such that the larger a speaker’s amplitude difference score, the more accurate listeners’ mappings were for both bright and dark sentences. Speakers whose bright sentences were louder than their dark sentences elicited more accurate responses in listeners.

Duration. Fig. 4c shows listeners' response accuracy collapsed across bright and dark trials as a function of each speaker's duration difference score. A linear regression equation regressing listener accuracy on this difference score accounted for a significant portion of variance in accuracy, R^2 = .09, F(1, 46) = 4.39, p = .042. Difference scores in pitch reliably predicted accuracy, β = -.30, t(46) = -2.10, p = .042, such that the smaller a speaker’s duration difference score, the more accurate listeners’ mappings were for both bright and dark sentences. Speakers whose bright sentences were shorter than their dark sentences elicited more accurate responses in listeners.

Figure 4a

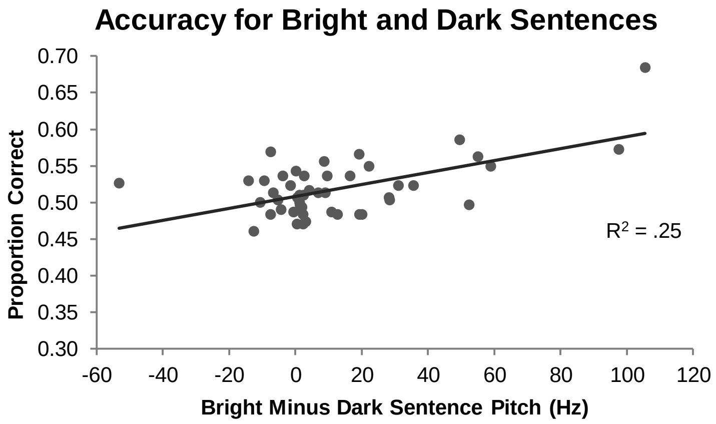


Figure 4b

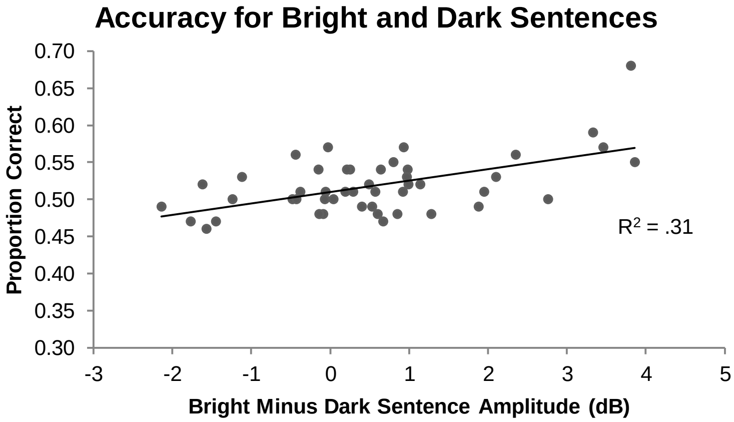


Figure 4c

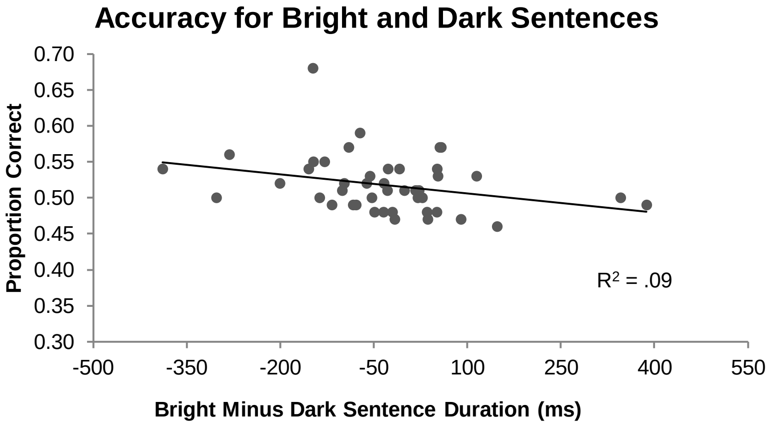


Figure 4a-c. Each data point represents average accuracy across listeners for one speaker's mean difference in pitch (Fig. 4a), amplitude (Fig. 4b), and duration (Fig. 4c) between sentences referring to bright versus dark shades Experiment 3. Chance listener performance is at .50.

That speakers' relative pitch, duration, and amplitude values between bright and dark sentences reliably predicted listeners' response accuracy is consistent with other findings demonstrating that listeners infer meaning from prosodic cues (Nygaard et al., 2009, Shintel et al., 2006; Shintel & Nusbaum, 2007; 2008). These results add to a growing body of evidence supporting the claim that prosody conveys information beyond linguistic structure and speakers' emotional state. The current findings are also consistent with demonstrations of systematic pitch-brightness mappings outside of language (Eitan & Timmers, 2009; Hubbard, 1996; Marks, 1974; Melara, 1989; Mondloch & Maurer, 2004) and imply that spoken language capitalizes on general cross-modal correspondences. Interestingly, differences between bright and dark sentences for all three acoustic measures, rather than in just pitch alone, reliably predicted listener accuracy. It is possible that because the provided color labels were completely devoid of semantic meaning, listeners might have recruited all acoustic cues available to aid in disambiguation. A second possibility is that pitch cues alone were too variable across speakers to be especially informative for disambiguating brightness, thus warranting the use of other acoustic information to facilitate choosing the target swatch.

Although previous work has provided evidence for systematic cross-modal associations in non-linguistic stimuli, the current findings demonstrate that these mappings also manifest in spoken language and that they offer functional significance in disambiguating meaning. I argue that prosody can therefore be conceptualized as a type of vocal gesture, as it provides referential details about objects and events in the world and resolves ambiguity in the accompanying linguistic content.

Experiment 4

A comparison of Experiments 1 and 2 suggest that speakers varied their prosody as a function of brightness only when color labels were novel rather than English words. The differing results from the two experiments suggest that speakers may be more likely to produce prosodic correlates to color brightness with increased demand to resolve referential ambiguity. In line with this possibility are findings suggesting that the likelihood with which speakers produce gestures increases with greater need to clarify linguistic ambiguity for the listener (Alibali et al., 2001; Holler & Beattie, 2003; Holler & Stevens, 2007; Jacobs & Garnham, 2007; McNeill, 1992). Using a referential communication task, Holler and Stevens (2007), for example, found that speakers conveying size information about objects to listeners gestured more when this information was new rather than known to the listener. Given that parallel patterns have been found in speaker's use of prosodic contours to highlight new and relevant information and that listeners are sensitive to these prosodic cues (e.g., Speer & Ito, 2011; Weber et al., 2006), one possibility is that speakers are more likely to produce prosodic cues to brightness when there is increased communicative need to provide referential information that cannot be resolved lexically. Under such circumstances, listeners may also be especially sensitive to potentially disambiguating acoustic information in the speech signal.

In Experiment 4, the demand to use prosody to convey referential detail was systematically manipulated 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 directly inform the extent to which communicative demand affects referential prosody use. Relative to the previous experiments, the communicative setting in Experiment 4 differed such that speakers were directing their utterances toward a simultaneously present listener whose goal was to select the intended target referent. In this more naturalistic communicative task, the effect of pragmatic context on referential prosody use, in addition to the effect of lexical ambiguity, could be assessed.

Given evidence that gesture use and pitch discrimination can vary as a function of personality characteristics (Hostetter & Potthoff, 2012) and musicality (Tervaniemi et al., 2005), participants also completed measures assessing these dimensions as an exploratory means by which to identify sources of variability in prosody perception and production. Identifying such sources of individual differences use may clarify the cognitive and social mechanisms underlying referential prosody use.

Method

Participants. Fifty-eight (29 speakers, 29 listeners) Emory University undergraduates participated for course credit. Participants were native English speakers with no history of speech or hearing disorders or color blindness.

Stimuli. Visual stimuli presented to the speakers and listeners were adapted from a subset of the swatches used in Experiments 1 and 2. One bright and one dark swatch from each of the six color spectra (red, orange, yellow, green, blue, purple) were chosen to be presented in pairs in Experiment 4. Color swatches were normed to ensure that they were (1) prototypical representations of the English color labels and (2) equally different in brightness within bright-dark color pairs. 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 employed in Experiments 1 and 2 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 represented the presented swatch. Swatches were selected as potential stimuli for Experiment 4 if they were labeled as the correct color and selected as representative of the color label by at least 70% of participants.

Results from the two tasks indicated that the bright and dark swatches of red, orange, and yellow spectra were more perceptually similar (closer to the midpoint of each spectrum) than the bright and dark swatches for green, blue, and purple spectra. The 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 the labeling task (n = 10) and yes-no 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 a pair of color swatches (one bright, one dark) from five[4] of the six color spectra and indicated on a response box whether the two presented swatches were equally bright. 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 Experiment 4 (see Appendix B for RGB coordinates of selected color swatches).

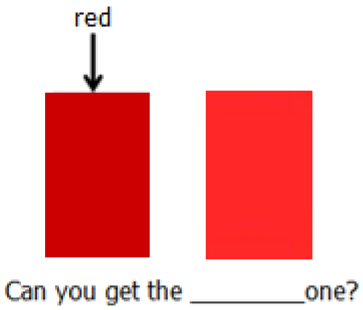
Procedure. Participants were randomly assigned as either a speaker or a listener to complete a referential communication task in pairs (e.g., Holler & Stevens, 2007;

[4] 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.

Keysar, Barr, Balin, & Brauner, 2000). One speaker and one listener were seated at adjacent 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 viewed two swatches, one dark, one bright, presented side by side on the computer screen. The speaker labeled one of the two swatches, which was indicated with an arrow and an English color label (appearing on the speaker's computer only) that corresponded to the swatch color (e.g., red, Fig. 5). Each swatch pair remained on the computer 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?"). English color terms (rather than novel words) were chosen as labels here to assess the use of referential prosody in a more naturalistic communicative setting.

Speakers were not explicitly told to vary their prosody and were instructed to indicate the color swatch as best as they can to the listener using only the target sentence and label. Utterances were recorded in a sound-attenuated room using an audio-technica ATR 20 microphone onto a Dell 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 PRAAT (Boersma & Weenink, 2012). To examine the acoustic features of the individual color labels in addition to the sentence-length utterances, each color label was segmented from the sentence utterance and amplitude normalized.

For a given trial, the listener viewed on his or her own computer screen the same two swatches as the speaker except without the accompanying arrow and text. 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[5].



Can you get the ________one?

Figure 5. Screenshot of speakers' view for a single trial in Experiment 4

To assess the extent to which communicative demand affects the recruitment of prosody to resolve referential ambiguity, the experiment consisted of trials that varied in ambiguity level. Trials were either ambiguous or unambiguous. 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. 6), 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. 6). Whether speakers were instructed to label the

[5] 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.

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. 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 a gray or yellow. Neither speakers nor listeners received corrective feedback during the practice trials or the experimental task.

To assess the extent to which individual differences in referential prosody perception and production may vary as a function of speaker- and listener-related dimensions, participants completed two self-report measures upon completion of the experimental task: the Ten Item Personality Inventory (TIPI; Gosling, Rentfrow, & Swann, 2003; see Appendix C) and the Empathy Quotient (EQ; Baron-Cohen & Wheelwright, 2004; see Appendix D). The TIPI is a self-report measure of the Big Five dimensions (openness to experience, conscientiousness, extraversion, agreeableness, emotional stability) and asks participants to rate how strongly they agree or disagree with the extent to which certain personality traits (e.g., reserved, quiet) apply to them. The EQ is a 60-item self-report measure designed to assess individual differences in cognitive-affective empathy and asks participants to rate how strongly they agree or disagree with statements such as “Other people tell me I am good at understanding how they are feeling and what they are thinking” and “I usually stay emotionally detached when watching a film.” Given evidence that prosody perception, and pitch perception in particular, varies with musicality (Tervaniemi et al., 2005; Thompson, Schellenberg, & Husain, 2004), participants also completed a questionnaire assessing their experience playing musical instruments (see Appendix E).

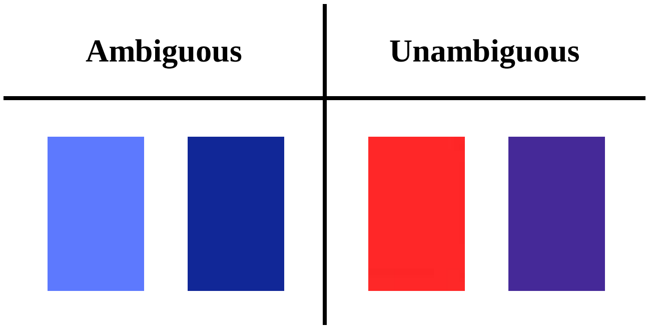


Figure 6. Examples of ambiguous and unambiguous trials in Experiment 4

Results and Discussion

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. Data from two 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, yielding a total of 24 speaker-listener pairs to be included in the reported analyses. Listeners reliably chose the correct corresponding color swatch for both unambiguous (*M* = .998, *SD* = .01; *t*(23) = 287.75, *p* < .001) and ambiguous (*M* = .83, *SD* = .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 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* $\eta^2 = .46$, and of Block, indicating improvement across blocks, $F(2, 46) = 4.38$, $p = .018$, *partial* $\eta^2 = .16$, as well as a significant interaction between these two variables, $F(2, 46) = 4.44$, $p = .017$, *partial* $\eta^2 = .16$ (see Fig. 7). All 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* $\eta^2 = .22$, but not unambiguous trials, $F(2, 46) = .49$, $p = .616$. Follow-up pairwise comparisons across blocks for ambiguous trials indicated a significant increase in listener accuracy between the first and second ($M_1 = .76$, $SD_1 = .23$; $M_2 = .85$, $SD_2 = .19$; $t(23) = 2.82$, $p = .010$) and first and third blocks ($M_1 = .76$, $SD_1 = .23$; $M_3 = .86$, $SD_3 = .24$; $t(23) = 2.54$, $p = .018$) of the task, suggesting that listeners learned to infer relevant acoustic information from speakers' utterances more systematically across blocks.

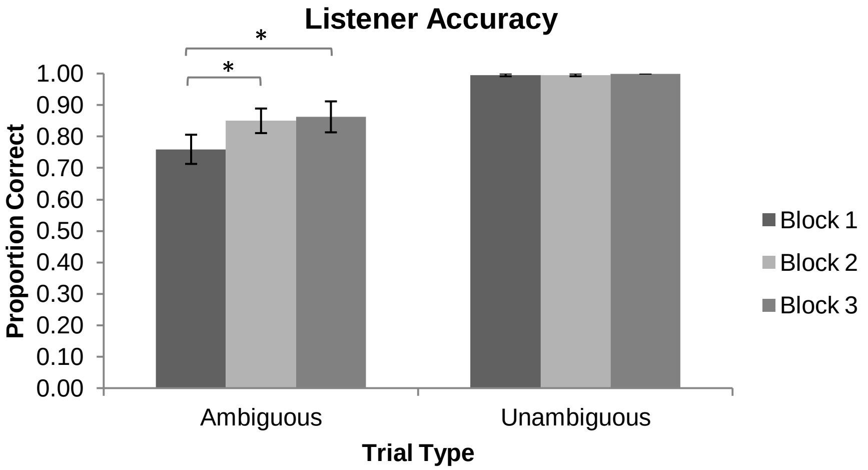


Figure 7. Listener accuracy as a function of block and trial type in Experiment 4. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

Speaker performance. Table 2 shows the F_0, amplitude, and duration of sentence-length utterances for bright and dark shades for both ambiguous and unambiguous trials. Because results patterned similarly overall at the sentence- and word-level, only sentence-level results are reported. Cases where the word-level results deviated from those at the sentence-level are noted when they occurred. As in Experiment 1, 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.

Table 2

Acoustic measurements of sentence-length utterances in Experiment 4

	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)	1510.75	1617.35	1241.51	1259.34

***F_0*.** Results indicated a significant effect of brightness, *F*(1, 23) = 8.13, *p* = .009, *partial* η^2 = .26, modified by a significant interaction between Trial Type and Brightness, *F*(1, 23) = 11.91, *p* = .002, *partial* η^2 = .34 (see Fig. 8). 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) = -.54, *p* = .593.

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

Duration. Sentences differed reliably in duration between trial types, *F*(1, 23) = 26.87, *p* < .001, *partial* η^2 = .54, such that sentences for ambiguous trials (*M* = 1562.31, *SD* = 329.78) were significantly longer than for unambiguous trials (*M* = 1249.56, *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* η^2 = .59, and Brightness, *F*(1, 23) = 7.94, *p* = .010, *partial* η^2 = .26. The interactions between Trial

Type and Brightness, *F*(1, 23) = 5.01, *p* = .035, *partial* η^2 = .18, and between Block and Brightness, *F*(2, 46) = 4.87, *p* = .023, *partial* η^2 = .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* η^2 = .19 (see Fig. 9a – 9b). 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 (*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.

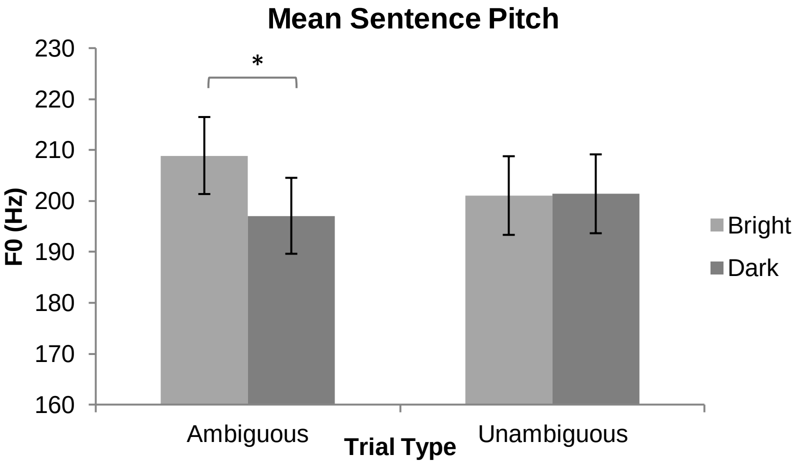


Figure 8. Mean sentence pitch as a function of trial type and brightness in Experiment 4. Error bars represent standard error of the mean for each condition, and indications of significance represent *p* < .05.

Figure 9a

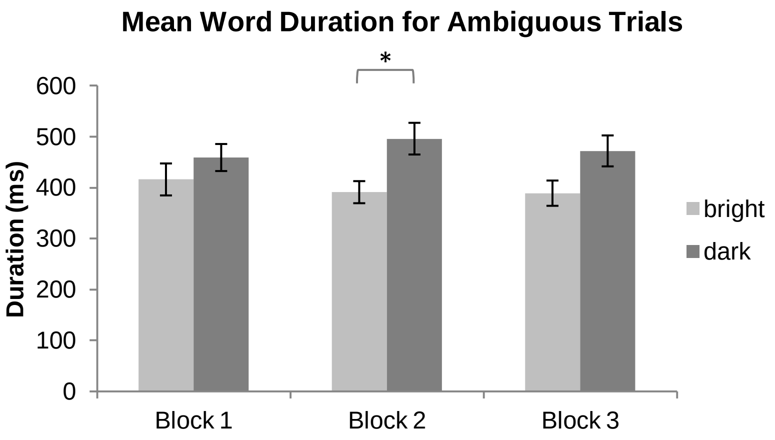


Figure 9b

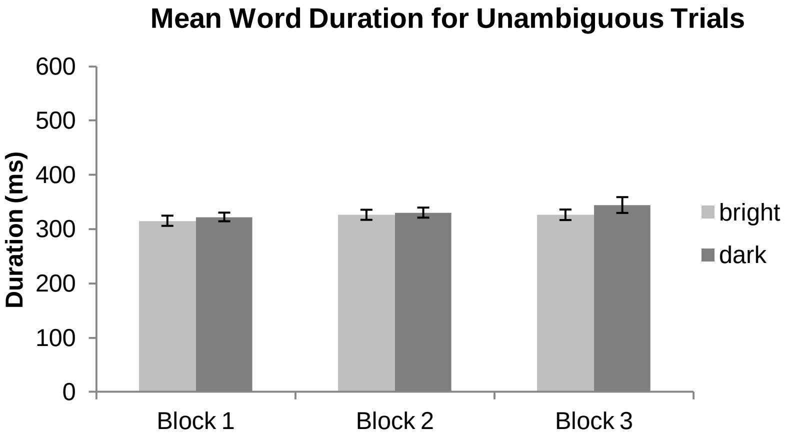


Figure 9a-b. Mean color word duration for ambiguous (Fig. 9a) and unambiguous trials (Fig. 9b) as a function of brightness and block in Experiment 4. Error bars represent standard error of the mean for each condition, and indications of significance represent $p < .05$.

Relation between listener and speaker performance. To assess the extent to which performance varied as a function of the robustness of the speakers' prosodic correlates to brightness, response accuracy was separately regressed on the mean difference between speakers' pitch, duration, and amplitude values for bright versus dark sentences. As in Experiment 3, 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.

***F_0*.** Fig. 10 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, $R^2 = .18$, $F(1, 23) = 4.74$, $p = .040$. Difference scores in pitch reliably predicted accuracy, $\beta = .42$, $t(23) = 2.18$, $p = .040$, 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, $R^2 = .12$, $F(1, 23) = 3.11$, $p = .070$, nor did difference scores reliably predict accuracy, $\beta = .35$, $t(23) = 1.77$, $p = .091$, suggesting that informative cues to brightness were not localized to the word level.

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 = .0002, *F*(1, 23) = .01, *p* = .945; duration, R^2 = .005, *F*(1, 12) = .12, *p* = .737), nor did they reliably predict accuracy (amplitude, *β* = .02, *t*(23) = .07, *p* = .945; duration, *β* = -.07, *t*(23) = -.34, *p* = .737), suggesting that listeners did not infer brightness from the relative difference in amplitude and duration between bright and dark sentences.

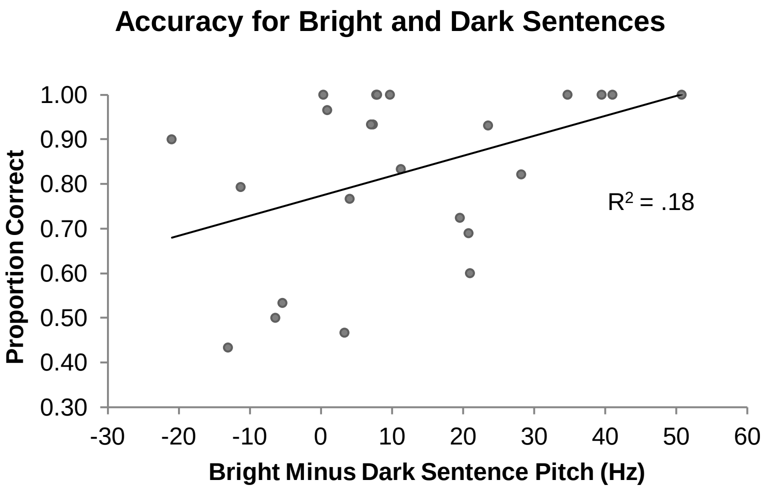


Figure 10. Each data point represents average accuracy across listeners for one speaker's mean difference in pitch between sentences referring to bright versus dark shades Experiment 4. Chance performance is at .50.

Taken together, the results for both listener accuracy and speaker performance suggest a prominent role of communicative demand in the use of prosodic cues to resolve referential ambiguity. 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. 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. The lack of communicative demand to employ prosodic cues to meaning may also have contributed to the non-significant differences in bright and dark acoustic cues found in Experiment 1. 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.

Across Experiments 3 and 4, pitch was the acoustic cue that most consistently predicted listener accuracy, suggesting that relative to duration and amplitude, pitch provided the most reliable disambiguating information about brightness. Given consistent demonstrations of listeners' sensitivity to the pitch-brightness mapping in non-linguistic stimuli (e.g., Eitan & Timmers, 2009, Marks, 1974), this finding was expected and suggests that the cross-modal correspondences that manifest in spoken language processing link specific visual and auditory dimensions. Although difference scores in duration and amplitude reliably predicted listener accuracy in Experiment 3, they did not in Experiment 4. Modulation in duration and amplitude in Experiment 4 was perhaps considered redundant for the listener when accompanied by pitch cues to brightness.

Unlike in Experiment 3 during which each listener heard utterances spoken by different speakers that changed across trials, each listener heard only one speaker in Experiment 4. One possibility is that whereas listeners employed all possible prosodic cues to brightness to overcome speaker variation in Experiment 3, this approach was unnecessary in Experiment 4. Another possibility is that even when speakers employed duration and amplitude cues in Experiment 4, they did so either inconsistently or in a manner that was not informative for the listener. That there was variability in listener accuracy across speakers with similar pitch difference scores (see Fig. 10) is suggestive of this possibility.

Individual differences in prosody perception and production. Exploratory analyses assessed whether task performance differed according to participant characteristics. Bonferroni-corrected Pearson's product-moment correlations were calculated to examine the extent to which speaker and listener task performance would vary as a function of participant empathy levels, personality characteristics, and amount of musical experience. Because only speakers' difference scores for pitch between bright and dark sentences predicted listener accuracy, correlations were calculated and reported using this variable as a measure of speaker task performance. For listeners, task performance was measured by response accuracy. As illustrated in Tables 3a – 3c, none of the measured participant characteristics correlated with either speaker prosodic modulation or listener response accuracy for ambiguous trials. However, listeners' Openness to Experience scores, described as a measure of their curiosity, creativity, open-mindedness and propensity to reflect (Gosling et al., 2003), were marginally correlated with listener accuracy (p = .10; see Table 3a), suggesting that listeners who scored highly on this attribute may have been more accepting of prosody as a potentially informative cue to brightness. Given that these analyses included data from only 24 speaker-listener pairs, a larger sample size may be necessary to detect a more robust correlative relation between participant characteristics and task performance. Variability in referential prosody use and may also be attributed to differences in lower-level auditory perceptual abilities, such as tone discrimination, or in levels of motivation to perform the task.

Table 3a.

Pearson product-moment correlation coefficients between participant characteristics and listener accuracy for ambiguous trials in Experiment 4

	Empathy	Extraversion	Agreeableness	Conscien-tiousness	Emotional Stability	Openness to Experience
Speaker Characteristics	.27	-.26	-.18	-.19	-.02	.09
Listener Characteristics	.02	-.06	.07	.10	-.04	.48

Note. p > .05 for all correlations

Table 3b.

Pearson product-moment correlation coefficients between speaker characteristics and sentence-level acoustic measures for ambiguous trials in Experiment 4

	Empathy	Extraversion	Agreeableness	Conscien-tiousness	Emotional Stability	Openness to Experience
Bright Minus Dark Pitch	.23	-.34	.08	-.09	.06	-.07

Note. p > .05 for all correlations

Table 3c.

Pearson product-moment correlation coefficients between participants' music experience and task performance in Experiment 4

	Number of Instruments	Hours Played per Week	Number of Years Played
Speaker Bright Minus Dark Pitch (sentence-level)	-.10	-.06	-.16
Listener Accuracy	.28	.19	.04

Note. p > .05 for all correlations

Summary of Results across Experiments

In four experiments, this dissertation used the mapping between pitch and color brightness to assess the possibility that referential prosody is an instantiation of general, systematic cross-modal mappings. Although the findings from Experiment 1 showed that speakers' prosody did not reliably differ across brightness levels when using English color labels, speakers' utterances in Experiment 2 varied as a function of brightness when using novel words such that labels for brighter shades were higher pitched, higher in amplitude, and shorter in duration. Taken together, the findings from these two studies suggest that referential prosody is more likely to be recruited when the accompanying linguistic content is ambiguous or underspecified.

From the utterances recorded in Experiment 2, listeners in Experiment 3 extracted prosodic cues to brightness and employed this information to reliably choose the corresponding target referent. In Experiment 4, speaker-listener pairs participated in a referential communication task during which speakers attempted to convey which of two color swatches they were indicating. Listeners reliably chose the correct corresponding swatch for ambiguous trials, when lexical information alone was insufficient to identify the target, suggesting that listeners recruited cues in the speakers' utterances to disambiguate between the two choice responses. Speakers' bright sentences were reliably higher pitched than dark sentences for ambiguous, but not unambiguous trials, suggesting that speakers did indeed provide meaningful acoustic cues to brightness when the accompanying linguistic content was underspecified. The findings from Experiment 4 also suggest that both lexical ambiguity and pragmatic context may affect the likelihood that referential prosody is recruited in spoken language. Taken together, results from the four reported experiments support a conceptualization of prosody as a source of referential information that is recruited to supplement linguistic content when there is communicative demand to do so.

General Discussion

A fundamental goal of research in the domain of spoken language processing is to characterize how listeners integrate *what* a speaker says with *how* the speaker says it. A parallel goal is to understand the cognitive mechanisms that underlie language users' ability to accomplish this integration efficiently and effectively. Examination of prosody in speech is a unique approach for achieving both objectives, as prosody contains suprasegmental acoustic cues that listeners must integrate during multiple stages of the speech perception process. Findings to date have highlighted the role of prosodic cues as conveying crucial information about syntactic structure and speakers' affective state. Notably, the literature on the role of prosody in speech perception has suggested that prosody overlays, but does not directly affect, processing of meaning.

The current work provides evidence that prosody can be conceptualized in an alternative way. Specifically, the present studies assess the possibility not only that prosody directly conveys information about linguistic reference but also that listeners infer meaning from these cues. Given consistent evidence for systematic auditory-visual mappings in non-linguistic stimuli, the current studies explore (1) the extent to which spoken language capitalizes on such cross-modal correspondences to convey meaningful referential detail and (2) under which communicative circumstances this might occur. Drawing from the results of the four reported experiments, I conclude that prosody can be conceptualized as a type of vocal gesture, as it is a source of referential detail that speakers can recruit to resolve ambiguity in the accompanying linguistic content.

The current results constrain theoretical models of spoken language processing and language evolution by clarifying the cognitive mechanisms by which language users express and infer meaning. In particular, these results provide evidence for theoretical perspectives that can account for language users' recruitment of systematic cross-modal correspondences to infer information from traditionally non-referential aspects of the speech signal. The current findings suggest that prosody in spoken language capitalizes on an inherent cross-modal perceptual system to maximize the efficiency and effectiveness of linguistic communication.

Systematic cross-modal mappings in spoken language

Accumulating evidence has documented human sensitivity to cross-modal correspondences across numerous combinations of semantic domains and sensory modalities. Pitch is implicated in many of the auditory-visual correspondences. In addition to the pitch-brightness mapping, individuals consistently and reliably associate pitch with visuo-spatial height (e.g., Melara & O’Brien, 1987; Walker et al., 2009), size (e.g., Evans & Treisman, 2010; Gallace & Spence, 2006), visual sharpness (e.g., Marks, 1987), and lightness in weight (e.g., Walker & Smith, 1985). Beyond auditory-visual correspondences, individuals have been found to systematically associate domains across other modalities, including shape and odor (e.g., Seo et al., 2014), shape and taste (e.g., Wan et al., 2014), and grapheme and color (e.g., Simner et al., 2005; Spector & Maurer, 2008).

Importantly, these systematic cross-modal mappings have consequences for perception. Using a series of speeded classification tasks, Evans and Treisman (2010) found that classification of a stimulus in one modality (e.g., pitch) was faster when a simultaneously presented irrelevant stimulus in another modality (e.g., vertical location) was congruent with respect to the hypothesized cross-modal mapping. Parallel patterns of results have been found in speeded discrimination tasks assessing sensitivity to the pitch-brightness, pitch-shape, and pitch-size mappings (Marks, 1987; Martino & Marks, 1999; Parise & Spence, 2008). Given these findings showing the reliability of correspondences involving pitch, it is perhaps unsurprising that the pitch-brightness mapping might manifest in spoken language, as pitch is a salient feature of an utterance’s prosodic contour. The current work thus adds to a growing literature demonstrating the use of prosody to express 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). Moreover, the present findings provide support for a mechanistic account suggesting that spoken language, and prosody in particular, capitalizes on cross-modal correspondences that are found in non-linguistic perceptual processing. One potential avenue for future work is to explore and characterize the range of semantic dimensions for which referential prosody is recruited to convey meaning.

Systematic cross-modal mappings also seem to manifest in metaphor. Metaphors such as *a bitter smile* or *a loud shirt,* for example, evoke sensory representations in multiple modalities. In an investigation of adults' sensitivity to auditory-visual metaphors, Marks et al. (1982) found that the word *sunlight* was rated as louder and brighter than *moonlight*, and *sneeze* as brighter and higher pitched than *cough*, suggesting that non-arbitrary cross-modal mappings may underlie these associations. One consideration is that sensory metaphors may be mediated by linguistic influences. Indeed, evidence from cross-cultural examinations of sensitivity to cross-modal correspondences suggests that the words used to describe polar dimensions (e.g., *high* or *low* to describe pitch) affect the extent to which individuals from different native language backgrounds exhibit sensitivity to mappings such as the pitch and visuo-spatial height correspondence (Dolscheid, Shayan, Majid, & Casasanto, 2013). However, pre-linguistic infants have been found to exhibit sensitivity to pitch-height and pitch-thickness mappings (Dolscheid et al., 2012; Walker et al., 2009), suggesting that sensitivity to cross-modal associations does not necessarily require linguistic expertise. Metaphors implicating these mappings may therefore reflect pre-linguistically available sensory correspondences.

Spence (2011) distinguishes among three types of cross-modal correspondences: structural, statistical, and semantically mediated, each of which has potentially different neural substrates and developmental trajectories. Structural cross-modal correspondences exist due to particular mappings between sensory cortices such that stimuli in one modality map non-arbitrarily to other sensory percepts. Cross-cortical activations may occur if the stimuli produce an increase in neural firing in each modality, as is the case with loud sounds and bright visual stimuli. Statistical correspondences are those that reflect co-occurrences between stimuli in the natural environment (e.g., big animals tend to make loud sounds), and semantically mediated correspondences are those for which common verbal labels are used to describe stimuli along ends of continua (e.g., *high* and *low* to describe pitch and spatial height).

The pitch-brightness mapping examined in the current study can perhaps be best characterized as a type of *structural* cross-modal correspondence. Ten-month-old infants have been found to exhibit sensitivity to pitch-brightness, but not pitch-size, associations in a violation-of-expectation procedure (Haryu & Kajikawa, 2012), suggesting that while the former association may represent an inherent bias, the latter may be learned by observing natural co-occurrences. That preverbal infants are sensitive to the pitch-brightness mapping supports its categorization as a structural, rather than statistical or semantically mediated cross-modal correspondence. Moreover, unlike the pitch-height mapping, the pitch-brightness association cannot be readily described using shared linguistic descriptors. Although future work will need to clarify the neural bases of the pitch-brightness association, the current findings, in the context of previous results, are consistent with the characterization of the pitch-brightness correspondence as a structural cross-modal mapping.

Neural substrates of cross-modal mappings in spoken language

Ramachandran and Hubbard (2001) have suggested that sensitivity to cross-modal correspondences is a consequence of cross-cortical activation and integration, whereby the acoustic and articulatory properties activate neural regions associated with the perceptual properties of the referent. Evidence from studies examining the neural basis for sensitivity to phonetic sound symbolism— non-arbitrary mappings between sound and meaning in natural language — provides support for this view (Asano et al., 2015; Imai et al., 2015; Kovic, et al., 2010; Revill et al., 2014). For example, Revill et al. (2014) found that relative to non-sound symbolic foreign words, sound symbolic words elicited increased activation in the left superior parietal cortex, an area associated with multi-sensory integration. Using electroencephalogram (EEG) measures, Asano et al. (2015) found that when presented with sound symbolically mismatched word-shape pairs, preverbal 11-month-old infants showed EEG patterns similar to an N400 effect, an index of semantic integration difficulty. General sensitivity to cross-modal mappings may therefore reflect an intrinsic cross-wiring of the nervous system that underlies neonatal perception (Mondloch & Maurer, 2004; Spector & Maurer, 2009). Systematic cross-modal correspondences between prosody and brightness in spoken language may potentially capitalize on these neural bases.

An alternative but not mutually exclusive possibility is that sensitivity to cross-modal correspondences, such as the pitch-brightness mapping, is a form of weak synaesthesia (Martino & Marks, 2001), a condition in which sensory stimuli (e.g., sounds, words) elicit simultaneous, often cross-modal percepts (e.g., colors, tastes). Unlike strong synaesthesia, weak synaesthesia does not elicit precise, often idiosyncratic cross-modal percepts (e.g., the sound of the letter *L* inducing the color red). Instead, weak synaesthesia is characterized by sensitivity to cross-modal associations that is shared by the general population and that does not manifest as specific and absolute percepts within particular sensory modalities. Conceptualized in this way, sensitivity to the pitch-brightness mapping may be considered a type of weak syneasthesia. Although recent evidence suggests that the synaesthesia may share neural correlates with sound symbolism sensitivity (Bankieris & Simner, 2015) and with sensitivity to cross-modal correspondences (Ward, Huckstep, & Tsakanikos, 2006), the precise relationship among the neural mechanisms underlying sound symbolism, cross-modal correspondences, and synaesthesia has yet to be determined.

Another proposed mechanism that could underlie systematic cross-modal associations is the matching of dimensions based on amodal representations of stimulus magnitude (Lourenco & Longo, 2011; Mondloch & Maurer, 2004; Smith & Sera, 1992; Spence, 2011). For example, the mapping of loud sounds to large objects may reflect a more general association between the “more” ends of amplitude and size. However, results from the current study cannot be fully attributed to magnitude mappings. First, unlike loudness and size, which are prothetic dimensions for which the polarity of more versus less is intrinsically determined, pitch and brightness are metathetic dimensions and lack intrinsic polarity (Smith & Sera, 1992). Second, whereas labels for brighter relative to darker shades were higher in pitch and amplitude in Experiment 2, they were shorter in duration. Thus, regardless of whether we conceptualize brighter (i.e., more white) or darker (i.e., more black) as the "more" end of the spectrum, the three acoustic measures map to brightness inconsistently. Without a systematic means by which to characterize pitch and brightness in more-less terms, a magnitude-based account cannot make definitive predictions across these domains.

Although the notion of *magnitude* is often considered synonymous with *intensity*, the two concepts might be different in the context of understanding mechanisms underlying cross-modal correspondences. Magnitude is limited in its ability to account for metathetic cross-modal mappings, as conceptualizing a dimension in terms of more versus less necessitates the assigning of dimension endpoints. Conceptualizing pitch and brightness in terms of intensity, however, bypasses this problem. Marks (1987) attributes sensitivity to cross-modal correspondences to a neural code for intensity such that the association between pitch and brightness might instead be attributed to similarity in the intensity of neural activity in the cortical regions activated in each respective modality. Behavioral evidence is consistent with notion. Indeed, Eitan and Timmers (2009) found that higher pitch and greater height were rated as more intense, suggesting that cross-modal correspondences between metathetic dimensions can be coded for similarly even without assigning more-less endpoints. Future work using imaging methods to assess the neural activation patterns in response to perceived cross-modal correspondences would further elucidate the extent to which these mappings can be attributed to neural representations of intensity.

Referential prosody and grounded views of spoken language processing

Evidence for systematic auditory-visual mappings in spoken language aligns with theories that assume language is grounded in multi-modal experiences (Barsalou, 1999; 2003; Glenberg & Kaschak, 2002; Zwaan et al., 2004). Grounded theories of language claim that linguistic symbols elicit simulations of perceptual experiences associated with external referents (Barsalou, 1999). Representations of these perceptual experiences are reactivated during linguistic processing and facilitate the comprehension of spoken language (Matlock, 2004; Šetić & Domijan, 2007; Zwaan et al., 2002).

Prosodic correlates to brightness may be an instantiation of the grounding of language in such multi-modal perceptual experiences. This does not imply, however, that language users need to have directly experienced simultaneous pitch-brightness co-occurrences to exhibit sensitivity to the pitch-brightness mapping. Grounded language accounts suggest that the simulations that occur during perception originate from bodily activity, which includes the brain states underlying stimulus perception. Consistent with an intensity-based characterization of cross-modal correspondences, perceptual simulations elicited by pitch-brightness stimuli might be driven instead by increases in the observer's alertness or arousal levels, or by parallel increases in cortical stimulation across sensory modalities. Sensitivity to cross-modal correspondences may also occur due to parallel effects each of the modalities has on the observer's emotional or affective state (Barsalou, 1999; Deroy & Spence, 2016). If so, such effects might be driven by similarities between stimuli in higher-order semantic dimensions, such as valence (good/bad), activity (fast/slow), and potency (weak/strong; Eitan & Timmers, 2009; Osgood, 1969).

Automaticity and context-dependence of referential prosody use

That the speakers in the current studies modulated their prosody as a function of referent brightness in Experiments 2 and 4 but not in Experiment 1 suggests that the extent to which referential prosody is recruited to disambiguate meaning varies with communicative context. That is, the process of using prosodic cues to meaning, at least in the pitch-brightness domain, may not be not automatic. One way to assess the automaticity of a process is to consider the extent to which it is implicitly and unconsciously executed. Within this framework, speakers' recruitment of prosody in Experiment 2 but not in Experiment 1 suggests that the use of prosody to convey reference was not automatic but instead varied depending on the degree of ambiguity in lexical information. This interpretation was confirmed in Experiment 4 during which speakers varied the extent to which they employed prosodic cues to reference within the same task depending on the referential ambiguity of the lexical information. However, this possibility does not necessarily imply that the detection of or sensitivity to the pitch-brightness correspondence on the part of listeners is not automatic. Cross-modal mappings have been found to modulate multi-sensory perception and integration such that processing is facilitated when cross-modally congruent versus incongruent (e.g., Evans & Treisman, 2010; Parise & Spence, 2008), suggesting that cross-modal correspondences likely have automatic effects. Rather, it is the recruitment of these mappings into spoken utterances that seems to occur at a decisional, goal-oriented level. Prosodic cues to meaning may therefore be stored in perceptual representations but are only accessed and employed when speakers find it pragmatically useful to do so.

Previous studies assessing the extent to which speakers recruit the pitch-height mapping to describe vertical motion found that speakers modulated their prosody accordingly even in the absence of any demand to resolve ambiguity (Shintel et al., 2006). However, this finding is not entirely inconsistent with the current pattern of results. The pitch-height mapping, unlike the pitch-brightness mapping, occurs in the natural environment and is semantically mediated, as both pitch and height are described as *high* and *low*. The pitch-height correspondence may thus be privileged and more likely to be recruited automatically in spoken language to provide referential detail. One potential avenue for future work is to systematically manipulate the communicative context for recruiting referential prosody using cross-modal correspondences from different sensory domains that differ in levels of previous exposure or salience. Another consideration is that the likelihood that prosodic cues to meaning are stored with lexical items may be higher for some words (e.g., valenced homophones; Nygaard & Lunders, 2002) than for others such that words with these prosodic cues stored are more likely to be produced with referential prosody. Thus, the likelihood with which referential prosody is recruited seems to be context-dependent and may vary with the extent to which the referents' grounded features are relevant to or activated in the current task (Barsalou, 1999; Lebois, Wilson-Mendenhall, & Barsalou, 2014).

Another potential source of differences in referential prosody use is variance associated with individual speaker or characteristics. That is, the extent to which language users employ and infer information from referential prosody may vary as a function of differences in personality traits or personal history. In contrast with recent findings from the co-speech gesture literature showing that gesture production correlated with speakers' extraversion and neuroticism levels (Hostetter & Potthoff, 2012), findings from exploratory analyses Experiment 4 found no reliable relation between prosody use or perception and the Big Five personality traits (extraversion, emotional stability, conscientiousness, agreeableness, and openness to experience) or empathy. Although previous work has found differences in cross-modal metaphor perception between individuals with and without musical training (Eitan & Timmers, 2009), no such evidence was found in the current study, perhaps due to underpowered analyses. The relation between referential prosody use and these characteristics thus warrants further investigation. Alternative sources of variation in prosody use and production may include differences in tendency to employ figurative language, artistic expertise, or motivation levels during the task. Because prosody perception entails sensitivity to fine-grained acoustic changes, individuals' perception and production of referential prosody may also vary with their lower-level auditory perception skills, such as pitch discrimination ability.

Parallels between referential prosody and co-speech gesture

Evidence from the current work suggests that prosody serves as an additional channel of referential detail that supplements the accompanying linguistic utterance. In this sense, prosody can be conceptualized as analogous to co-speech gesture (Perlman, 2010). Language users produce co-speech gestures that convey information that is non-redundant with the accompanying speech (Goldin-Meadow & Singer, 2003; McNeill, 1992). That speakers' prosody varied as a function of brightness in Experiments 2 and 4 but not Experiment 1 is consistent with findings in the gesture literature suggesting that speakers produce extra-linguistic cues to clarify ambiguity in the accompanying propositional content (Holler & Beattie, 2003; McNeill, 1992). Rather than provide information that is redundant with the accompanying speech, prosody, like gesture, offers referential information that is supplementary or non-redundant.

Another parallel between gesture and prosody is that both sources of information are encoded by the listener and can facilitate or interfere with listener comprehension of spoken language. For example, listeners have been found to retell a story more accurately and in more detail after hearing a speaker tell the story while producing gestures that match rather than mismatch with the accompanying speech (McNeill, Cassell, & McCullough, 1994). Similar findings have been found for when prosody and speech mismatch such that listeners are able to identify the correct meaning of spoken novel adjectives more accurately when prosodic cues matched rather than mismatched the referent (Nygaard et al., 2009). Reliably above-chance listener performance in Experiments 3 and 4 is also consistent with the notion that prosody directly influences the process of inferring meaning. Taken together, these and similar findings (e.g., Beattie & Shovelton, 1999; Broaders & Golden Meadow, 2010; Shintel et al., 2006) suggest that both gesture and prosody are sources of referential information that affect spoken language comprehension.

Co-speech gestures can be grouped into different categories according to what types of information they convey (McNeill, 1992). Perhaps most analogous to referential prosody are *representational gestures*, which include *iconic* gestures, or gestures that resemble the referent in form, *deictic* gestures, which indicate a location or path, and *metaphoric* gestures, which represent abstract referents (e.g., fairness). Because of their imagistic nature, representational gestures, and specifically iconic and deictic gestures, have been found to be particularly well-suited for expressing spatial information, such as size, shape, location, or directionality. Indeed, speakers gesture more when describing spatial versus non-spatial information, with facilitative effects on listener comprehension

(Alibali, 2005; Beattie & Shovelton, 2002; Driskell & Radtke, 2003; Hostetter, 2011). Driskell and Radtke (2003), for example, found that listeners identified spatial words (e.g., adjacent, cube) described by a speaker in fewer attempts when the speaker gestured than when the speaker did not gesture but showed no such advantage when the speaker described non-spatial words (e.g., warm, dark). This finding suggests that gestures enhance the communicative effectiveness of the accompanying linguistic description in the spatial realm. Gestures accompanying non-spatial descriptions, however, may be less informative, as non-spatial information may be more effectively and efficiently communicated using linguistic content or prosody.

Although both gesture and prosody can also serve as intensifiers of linguistic content (e.g., fast speech rate or rapid hand movements to convey the speed of motion event), prosody may do this more efficiently for a wider range of perceptual and semantic domains. Beyond the pitch-brightness mapping examined in the current study, prosody has been found to convey valence, strength, and heat information that listeners use to infer meaning (Herold et al., 2011; Nygaard et al., 2009). Such referential details can be considered abstract in nature and are perhaps more efficiently expressed using prosodic cues rather than gesture. Characterized in this way, the distinction between prosody as a source of referential versus emotional detail may be blurred, as valence-related information is likely to be emotionally grounded. Indeed, an alternative but not mutually exclusive means by which to parse the roles between gesture and prosody is to characterize gesture as a source of semantic reference and prosody as a source of emotionally-based information. However, that prosody can convey any information regarding semantic reference, even if it is rooted in emotional experience, has been overlooked in traditional characterizations of prosody. Taken together, the above findings suggest that prosody and gesture can be recruited to represent different aspects of an idea that is to be expressed. Which of the two channels of information is most useful may depend on how readily the modality (visuo-spatial versus auditory) lends itself to representing the semantic content (Levinson & Holler, 2014).

Prosody as a source of referential information: Implications for language evolution

That prosody has persisted over the course of language evolution as a source of referential information in spoken language suggests that it plays an integral role in effective communicative exchange. Prosody extends the range of available possibilities for expressing information beyond those that are afforded by propositional content alone. Evidence from the current study is consistent with this notion and suggests that prosody optimizes efficiency and effectiveness for conveying referential detail.

Prosody also appears to play several important roles in language acquisition. For example, infant-directed speech (IDS) is characterized by higher pitch, larger changes in pitch, exaggerated vowels, and increased repetition, all melodic qualities that infants and children prefer over adult-directed speech (ADS; Fernald, 1991). The same prosodic contours tend to be employed to modulate infant attention and arousal across cultures (Fernald et al., 1989). For example, soothing versus prohibitive utterances differ markedly in acoustic characteristics in ways that seem to directly impact the central nervous system. This provides an advantage in parent communication to infants by inducing desired responses prior to the infants' comprehension of language (Fernald, 1992). Conceptualized in this way, prosodic modifications in vocalizations have become selected for and conventionalized in order to convey and induce emotional states (Falk, 2004; Hauser & McDermott, 2003). That caretakers often sing their children to sleep also suggests that prosodic modulations served as a comfort-inducing mechanism (Trehub, 2003). According to account, song is considered a precursor to language, providing the foundation for speech as a "musical protolanguage" (Darwin, 1871; Hauser & McDermott, 2003; Masataka, 2007). Given the melodic qualities of prosody, one possibility is that prosodic cues were, like song, an evolutionary precursor to linguistic expression. That the pitch-brightness mapping manifests not only in spoken language but also in human music perception (Eitan &Timmers, 2009) and in non-human primate perception (Ludwig, Adachi, & Matsuzawa, 2011) is consistent with this possibility.

The prosodic characteristics of infant-directed speech can also facilitate language learning by providing children with syntactic (Cutler & Norris, 1988; Shi, Werker, & Morgan, 1999) and semantic (Herold et al., 2011) cues. In a picture-book reading task, mothers of two-year-old children were found to spontaneously modulate their prosody when using dimensional adjectives (e.g., big, small, hot, cold; Herold et al., 2011), suggesting that caregivers recruit prosody to signal linguistic reference. The pitch-brightness mapping can be conceptualized as an instantiation of this idea of non-arbitrary correspondence between form and referent that facilitates language comprehension and acquisition. Early in language evolution, spoken protolanguages may have capitalized on inherent cross-modal connections and/or iconicity, resemblance between form and referent, via prosody to balance the apparent referential power of an arbitrary linguistic system with ease of learning of a non-arbitrary system (Gasser, 2004; Monaghan & Christiansen, 2006). Indeed, evidence that infants as young as four months of age exhibit sensitivity to sound-shape correspondences (e.g., *bouba* as a label for a round versus pointy object) suggests that non-arbitrary mappings between sound and meaning are pre-linguistic and may bootstrap early word learning by facilitating word-referent associations (Ozturk, Krehm, & Vouloumanos, 2013).

If the evolutionary origins of prosody are rooted in conveying and modulating affective state, one question is how prosody came to also convey referential information. Ohala (1994) describes one view, the *frequency-code hypothesis*, which posits that because voice pitch conveys information regarding the size of the signaler, it is an indicator of related signaler characteristics, such as aggression, assertiveness, and dominance. Although this hypothesis pertains to the potential for prosody to convey details describing the speaker rather than the referent, it presents the possibility that prosodic cues may have served as an early precursor to semantic reference. With increased control of the human vocal tract (Fitch, 2000) and increased demand for referential specificity (Monaghan, Shillcock, Christiansen, & Kirby, 2014), a vocal, symbolic language system may have emerged from such use of prosody.

Another critical question is why prosody and co-speech gesture, as channels of referential information, have continued to be integral to human communication even as language has evolved into a highly functional, largely arbitrary system. One possibility is that neither prosody nor gesture readily offers sufficient communicative precision on its own and have co-evolved to serve non-redundant, supplementary roles in disambiguating linguistic content. This notion is consistent with the above-described findings suggesting that whereas co-speech gestures readily convey informative spatial detail, prosody provides cues to non-spatial information, such as valence and intensity. Irrespective of the evolutionary *origins* of language, the current findings suggest that the vocal and manual gesture systems may have co-developed together to augment language, ensuring maximal communicative flexibility (Perniss, Thompson, & Vigliocco, 2010; Perniss & Vigliocco, 2014).

Conclusion

The current work provides evidence that prosody can be conceptualized as a type of vocal gesture that capitalizes on systematic cross-modal correspondences to convey information about objects and events in the world. Challenging traditional conceptualizations of prosody as conveying non-referential detail, the present findings suggest not only that prosody can be recruited to convey disambiguating information about linguistic reference but also that listeners infer meaning from these cues. Manifestation of the pitch-brightness mapping in spoken language implies that prosody is an instantiation of an inherently cross-modal perceptual experience and may have co-evolved with the lexicon to optimize efficiency and flexibility in a multi-modal communicative system.

References

Alibali, M. W. (2005). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition and Computation*, *5*(4), 307-331. doi: 10.1207/s15427633scc0504_2

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. doi:10.1006/jmla.2000.2752

Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., & Thierry, G. (2015). Sound symbolism scaffolds language development in preverbal infants. *Cortex*, *63*, 196-205. doi: 10.1016/j.cortex.2014.08.025

Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science, 8*(2), 53–57. doi: 10.1111/1467-8721.00013

Bachorowski, J.A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic Properties of speech are associated with emotional intensity and context. *Psychological Science*, *6*(4), 219-224. doi:10.1111/j.1467-9280.1995.tb00596.x

Bankieris, K., & Simner, J. (2015). What is the link between synaesthesia and sound symbolism?. *Cognition, 136,* 186-195. doi: 10.1016/j.cognition.2014.11.013

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614-36. doi:10.1037/0022-3514.70.3.614

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger Syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Development Disorders, 34*(2), 163-175. doi: 10.1023/B:JADD.0000022607.19833.00

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

Barsalou, L.W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes, 18(5-6)*, 513-562. doi:10.1080/01690960344000026

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica, 123,* 1–30. doi:10.1515/

Beattie, G., & Shovelton, H. (2002). An experimental investigation of some properties of individual iconic gestures that mediate their communicative power. *British Journal of Psychology, 93*(2), 179-192. doi: 10.1348/000712602162526

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer [Computer program]. Version 5.3.43, retrieved 27 March 2013 from http://www.praat.org/

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception and Psychophysics*, *61*(2), 206-219. doi:10.3758/BF03206883

Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, *10*(2), 137-167. doi: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. doi:10.1177/0956797610366082

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(3), 521-533. doi:10.1037/0278-7393.20.3.521

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

Clifton, C., Carlson, K., & Frazier, L. (2002). Informative prosodic boundaries. *Language and Speech*, *45*(2), 87-114. doi:10.1177/00238309020450020101

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, *31*(2), 218-236. doi:10.1016/0749-596X(92)90012-M

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*(3-4), 133-142. doi:10.1016/0885-2308(87)90004-0

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: a literature review. *Language and Speech*, *40*(2), 141-201. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9509577

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. doi:10.1037//0096-1523.14.1.113

Deroy, O., & Spence, C. (2016). Crossmodal correspondences: four challenges. *Multisensory Research*, *29*(1-3), 29-48. doi:10.1163/22134808-00002488

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. doi:10.1177/0956797612457374

Driskell, J. E., & Radtke, P. H. (2003). The effect of gesture on speech production and comprehension. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *45*(3), 445-454. doi:10.1518/hfes.45.3.445.27258

Eitan, Z., & Timmers, R. (2009). Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition*, *114*(3), 405-422. doi:10.1016/j.cognition.2009.10.013

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), 1-12. doi:10.1167/10.1.6

Falk, D. (2004). Prelinguistic evolution in early hominins: Whence motherese?. *Behavioral and Brain Sciences*, *27*(4), 491-503. doi:10.1017/S0140525X04000111

Fernald, A. (1991). Prosody in speech to children: Prelinguistic and linguistic functions. *Annals of Child Development*, *8*, 43-80.

Fernald A. (1992). Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In: JH Barkow, L Cosmides and J Tooby (eds), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 391-428). New York: Oxford University Press.

Fernald, A., & Kuhl, P.K. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development, 10*(3), 279–293. doi:10.1016/0163-6383(87)90017-8

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477-501. doi:10.1017/S0305000900010679

Fitch, W. T. (2000). The evolution of speech: a comparative review. *Trends in Cognitive Sciences*, *4*(7), 258-267. doi:10.1016/S1364-6613(00)01494-7

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception and Psychophysics*, *68*(7), 1191-1203. doi:10.3758/BF03193720

Gasser, M. (2004). The origins of arbitrariness in language. *Proceedings of the Cognitive Science Society* (pp. 434-439). Hilllsdale, NJ: Lawrence Erlbaum.

Glenberg, A. M., & Kaschak, M.P. (2002). Grounding language in action. *Psychonomic Bulletin and Review, 9*(3), 558-565. doi: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. doi:10.1037/0012-1649.39.3.509

Goldinger, S.D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166-1183. doi:10.1037/0278-7393.22.5.1166

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality, 37*(6), 504-528. doi:10.1016/S0092-6566(03)00046-1

Greasley, P., Sherrard, C., & Waterman, M. (2000) Emotion in language and speech: Methodological issues in the coding of natural data. *Language and Speech, 43*(4), 355-375. doi: 10.1177/00238309000430040201

Grice, P. (1989). *Studies in the Way of Words.* Cambridge: 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. doi:10.1016/j.infbeh.2012.07.015

Hauser, M. D., & McDermott, J. (2003). The evolution of the music faculty: A comparative perspective. *Nature Neuroscience*, *6*(7), 663-668. doi:10.1038/nn1080

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. doi:10.1016/j.jecp.2010.09.005

Herold, D. S., Nygaard, L. C., & Namy, L. L. (2011). Say it like you mean it: Mothers' use of prosody to convey word meaning. *Language and Speech, 55*(3), 423-436. doi:10.1177/0023830911422212

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. doi:0.1075/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. doi:10.1177/0261927X06296428

Hostetter, A.B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin, 137*(2), 297-315. doi:10.1037/a0022128

Hostetter, A. B., & Alibali, M. W. (2007). Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture*, *7*(1), 73-95. doi:10.1075/gest.7.1.05hos

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.* John Benjamins.

Hostetter, A. B., & Potthoff, A. L. (2012). Effects of personality and social situation on representational gesture production. *Gesture*, *12*(1), 62-83. doi:10.1075/gest.12.1.04hos

Hubbard, T. L. (1996). Synesthesia-like mappings of lightness, pitch, and melodic interval. *The American Journal of Psychology*, 219-238. doi: 10.2307/1423274

Imai, M., Miyazaki, M., Yeung, H.H., Hidaka, S., Kantartzis, K., Okada, H., & Kita, S. (2015). Sound symbolism facilitates word learning in 14-month-olds. *Plos One, 10*(2): e0116494. doi:10.1371/journal.pone.0116494

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

Jusczyk, P.W., & Aslin, R.N. (1995). Infants’ detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29*(1), 1–23. doi:10.1006/cogp.1995.1010

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants ' preference for the predominant patterns of English words stress, *Child Development*, *64*(3), 675-687. doi:10.2307/1131210

Kempe, V., Schaeffler, S., & Thoresen, J. C. (2010). Prosodic disambiguation in child-directed speech. *Journal of Memory and Language*, *62*(2), 204-225. doi:10.1016/j.jml.2009.11.006

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. doi:10.1111/1467-9280.00211

Kovic, V., Plunkett, K., & Westermann, G. (2010). The shape of words in the brain. *Cognition*, *114*(1), 19-28. doi: 10.1016/j.cognition.2009.08.016

Kunihira, S. (1971). Effects of the expressive voice on phonetic symbolism. *Journal of Verbal Learning and Verbal Behavior, 10*(4), 427– 429. doi:10.1016/S0022-5371(71)80042-7

Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*, *19*(5), 633-653. doi:10.1080/02699930441000445

Lebois, L. A., Wilson-Mendenhall, C. D., & Barsalou, L. W. (2015). Are automatic conceptual cores the gold standard of semantic processing? The context-dependence of spatial meaning in grounded congruency effects. *Cognitive Science, 39*(8), 1764-1801. doi: 10.1111/cogs.12174

Lehiste, I. (1970). *Suprasegmentals.* Cambridge, MA: MIT Press.

Leinonen, L., Hiltunen, T. I. Linnankoski, I., & Laakso, M.J. (1997). Expression of emotional–motivational connotations with a one-word utterance. *Journal of the Acoustical Society of America, 102*(3), 1853–1863. doi:10.1121/1.420109

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B*, *369*(1651), 20130302. doi: 10.1098/rstb.2013.0302

Lourenco, S.F. & Longo, M.R. (2011). Origins and development of generalized magnitude representation. In S. Dehaene & E. Brannon (Eds.), *Space, Time and Number in the Brain: Searching for the Foundations of Mathematical Thought* (pp. 225–244). London: Academic.

Ludwig, V. U., Adachi, I., & Matsuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (Pan troglodytes) and humans. *Proceedings of the National Academy of Sciences*, *108*(51), 20661-20665. doi:10.1073/pnas.1112605108

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

Marks, L. E. (1982). Bright sneezes and dark coughs, loud sunlight and soft moonlight. *Journal of Experimental Psychology: Human Perception and Performance, 8*(2), 177-193. doi: 10.1037/0096-1523.8.2.177

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. doi:10.1037/0096-1523.13.3.384

Marks, L. E., Hammeal, R. J., Bornstein, M. H., & Smith, L. B. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development*, 1-100. doi:10.2307/1166084

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, *28*(7), 903-923.doi: 10.1068/p2866

Martino, G., & Marks, L. E. (2001). Synesthesia: Strong and weak. *Current Directions in Psychological Science*, *10*(2), 61-65. doi:10.1111/1467-8721.00116

Masataka, N. (2007). Music, evolution and language. *Developmental Science, 10*(1), 35-39.doi: 10.1111/j.1467-7687.2007.00561.x

Matlock, T. (2004). Fictive motion as cognitive simulation. *Memory and Cognition*, *32*(8), 1389-1400. doi:10.3758/BF03206329

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, *23*(2), 209-237. doi:10.1080/02699930802204677

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

McNeill, D., Cassell, J., & McCullough, K. E. (1994). Communicative effects of speech-mismatched gestures. *Research on Language and Social Interaction*, *27*(3), 223-237. doi:10.1207/s15327973rlsi2703_4

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

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, *116*(4), 323-336. doi:10.1037/0096-3445.116.4.323

Monaghan, P., & Christiansen, M.H. (2006). Why form-meaning mappings are not entirely arbitrary in language. *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1838-1843). Mahwah, NJ: Lawrence Erlbaum.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130299. doi:10.1098/rstb.2013.0299

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. doi:10.3758/CABN.4.2.133

Morton, J. B., Trehub, S. E., & Bruce, J. (2001). Children’s understanding of emotion in speech. *Emotion*, *72*(3), 834-843. doi:10.1111/1467-8624.00318

Nygaard, L.C., Burt, S.A, & Queen, J. S. (2000). Surface form typicality and asymmetric transfer in episodic memory for spoken words. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(5), 1228-1244. doi:10.1037/0278-7393.26.5.1228

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. doi:10.1111/j.1551-6709.2008.01007.x

Nygaard, L. C., & Lunders, E. R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory and Cognition*, *30*(4), 583-593. doi:10.3758/BF03194959

Nygaard, L. C., & Queen, J. S. (2008). Communicating emotion: linking affective prosody and word meaning. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(4), 1017-1030. doi:10.1037/0096-1523.34.4.1017

Ohala, J.J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. Leanne Hinton, Johanna Nichols, John J. Ohala (Eds.), Sound Symbolism (pp. 325–347), Cambridge University Press: Cambridge.

Osgood, C.E. (1969). On the whys and wherefores of E. P. A. *Journal of Personality and Social Psychology*, *12*(3), 194-199. doi:10.1037/h0027715

Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, *114*(2), 173-186. doi:10.1016/j.jecp.2012.05.004

Palmeri, T J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *19*(2), 309-328. doi:10.1037/0278-7393.19.2.309

Parise, C., & Spence, C. (2008). Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters*, *442*(3), 257-261. doi:10.1016/j.neulet.2008.07.010

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*(2), 107-120. doi:10.1007/s10919

Perlman, M. (2010). Talking fast: The use of speech rate as iconic gesture. In Parrill, F., Tobin, V., & Turner, M. (Eds.), *Meaning, Form, and Body.* Stanford: CSLI Publications.

Perlman, M., Clark, N., & Falck, M.J. (2014). Iconic prosody in story reading. *Cognitive Science*, 1-21. doi:10.1111/cogs.12190

Perniss, P., Thompson, R., & Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology, 1,* 1-15. doi:10.3389/fpsyg.2010.00227

Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: from a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130300. doi:10.1098/rstb.2013.0300

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, *90*(6), 2956-2970.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia- A window into perception, thought and language. *Journal of Consciousness Studies*, *8*(12), 3-34.

Reinisch, E., Jesse, A., & Nygaard, L. C. (2012). Tone of voice guides word learning in informative referential contexts. *Quarterly Journal of Experimental Psychology*, *66*(6), 1227-1240. doi:10.1080/17470218.2012.736525

Revill, K. P., Namy, L. L., DeFife, L. C., & Nygaard, L. C. (2014). Cross-linguistic sound symbolism and crossmodal correspondence: Evidence from fMRI and DTI. *Brain and Language, 128*(1), 18-24. doi:10.1016/j.bandl.2013.11.002

Scherer, K. (1994). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143–165. doi:10.1037//0033-2909.99.2.143

Schirmer, A., Kotz, S. A., & Friederici, A.D. (2002). Sex differentiates the role of emotional prosody during word processing. *Cognitive Brain Research, 14*(2), 228-233. doi:10.1016/S0926-6410(02)00108-8

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

Seo, H. S., Arshamian, A., Schemmer, K., Scheer, I., Sander, T., Ritter, G., & Hummel, T. (2010). Cross-modal integration between odors and abstract symbols. *Neuroscience Letters*, *478*(3), 175-178. doi:10.1016/j.neulet.2010.05.011

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

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2), 193-247. doi:10.1007/BF01708572

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*(2), B11-B21. doi:10.1016/S0010-0277(99)00047-5

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. doi: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. doi: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. doi:10.1016/j.jml.2006.03.002

Simner, J., Ward, J., Lanz, M., Jansari, A., Noonan, K., Glover, L., & Oakley, D. A. (2005). Non-random associations of graphemes to colours in synaesthetic and non-synaesthetic populations. *Cognitive Neuropsychology*, *22*(8), 1069-1085. doi:10.1080/02643290500200122

Singh, L., Morgan, J. L., & Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk? *Infancy*, *3*(3), 365-394. doi:10.1207/S15327078IN0303_5

Smith, L.B., & Sera, M.D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology*, *24*(1), 99-142. doi: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. doi:10.1016/S0749-596X(02)00519-3

Spector, F., & Maurer, D. (2008). The colour of Os: naturally biased associations between shape and colour. *Perception, 37*(6), 841-847. doi:10.1068/p5830

Spector, F., & Maurer, D. (2009). Synesthesia: a new approach to understanding the development of perception. *Developmental Psychology, 45*(1), 175-189. doi:10.1037/a0014171

Speer, S. R., & Ito, K. (August, 2011). Prosodic properties of contrastive utterances in spontaneous speech. Paper presented at the 17th International Congress of Phonetic Sciences, Hong Kong, 1890-1893.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, and Psychophysics*, *73*(4), 971-995. doi:10.3758/s13414-010-0073-7

Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs. non-musicians: An event-related potential and behavioral study. *Experimental Brain Research*, *161*(1), 1-10. doi:10.1007/s00221-004-2044-5

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53-71. doi:10.1207/s15327078in0701_5

Thompson, W. F., & Balkwill, L. (2006). Decoding speech prosody in five languages. *Semiotica*, 158*(4)*, 407-424. doi:10.1515/SEM.2006.017

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: do music lessons help?. *Emotion*, *4*(1), 46-64. doi:10.1037/1528-3542.4.1.46

Trehub, S. E. (2003). The developmental origins of musicality. *Nature Neuroscience, 6*(7), 669-673.doi: 10.1038/nn1084

Wagner, M., & Watson, D. G. (2012). Experimental and theoretical advances in prosody:

A review. *Language and Cognitive Processes*, *25*(7-9), 905-945. doi:10.1080/01690961003589492

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2009). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, *21*(1), 21-25. doi:10.1177/0956797609354734

Walker, P., & Smith, S. (1985). Stroop interference based on the multimodal correlates of haptic size and auditory pitch. *Perception*, *14*(6), 729-736. doi: 10.1068/p140729

Wan, X., Woods, A. T., van den Bosch, J. J., McKenzie, K. J., Velasco, C., & Spence, C. (2014). Cross-cultural differences in crossmodal correspondences between basic tastes and visual features. *Frontiers in Psychology*, *5*, 1365. doi:10.3389/fpsyg.2014.01365

Ward, J., Huckstep, B., & Tsakanikos, E. (2006). Sound-colour synaesthesia: To what extent does it use cross-modal mechanisms common to us all?. *Cortex*, *42*(2), 264-280.doi:10.1016/S0010-9452(08)70352-6

Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, *19*(6), 713-755. doi:10.1080/01690960444000070

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. doi:10.1177/00238309060490030301

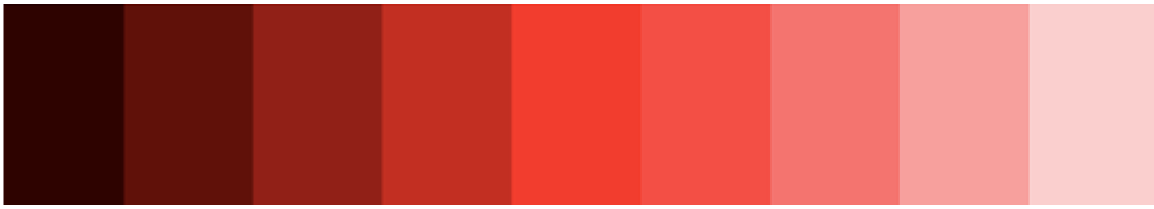
Wurm, L. H., Vakoch, D.A., Strasser, M. R., Calin-Jageman, R., & Ross, S. E. (2001).

Speech perception and vocal expression of emotion. *Cognition and Emotion*, *15*(6), 831-852. doi:10.1080/02699930143000086

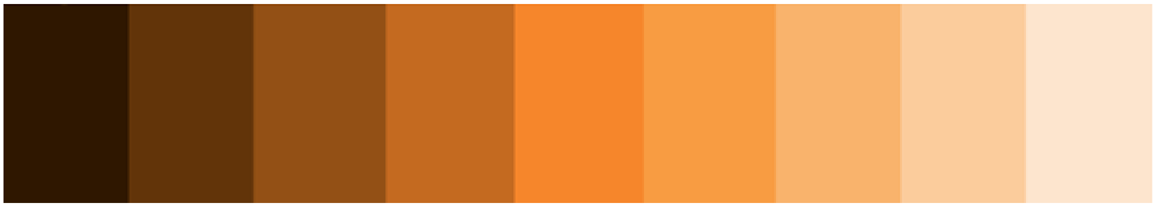
Zwaan, R. A., Madden, C. J., Yaxley, R. H., & Aveyard, M. E. (2004). Moving words: Dynamic representations in language comprehension. *Cognitive Science*, *28*(4), 611-619. doi:10.1016/j.cogsci.2004.03.004

Zwaan, R.A., Stanfield, R. A, & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science*, *13*(2), 168-171. doi:10.1111/1467-9280.00430

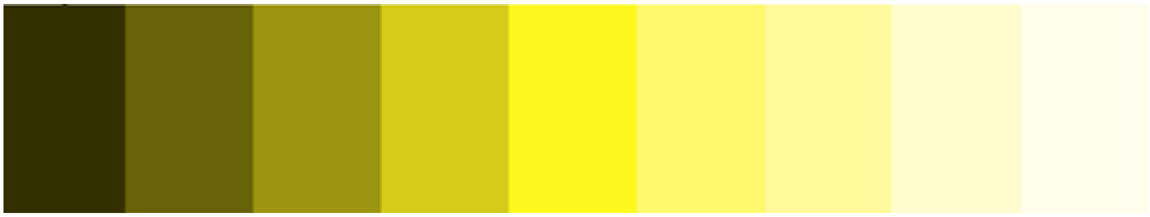
Appendix A



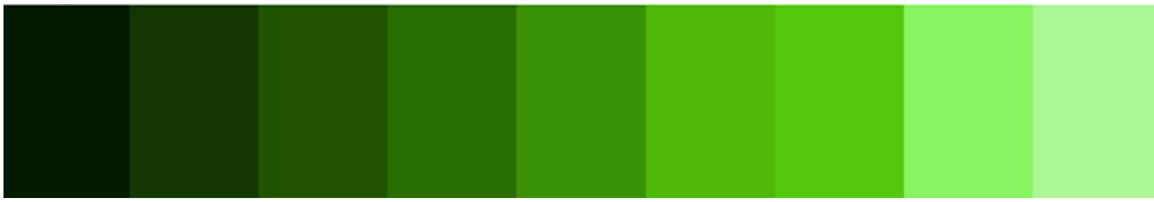
R: 51	R: 102	R: 153	R: 204	R: 255	R: 255	R: 255	R: 255	R: 255
G: 0	G: 0	G: 0	G: 0	G: 0	G: 51	G: 102	G: 153	G: 204
B: 0	B: 0	B: 0	B: 0	B: 0	B: 51	B: 102	B: 153	B: 204



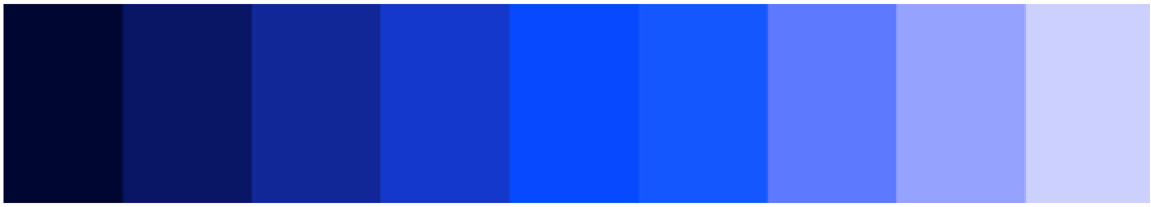
R: 51	R: 102	R: 153	R: 204	R: 255	R: 255	R: 255	R: 255	R: 255
G: 25	G: 51	G: 76	G: 102	G: 128	G: 153	G: 178	G: 204	G: 229
B: 0	B: 0	B: 0	B: 0	B: 0	B: 51	B: 102	B: 153	B: 204



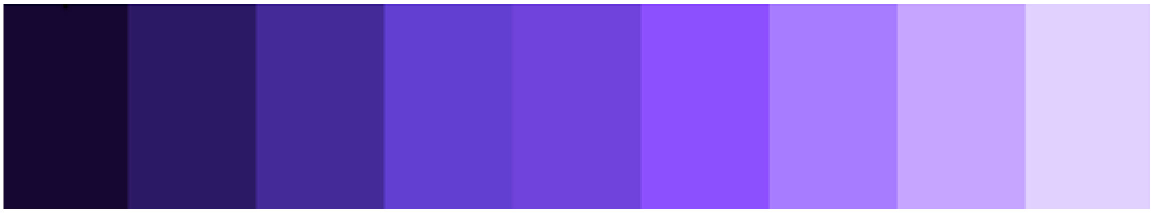
R: 51	R: 102	R: 153	R: 204	R: 255	R: 255	R: 255	R: 255	R: 255
G: 51	G: 102	G: 153	G: 204	G: 255	G: 255	G: 255	G: 255	G: 255
B: 0	B: 0	B: 0	B: 0	B: 0	B: 80	B: 122	B: 153	B: 204



R: 0	R: 0	R: 0	R: 0	R: 0	R: 51	R: 102	R: 153	R: 204
G: 51	G: 102	G: 153	G: 204	G: 255	G: 255	G: 255	G: 255	G: 255
B: 0	B: 0	B: 0	B: 0	B: 0	B: 51	B: 102	B: 153	B: 204



R: 0	R: 0	R: 0	R: 0	R: 0	R: 51	R: 102	R: 153	R: 204
G: 0	G: 0	G: 0	G: 0	G: 0	G: 51	G: 102	G: 153	G: 204
B: 51	B: 102	B: 153	B: 204	B: 255	B: 255	B: 255	B: 255	B: 255



R: 25	R: 51	R: 76	R: 102	R: 127	R: 153	R: 178	R: 204	R: 229
G: 0	G: 0	G: 0	G: 0	G: 0	G: 51	G: 102	G: 153	G: 204
B: 51	B: 102	B: 153	B: 204	B: 225	B: 255	B: 255	B: 255	B: 225

Note: Color shades were created by adjusting RGB coordinates using http://www.rapidtables.com/web/color/RGB_Color.htm.

Appendix B

	Ambiguous			Unambiguous	
	🟥	🟥		🟩	🟦
R:	255	204	R:	102	0
G:	40	0	G:	102	153
B:	40	0	B:	255	0
	🟧	🟧		🟧	🟩
R:	255	228	R:	255	0
G:	58	0	G:	58	0
B:	170	91	B:	170	153
	🟩	🟩		🟪	🟧
R:	102	0	R:	178	255
G:	102	0	G:	255	58
B:	255	153	B:	102	170
	🟦	🟦		🟥	🟪
R:	102	0	R:	255	76
G:	255	153	G:	40	153
B:	102	0	B:	40	0
	🟪	🟪		🟦	🟥
R:	178	76	R:	102	204
G:	255	153	G:	255	0
B:	102	0	B:	102	0

Appendix C

Ten-Item Personality Inventory (TIPI)

Here are a number of personality traits that may or may not apply to you. Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

Disagree strongly	Disagree moderately	Disagree a little	Neither agree or disagree	Agree a little	Agree moderately	Agree strongly
1	2	3	4	5	6	7

I see myself as:

1. _____ Extraverted, enthusiastic.

2. _____ Critical, quarrelsome.

3. _____ Dependable, self-disciplined.

4. _____ Anxious, easily upset.

5. _____ Open to new experiences, complex.

6. _____ Reserved, quiet.

7. _____ Sympathetic, warm.

8. _____ Disorganized, careless.

9. _____ Calm, emotionally stable.

10. _____ Conventional, uncreative.

Appendix D

Empathy Quotient

How to Fill Out the Questionnaire

Below is a list of statements. Please read each statement carefully and rate how strongly you agree or disagree with it by circling your answer. There are no right or wrong answers, or trick questions.

IN ORDER FOR THE SCALE TO BE VALID, YOU MUST ANSWER EVERY QUESTION.

Examples

E1. I would be very upset if I couldn't listen to music every day.	[✓ strongly agree]	slightly agree	slightly disagree	strongly disagree
E2. I prefer to speak to my friends on the phone rather than write letters to them.	strongly agree	slightly agree	[✓ slightly disagree]	strongly disagree
E3. I have no desire to travel to different parts of the world.	strongly agree	[✓ slightly agree]	slightly disagree	strongly disagree
E4. I prefer to read than to dance.	strongly agree	slightly agree	slightly disagree	[✓ strongly disagree]

Questionnaire

1. I can easily tell if someone else wants to enter a conversation.	strongly agree	slightly agree	slightly disagree	strongly disagree
2. I prefer animals to humans.	strongly agree	slightly agree	slightly disagree	strongly disagree
3. I try to keep up with the current trends and fashions.	strongly agree	slightly agree	slightly disagree	strongly disagree
4. I find it difficult to explain to others things that I understand easily, when they don't understand it the first time.	strongly agree	slightly agree	slightly disagree	strongly disagree
5. I dream most nights.	strongly agree	slightly agree	slightly disagree	strongly disagree
6. I really enjoy caring for other people.	strongly agree	slightly agree	slightly disagree	strongly disagree
7. I try to solve my own problems rather than discussing them with others.	strongly agree	slightly agree	slightly disagree	strongly disagree
8. I find it hard to know what to do in a social situation.	strongly agree	slightly agree	slightly disagree	strongly disagree
9. I am at my best first thing in the morning.	strongly agree	slightly agree	slightly disagree	strongly disagree

10. People often tell me that I went too far in driving my point home in a discussion.	strongly agree	slightly agree	slightly disagree	strongly disagree
11. It doesn’t bother me too much if I am late meeting a friend.	strongly agree	slightly agree	slightly disagree	strongly disagree
12. Friendships and relationships are just too difficult, so I tend not to bother with them.	strongly agree	slightly agree	slightly disagree	strongly disagree
13. I would never break a law, no matter how minor.	strongly agree	slightly agree	slightly disagree	strongly disagree
14. I often find it difficult to judge if something is rude or polite.	strongly agree	slightly agree	slightly disagree	strongly disagree
15. In a conversation, I tend to focus on my own thoughts rather than on what my listener might be thinking.	strongly agree	slightly agree	slightly disagree	strongly disagree
16. I prefer practical jokes to verbal humor.	strongly agree	slightly agree	slightly disagree	strongly disagree
17. I live life for today rather than the future.	strongly agree	slightly agree	slightly disagree	strongly disagree
18. When I was a child, I enjoyed cutting up worms to see what would happen.	strongly agree	slightly agree	slightly disagree	strongly disagree
19. I can pick up quickly if someone says one thing but means another.	strongly agree	slightly agree	slightly disagree	strongly disagree
20. I tend to have very strong opinions about morality.	strongly agree	slightly agree	slightly disagree	strongly disagree
21. It is hard for me to see why some things upset people so much.	strongly agree	slightly agree	slightly disagree	strongly disagree
22. I find it easy to put myself in somebody else’s shoes.	strongly agree	slightly agree	slightly disagree	strongly disagree
23. I think that good manners are the most important thing a parent can teach their child.	strongly agree	slightly agree	slightly disagree	strongly disagree
24. I like to do things on the spur of the moment.	strongly agree	slightly agree	slightly disagree	strongly disagree
25. I am good at predicting how someone will feel.	strongly agree	slightly agree	slightly disagree	strongly disagree
26. I am quick to spot when someone in a group is feeling awkward or uncomfortable.	strongly agree	slightly agree	slightly disagree	strongly disagree
27. If I say something that someone else is offended by, I think that that’s their problem, not mine.	strongly agree	slightly agree	slightly disagree	strongly disagree
28. If anyone asked me if I liked their haircut, I would reply	strongly agree	slightly agree	slightly disagree	strongly disagree

truthfully, even if I didn’t like it.				
29. I can’t always see why someone should have felt offended by a remark.	strongly agree	slightly agree	slightly disagree	strongly disagree
30. People often tell me that I am very unpredictable.	strongly agree	slightly agree	slightly disagree	strongly disagree
31. I enjoy being the center of attention at any social gathering.	strongly agree	slightly agree	slightly disagree	strongly disagree
32. Seeing people cry doesn’t really upset me.	strongly agree	slightly agree	slightly disagree	strongly disagree
33. I enjoy having discussions about politics.	strongly agree	slightly agree	slightly disagree	strongly disagree
34. I am very blunt, which some people take to be rudeness, even though this is unintentional.	strongly agree	slightly agree	slightly disagree	strongly disagree
35. I don’t tend to find social situations confusing.	strongly agree	slightly agree	slightly disagree	strongly disagree
36. Other people tell me I am good at understanding how they are feeling and what they are thinking.	strongly agree	slightly agree	slightly disagree	strongly disagree
37. When I talk to people, I tend to talk about their experiences rather than my own.	strongly agree	slightly agree	slightly disagree	strongly disagree
38. It upsets me to see an animal in pain.	strongly agree	slightly agree	slightly disagree	strongly disagree
39. I am able to make decisions without being influenced by people’s feelings.	strongly agree	slightly agree	slightly disagree	strongly disagree
40. I can’t relax until I have done everything I had planned to do that day.	strongly agree	slightly agree	slightly disagree	strongly disagree
41. I can easily tell if someone else is interested or bored with what I am saying.	strongly agree	slightly agree	slightly disagree	strongly disagree
42. I get upset if I see people suffering on news programmes.	strongly agree	slightly agree	slightly disagree	strongly disagree
43. Friends usually talk to me about their problems as they say that I am very understanding.	strongly agree	slightly agree	slightly disagree	strongly disagree
44. I can sense if I am intruding, even if the other person doesn’t tell me.	strongly agree	slightly agree	slightly disagree	strongly disagree
45. I often start new hobbies but quickly become bored with them and move on to something else.	strongly agree	slightly agree	slightly disagree	strongly disagree
46. People sometimes tell me that I have gone too far with teasing.	strongly agree	slightly agree	slightly disagree	strongly disagree

47. I would be too nervous to go on a big rollercoaster.	strongly agree	slightly agree	slightly disagree	strongly disagree
48. Other people, often say that I am insensitive, though I don’t always see why.	strongly agree	slightly agree	slightly disagree	strongly disagree
49. If I see a stranger in a group, I think that it is up to them to make an effort to join in.	strongly agree	slightly agree	slightly disagree	strongly disagree
50. I usually stay emotionally detached when watching a film.	strongly agree	slightly agree	slightly disagree	strongly disagree
51. I like to be very organized in day-to-day life and often make lists of the chores I have to do.	strongly agree	slightly agree	slightly disagree	strongly disagree
52. I can tune into how someone else feels rapidly and intuitively.	strongly agree	slightly agree	slightly disagree	strongly disagree
53. I don’t like to take risks.	strongly agree	slightly agree	slightly disagree	strongly disagree
54. I can easily work out what another person might want to talk about.	strongly agree	slightly agree	slightly disagree	strongly disagree
55. I can tell if someone is masking their true emotions.	strongly agree	slightly agree	slightly disagree	strongly disagree
56. Before making a decision I always weigh up the pros and cons.	strongly agree	slightly agree	slightly disagree	strongly disagree
57. I don’t consciously work out the rules of social situations.	strongly agree	slightly agree	slightly disagree	strongly disagree
58. I am good at predicting what someone will do.	strongly agree	slightly agree	slightly disagree	strongly disagree
59. I tend to get emotionally involved with a friend’s problems.	strongly agree	slightly agree	slightly disagree	strongly disagree
60. I can usually appreciate the other person’s viewpoint, even if I don’t agree with it.	strongly agree	slightly agree	slightly disagree	strongly disagree

Appendix E

Musicality Assessment

Do you or have you ever played a musical instrument? Circle one: Y/ N

If yes, please list the following details about each instrument played:

Name of instrument	Age when started to play	Total years spent playing	Hours spent playing/week	Still playing instrument? Y /N