From: Elsevier - Article Status <Article_Status@elsevier.com>
Date: September 16, 2014 12:02:10 PM EDT
To: <jroche3@kent.edu>
Subject: **Article tracking [SPECOM_2238] - Share your article**

Please note this is a system generated email from an unmanned mailbox.
If you have any queries we really want to hear from
you via our 24/7 support at http://help.elsevier.com

Article title: “Your Tone Says It All”: The Processing and Interpretation of Affective Language
Reference: SPECOM2238
Journal title: Speech Communication
Corresponding author: Dr. Jennifer M. Roche
First author: Dr. Jennifer M Roche
Final version published online: 16-SEP-2014
Full bibliographic details: Speech Communication (2015), pp. 47-64
DOI information: 10.1016/j.specom.2014.07.004

Dear Dr. Roche,

We are pleased to inform you that the final version of your article with full bibliographic details is now available online.

To help you access and share your article, we are providing you with the following personal article link, which will provide free access to your article, and is valid for 50 days, until November 5, 2014

http://authors.elsevier.com/a/1PivLc7UHmkhI

Please use this link to download a personal copy of your article for your own archive. You're also welcome to email the link to your co-authors and colleagues, or post the link on your Facebook, Google +, Twitter or other social media profile, to tell your network about your new publication. Anyone who clicks on the link until November 5, 2014, will be taken to the final version of your article on ScienceDirect for free. No sign up or registration is needed - just click and read!

As an author, you may use your article for a wide range of scholarly, non-

commercial purposes, and share and post your article online in a variety of ways. For more information, please see www.elsevier.com/copyright.

Yours sincerely,
Elsevier Author Support

TRACK YOUR ARTICLE
To track the status of your article throughout the publication process, please use our article tracking service:

http://authors.elsevier.com/TrackPaper.html?trk_article=SPECOM2238&trk_surname=Roche

For detailed article tracking instructions please go to:
http://help.elsevier.com/app/answers/detail/a_id/90

ADVANCING WOMEN
Advancing women in science and libraries in the developing world:
Every year, the Elsevier Foundation provides grants to institutions around the world, with a focus on support for the world's libraries and for scholars in the early stages of their careers. Since 2002, The Elsevier Foundation has awarded more than 60 grants worth millions dollars to non-profit organizations focusing on helping the world's libraries, nurse faculties, and women scholars during their early and mid-careers. Maybe we can help you.

See the latest call for funding applications at: www.elsevierfoundation.org

HAVE A QUERY?
We have 24/7 support to answer all of your queries quickly.
http://help.elsevier.com

SENDER INFORMATION
This e-mail has been sent to you from Elsevier Limited, The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, United Kingdom. To ensure delivery to your inbox (not bulk or junk folders), please add Article_Status@elsevier.com to your address book or safe senders list.

PRIVACY POLICY
Please read our privacy policy.
http://www.elsevier.com/privacypolicy

[T-18-20140301]

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 66 (2015) 47–64

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

"Your Tone Says It All": The processing and interpretation of affective language ☆

Jennifer M. Roche [a,*], Brett Peters [b], Rick Dale [c]

[a] School of Health Sciences, Kent State University, United States
[b] Clinical and Social Sciences in Psychology, University of Rochester, United States
[c] Cognitive and Information Sciences, University of California, Merced, United States

Received 20 September 2012; received in revised form 28 July 2014; accepted 28 July 2014
Available online 14 August 2014

Abstract

Pragmatic interpretation of intent is essential for successful communication. The current studies evaluate the impact of affective prosody on the processing and interpretation of affectively spoken language. A production study provided further evidence of talker variability in the production of the emotionally-laden categories of Innuendo, Irritation, Compassion and Neutral, indicating a great deal of within and between talker variability, as well as talker systematicity within affect categories. Despite this talker variability, in a listening task, participants were asked to categorize the intent of the talkers statements (from the production study) to determine the relative accuracy of responding, while also tracking the perception of intent as it unfolded over time (i.e., via computer Wii-mote x, y coordinates). The results from the online measurement of the perception of intent indicated that even though our listeners were accurate in categorizing intent (~70% mean accuracy), the "dynamic signature" of their responses was laden with a great deal hesitation and indecision for some, but not all talkers. This suggests that during the perception of intent, the cognitive system is flexible enough to handle talker variability, but during perception, uncertainty will change the manner in which the intent is processed.

Keywords: Affect; Speech production; Speech perception; Prosody; Pragmatics; Intent

1. Introduction

The interpretation of intent often goes beyond a single word, and its explicit meaning. Discourse often has embedded meanings that require attention to context and the appropriate decoding of paralinguistic information to facilitate a felicitous response. For example, a listener must not only pay attention to prosodic cues (linguistic and affective), but must also attend to speaker specific cues, in hopes to prompt the listener with a means to appropriately respond given the context and the speaker's intentions.

A great deal of research has focused primarily on talker variability, showing that talkers variably produce speaking rates, have different levels of spoken word intelligibility, have a range in voice quality, are not always systematic in their vowel production (Bachorowski and Owren, 1999; Mullennix et al., 1989; Mullennix and Pisoni, 1990; Pisoni, 1992), and, at the most basic level, have biological differences in the vocal tracts that provide very strong cues to the gender of the speaker (e.g., due to vocal tract length; Goldstein, 1980; Nordström, 1977). Listeners also have little difficulty in determining the race and even age of talkers (Ryalls et al., 1997). Though these cues may not necessarily

☆ This research was supported in part by a grant from the National Science Foundation to Rick Dale (NSF HSD-0826825).
* Corresponding author. Address: Speech Pathology & Audiology, School of Health Sciences, Kent State University, Kent, OH, United States. Tel.: +1 330 672 0244.
E-mail address: jroche3@kent.edu (J.M. Roche).
URL: https://www.jennyroche.com (J.M. Roche).

http://dx.doi.org/10.1016/j.specom.2014.07.004
0167-6393/

contribute to the interpretation of intent, they are likely to guide the listener in deciding how to act upon that intent. In fact, some researchers suggest that the variability that exists between talkers is a prominent and necessary component of speech perception (e.g., Newman et al., 2001).

However, at the pragmatic level, it is possible that other factors, such as cultural experiences, may also strongly shape how interlocutors produce and perceive affect (e.g., Hawk et al., 2009; Ishii and Kitayama, 2002; Kitayama and Ishii, 2003; Kitayama et al., 2006). A speaker's ability (or inability) to produce affect has been shown to produce a negative effect on a listener's ability to perceive affect properly (Mullennix et al., 2002). Therefore, cues to speaker identity and ability to produce affect have important social ramifications with regards to how the listener might address and interact with the speaker.

Since communication is often layered with affective information that interacts with psycholinguistic processes, affective paralinguistic cues may further promote socially acceptable behavioral responses and the understanding of hidden meanings (Attardo et al., 2005; Nygaard and Lunders, 2002). Considering social cues and a speaker's ability to produce affective language may shed light on how we easily (and sometimes not so easily) are able to correctly identify speaker affect in novel situations, and with new people. Therefore, the presence of affective cues in speech should help guide social exchanges, imbued with varying emotions and intent, as it is integrated in a rich social context. Talker variability in spoken word production may help the listener better decode the message, but it is possible that talker variability as it relates to affect perception may make the interpretation of the message more difficult as research has shown that elicited emotions are often a blend of several disparate emotions (Scherer and Ceschi, 1997).

Difficulty may arise during affect speech perception because both the talker and listener must be aware of how the affective cues influence language. If the affective cues are not salient, or are somehow misinterpreted, then conflict may arise (e.g., both parties feeling negatively towards the other because they misunderstood something about the situation). Therefore, it is of particular importance for speakers to pay attention to context and use appropriate cues, in hopes that the listener will properly integrate the relevant cues during the interpretation of intent. Given the centrality of these cues in everyday interaction, it is important to understand the underlying mechanisms involved in processing them. In the current study we capitalize on the fact that speakers produce affect differently and assess how affective talker variability impacts the way listeners perceive the speaker's intent.

2. Background

Communication is often driven by behaviors related to the expression of affect cues. The tendency to respond affectively is important to decrease social distance, and maintain and develop social relationships. Additionally, responding affectively may promote the coordination of social activities, provide cues to others about how to respond in a socially appropriate manner, and may help promote the interpretation of another's behaviors that help regulate interpersonal interactions (Fischer and Manstead, 2008; Fridlund, 1994; Hawk et al., 2009; Keltner and Haidt, 1999; Scherer, 1980, 1988, 1994; van Kleef et al., 2004). The interpretation of affect is often multimodal (e.g., facial, gestural, postural, and vocal; Guerrero and Floyd, 2006). Other sensory modalities are highly interactive among one another (e.g., auditory and visual information during speaking). However, vocal expressions of affect may have general detectability advantages over the other modalities, because their expression has the ability to draw attention "omni-directionally and over long distances" (Hawk et al., 2009; pp. 294). In the current paper, we focus on this auditory channel during higher-level spoken language acts that require the interpretation of intent beyond the literal meaning of the words spoken.

Since "we don't always say what we mean, or mean what we say," (Galloway, 1974), we may rely on the vocal cues to disambiguate our intentions (e.g., Attardo et al., 2005; Nygaard and Lunders, 2002). A significant amount of work has been conducted to evaluate how we produce and perceive affective cues in speech, and motivates our studies here. During vocal production, affect has been evaluated based on a number of emotion/affect categories (e.g., ranging from basic emotions to more subtle pragmatic categories like sarcasm; e.g., Cheang and Pell, 2008; Rockwell, 2000; Scherer, 1986, 2003; Scherer and Banziger, 2004). The categories have been extensively evaluated for their relevant acoustic correlates across talkers during the production of single-word utterances (e.g., see Bachorowski, 1999; Banse and Scherer, 1996; Leionenen et al., 1997; Scherer, 2003; Scherer and Banziger, 2004), in addition to a number of studies evaluating nonsense sentential structures (e.g., Banse and Scherer, 1996; Scherer et al., 2010).

Studies of affective prosody are usually carefully controlled for lexical and semantic content. The evaluation of single words is practical, because single words carry the majority of the affective prosodic variation and it has also been shown that most of the affective information is carried in the vowel (Kaiser, 1962). However, the interpretation of affective speech in natural settings minimally, at best, involves the integration of lexical, semantic and prosodic content towards the interpretation of intent (e.g., for a review of natural vocal expression see Scherer, 2003). Here we consider that a single word in an utterance may carry a greater degree of affective prosody, but the surrounding words (with the interaction of their meaning) may contribute to and also have prosodic markers necessary to decode intent, especially when contextual cues may be less salient (e.g., on a cell-phone, which could require listeners to compare featural information held in the pre-categorical acoustic sensory store; Crowder and Morton, 1969; MacMillan et al., 1988).

The interpretation of intent is highly interactive and may be greatly affected by non-linguistic vocal cues on language (i.e., meaning or purpose behind the affective expression), as it interacts with the peripheral affective cues. Initially, some suggestions have been made that paralinguistic and linguistic content are processed independently of each other, suggesting affect is added noise that will only hinder the encoding of the linguistic information (Forster, 1979; Massaro and Cohen, 2000). However, this is by far not the common viewpoint among emotion researchers. Emotion researchers, unlike most traditional linguists and psycholinguists, advocate the importance of paralinguistic cues to understand linguistic content (e.g., Ladd et al., 1986; Majid, 2012; Nygaard and Lunders, 2002; Nygaard and Queen, 2008; Scherer et al., 1984). Rightly so, individuals are faced with a great deal of variability in conversational settings on a regular basis and may in fact preserve all possible cues (e.g., semantic, prosodic, social, and cultural) to form "exemplar-based representations" that promote perception and aid production (i.e., for a review see Nygaard and Queen, 2008). Most traditional linguistic perspectives have often neglected the important paralinguistic cues to language production, leaving affect and emotion to be studied separately (as described by Nygaard and Lunders (2002)). Integrating the two (affect and language), as they exist in the natural environment, have lead to new and interesting ways to consider the cognitive mechanisms behind affect production and perception.

For example, Nygaard and Queen (2008) and Nygaard and Lunders (2002) have found that words spoken with affective intonations are better remembered. Halberstadt et al. (1995) also found that individuals are faster at making lexical decisions when affective information is congruent with a spoken word. In fact, affective prosody may provide important cues that allow interlocutors to disambiguate speech and understand language (Martinez-Castilla and Peppe, 2008; Morton and Trehub, 2001).

Relatedly, Egidi and Nusbaum (2012) found increased N400 responses to spoken language that had incongruent story resolutions (e.g., positive story with a negative outcome). This is a particularly interesting finding because the N400 response is associated with semantic processing of incongruent information. This suggests that affect is processed in conjunction with the meaning. Schirmer et al. (2005) report that listeners showed smaller N400 responses to congruent language/prosody for joy and sadness relative to incongruently spoken joyful or sad words (e.g., happy word, sad tone of voice). Egidi and Gerrig (2009) also provided evidence that negatively valenced stories took longer to process than positively valenced stories during reading comprehension. The findings from these studies suggest that the effect of affect on language is readily integrated during natural language processing. As provided by these examples, evaluating the interaction between language and affect is an important and a very real phenomenon. Thus, focusing on affective contribution to single words and nonsense sentences may miss the richness in a signal for decoding intent indexed by individual talker variability. These interactions of intent and variability may drive perception to appropriately interpret intent in social discourse (as seen in a review by Bachorowski (1999) and Attardo et al. (2005)).

In the current paper, to tap into real-time processing of intent, we utilize a relatively new method for tracking responses semi-continuously while participants process language. Recently, tasks that use computer-mouse cursor movements are being used to unveil online processing of language, and there has been some suggestion that there is a perception–action link associated with computer-mouse movement data during learning, language comprehension, and other tasks (see primarily Spivey et al., 2005; see also Dale et al., 2007, 2008; Farmer et al., 2007; see Freeman and Ambady, 2010 and Freeman et al., 2011, for reviews). Some have argued, using evidence from this and related research, that there is a continuous flow of information as cognitive processing unfolds (Spivey, 2007). In particular, the patterns of responding via computer-mouse trajectories could reveal that cognitive competition is present during processing, such as between two possible response options during language processing (e.g., as in a specific sentential parse; see Farmer et al., 2007). Whether or not processing is continuous, the various velocity and complexity measures obtained from computer-mouse trajectories provide a useful description about these processes as responding unfolds. We use this method here. If participants' responses to a speaker's intent are based on variability in affective prosody between and within talkers, then there may be interesting "signatures" of processing in the response dynamics, thus serving as a window into how listeners comprehend intent.

Many of the studies reviewed here have made a significant contribution towards the understanding of affect as it relates to language production and comprehension. However, many methods have evaluated single word and nonsense sentential processing primarily within the *basic* emotion/affect categories (e.g., see Bodenhausen and Moreno, 2000; Juslin and Laukka, 2001, 2003; Liscombe et al., 2003; Scherer, 2003; Swerts and Hirschberg, 2010) during offline processing. The purpose of the current study is to go beyond the offline measures and basic emotion categories to evaluate the effects on listener comprehension as they interpret speaker intent (i.e., while adapting to talker variability) during real time processing. In our studies, participants process whole sentences that are loaded with different affective intents. By utilizing the mouse-tracking measure, we explore the manner in which this comprehension process unfolds.

We begin with exploring the prosody of statements imbued with affective cues signaling intent to examine both talker variability and talker systematicity (similarities) across talkers within our stimulus set. That is, all talkers produce some variability in their expressions, but should also maintain some affective and linguistic systematicity to make their utterances more interpretable by their listen-

ers. We then look at how listeners process these statements (produced by different talkers) by using them as stimuli in a separate perception experiment. The purpose of the perception experiment is to determine (1) the relative accuracy listeners exhibit while decoding intent between and within multiple talkers (offline), (2) how affective language perception unfolds over time (online), and (3) if the perception of intent is differentially processed between and within talkers (cognitive mechanisms). Our overall goal in this paper, consistent with our background review above, is to provide further evidence that affective prosody is not simply noise in the linguistic signal, and that listeners may differentially handle talker variability while interpreting intent. We specifically find that listeners *can* categorize pragmatic intent without any visual cues whatsoever, based purely on the processing of acoustic, affective cues.

Though basic emotions have been studied most often, for the purposes of the current study, *Neutral*, *Compassion*, *Irritation* and (*Sexual*) *Innuendo* were chosen because they: (1) span a range of positive, negative, and neutral valences, (2) embody a range cognitive appraisals, (3) have a clear perlocutionary effect, and (4) have an abstract nature and potentially pragmatic interpretation (require going beyond the literal interpretation; Austin, 1962) in the absence of context. *Irritation* is negatively valenced and unpleasant, often impedes a desired goal, creates a feeling of hopelessness (lack of control and low probability of obtaining a goal), and has a perlocutionary effect of portraying displeasure of the action of another (or event), with the hopes to dissuade the current behavior (Grundy, 2008; Sacharin et al., 2012). *Compassion*, is positively valenced, implies a position of greater control and power to help another, and is intended to portray support, understanding and concern within a particular context (e.g., act of consoling, in hopes to reduce sadness; Sacharin et al., 2012). *Neutral* provides a baseline, with no clear perlocutionary effect. *Innuendo* has a context-dependent valence (can be positive or negative), while having a prelocutionary effect of expressing interest in hopes to procure a sexual partner but with a means to save face if it is not received well by another (Bell, 1997). Each of these emotion-laden categories were produced by four talkers in the context of five neutrally valenced sentences (e.g., "The elephants carried the supplies.") to ensure the perception of these statements were due to the prosodic contribution and not the linguistic contribution; Seibert and Ellis, 1991).

Scherer and colleagues proposed a notion of push and pull factors in emotion expression (Bänziger and Scherer, 2007; Banziger et al., 2012; Scherer, 1988). Push refers to the involuntary and internal state changes (as related to vocal responses, for example) that occur during emotion expression as represented in the outward expression of emotion. Alternatively, pull factors are voluntary responses to emotional expression that may be exhibited in order to express a particular perlocutionary effect. Therefore, the purpose of having our talkers produce affect in the current paradigm was to evaluate the *pull* factor of emotional expression as a means to portray the intention in a pragmatic form, but not necessarily the internal state changes within the talker (i.e., we make no assumptions that our talkers ever felt compassionate, neutral, irritated, or aroused by innuendo).

We performed acoustic analyses on these expressions from four talkers, but make no attempt to generalize these acoustic cues to a general featural representation of each of these emotion-laden categories, or make any new statements regarding talker variability. In order to do this, we would need a larger scaled production study to speak directly to the acoustic properties of *Innuendo*, *Irritation*, and *Compassion*. However, the interpretation of these types of intentions clearly goes beyond what is literally stated and has very real social consequences that often require the perceiver to respond in a particular manner. This is not to say that the basic emotions would not have clear social consequences, but there is a level of complexity that is inherently interesting about these categories. Specifically, it could be that the production and perception of these expressions may specifically be shaped by cultural and social experiences, forcing the listener towards a pragmatic interpretation, whereas the basic emotions may indicate cues about the internal state of the listener and not necessarily adding any pragmatic information (Ekman, 1992).

Our perception results speak directly to this hypothesis. As a sanity check, we confirm that our talkers do in fact exhibit a great deal of talker variability, but also show systematicity which may be necessary cues listeners use to successfully process the statements (see acoustic analysis). In other words, each talker was not "created equal." The reason listeners may not have responded equally across talkers, may be directly related to the speaker's ability to produce these affective expressions. Specifically, the dynamics of the perceiver's responses reveal interesting patterns of comprehension during successful and unsuccessful interpretation of intent. We provide evidence that when listeners respond accurately (on average, above 70%, well above chance), the listener will respond to the type of affective expression and the talker differentially. For example, listeners had the most difficulty with statements imbued with *Compassion*, especially for one talker (Female 2). On average, Female 2 received an average accuracy score of about 55%, with Neutral receiving the highest average of accuracy (above 80%) and lowest accuracy for Compassion (around 35%, with 25% being at chance). Additionally, when participants responded correctly to *Compassion* they hesitated more, suggesting they may have trusted the saliency of the acoustic cues less for compassion.

3. Analysis of affective prosody

Scherer (2003), among others, has evaluated cues related to affective expressions by examining their acoustic correlates. Some prosodic cues used to communicate affective meaning beyond literal language are changes in pitch, speech rate, and intensity (e.g., Banse and Scherer, 1996;

Bachorowski, 1999; Hammerschmidt and Jurgen, 2007; Mozziconacci, 2001; Scherer, 1986, 2003). The various combinations of acoustic features provide an extra source of information about the intentions of a speaker (Scherer, 2003). These measures are most common within the affective speech literature, but researchers have also evaluated other measures related to F_0, amplitude, and spectral cues (e.g., jitter: frequency perturbation, and shimmer: amplitude perturbation; harmonics; Bachorowski, 1999; Scherer, 2003). Since we do not intend to directly discuss the causes or implications of acoustic variability between talkers directly, but intend to show how listeners handle such variability, we only evaluate F_0, amplitude and duration cues (since these are most prominent in the literature).

Additionally, single syllable/word utterances and nonsense sentences have been primarily used in similar production studies (e.g., as noted by Bachorowski, 1999). However, *real* sentence-long utterances may interact with prosodic cues in a more meaningful way, beyond that found with single-word utterances and nonsense sentences. Therefore, it should be considered that the overall communicative utility of the affective cues as it relates to each word within the span of a sentence might provide a richer context of interpretation. Evaluating the acoustic variability within a sentential context should help us capture the dynamic unfolding of the acoustic signal in a larger unit of the communicative medium.

Systematic trends of acoustic cues have been reported for basic emotions (e.g., vocal joy: an increase in intensity, pitch and duration). This description provides insight into the acoustic make up of a specific *basic* affective category (basic emotion categories: joy, sadness, fear, surprise, anger and disgust; Ekman et al., 1972), but the description does not generally provide information about individual talker variability and speaker intentions, especially as it relates to pragmatic meaning during naturalistic communication. The purpose of the production analysis is to verify the existence of talker variability and provide an initial description of the acoustic cues that may be related to the affective categories chosen as they are tied to sentential meaning and affective cues (specifically for *Compassion*, *Neutral*, *Irritation* and *Innuendo*). It should be noted, that the evaluation of these categories are only based on four talkers, and the purpose of the current study is not to generalize specific acoustic cues to an account of the featural properties of *Compassion*, *Innuendo* and *Irritation*, however we do provide a description of the prosodic aspects of each affective category (see above). Therefore, the current analysis is intended to set the stage for a larger scale evaluation of the acoustics related to these affective categories.

4. Methods

4.1. Participants

Four individuals from the city of Memphis volunteered for participation (2 males and 2 females; mean age: 28.5 years). Participants provided self-reports that they were native speakers of American English, with a southern accent and no reports of diagnosed speech or hearing impairments.

Table 1
Normed neutral statements selected from the Seibert and Ellis (1991) study.

Production study statements
1. Elephants carried the supplies
2. You have to take the ferry to get to the island
3. The Pacific Ocean has fish
4. There are 60 min in 1 h
5. Most oil paintings are done on canvas

4.2. Materials

The talkers were placed at a comfortable viewing distance from a 20″ iMac computer screen. A standalone microphone (MXL 990 Condenser Mic) was placed directly in front of each talker and was used to collect auditory recordings.

4.3. Stimuli

4.3.1. Sentences

The sentential stimuli chosen were selected from a set of statements normed for neutrality (see Table 1, Seibert and Ellis, 1991). Each statement was randomly presented to each of the talkers. Non-valenced (i.e., neutral) sentences were chosen so that the affective prosody changed the interpretation, but the lexical content did not. We did not want the words in each statement to interact with the connotations, as we attempt to evaluate how the interpretation of language changes as a function of the acoustic properties of affect.

The sentences chosen included five non-valenced statements, spoken with four affective intents (5 statements × 4 Affective Intents [*Compassion*, *Innuendo*, *Irritation* and *Neutral*] × 4 Talkers [2 males and 2 females] = 80 productions, 20 per talker).

4.3.2. Images

Images were found that closely matched the sentential meaning. For example, an image was searched, via medium sized Google images[1] that closely matched the context of the statement "Elephant's carried the supplies." This was important to help the talker "imagine" a context in which this statement could *work*. Each image was resized to 300 × 400 pixels using Adobe Photoshop before

[1] The use of IAPS (International Affective Picture System; Lang et al., 2008) was originally considered. However, the researchers were not easily able to find images that closely matched the sentential context and affective expression. Therefore, the researchers resorted to finding Google images that were closely related to the sentential and affective context.

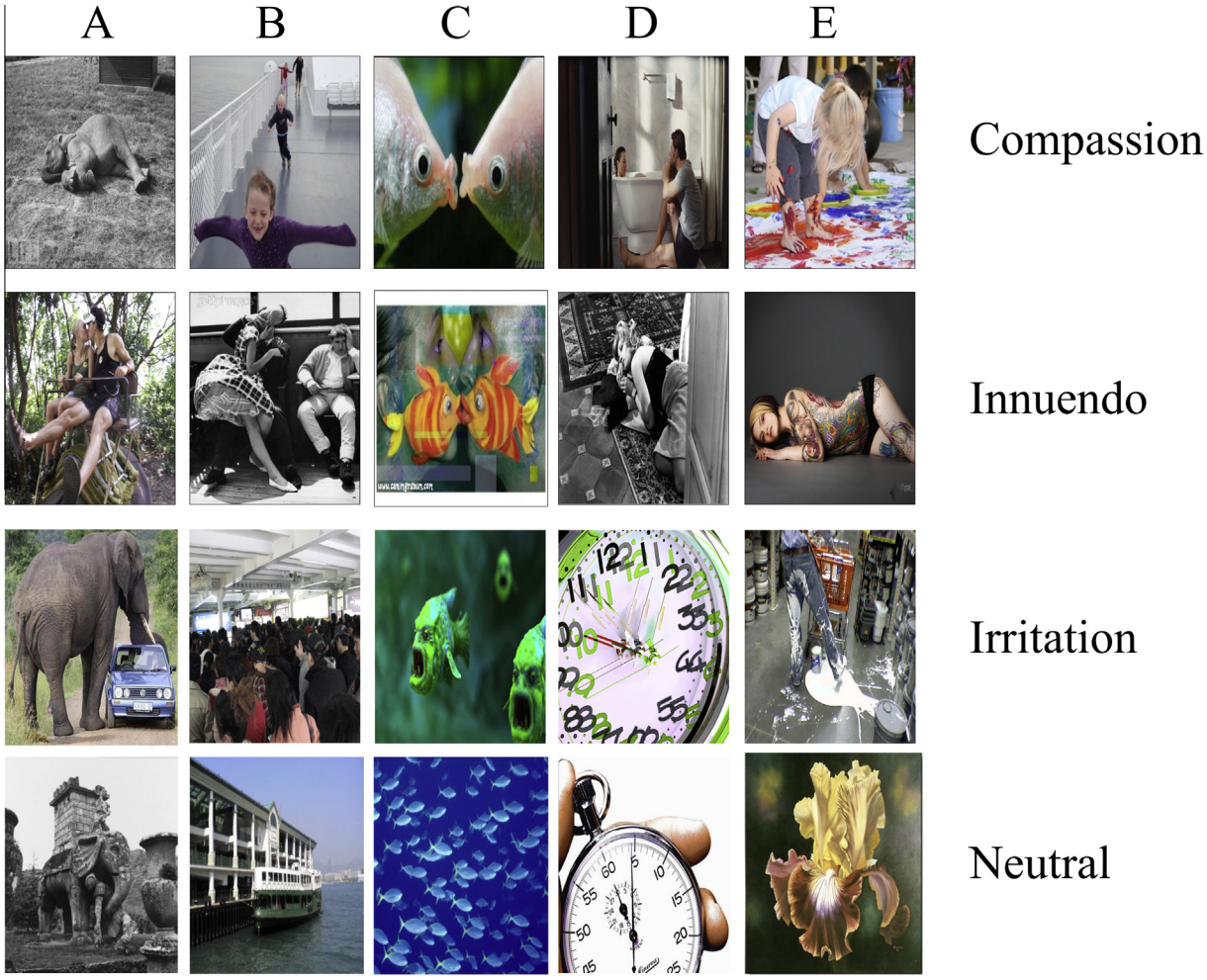

Fig. 1. Images used in the production study. (A) Elephants carried the supplies. (B) You have to take the ferry to get to the island. (C) The Pacific Ocean has fish. (D) There are 60 min in 1 h. (E) Most oil paintings are done on canvas.

presentation in the middle of black background on a 20″ Mac computer screen (see Fig. 1 for each image)[2]

4.3.3. Norming study

Each of the images paired with the statement by affective expression was then presented to paid participants via Amazon's Mechanical Turk (an online system that pays individuals to complete various types of data entry tasks to be normed for within category representation). Mechanical Turk has been demonstrated to produce reliable responses from its users (Snow et al., 2008; Sorokin and Forsythe, 2008; Gibson et al., 2011). Participants were asked: "How well could you say this sentence, in this tone of voice, while looking at this image?" Fifteen Mechanical Turk users rated each of the images on a 5-point Likert scale (1 = very well to 5 = not very well at all). Each of these ratings was submitted to a reliability analysis and the intraclass correlation, also known as the reliability coefficient, was used to evaluate how strongly the responses provided by the MTurkers resembled each other (ICC = .921 for all images, high ICCs represent strong agreement for the images selected; Russ et al., 2008; see Table 2 for mean ratings for each of the images). The results from the reliability analysis suggest that each of the images chosen fits well within the category expected.

5. Procedure

The talkers were separately and randomly presented with each of the statements by image by affective intent, via a MATLAB PsychToolBox-3 program (Brainard, 1997). Stimulus presentation included an affect-inducing picture, to help prime the talker with the affective expression to be spoken. Text labels and statements were presented above and below each image (see Fig. 2 for an example).

Talkers were never asked to guess the affect, but were explicitly told which expression to produce (on the experimental screen in text format). The image was presented for the sole purpose to help the talker visualize a scenario in

[2] Although we acknowledge that the variability produced in experiment 1 and listened to in experiment 2 may be due to the differences between the pictures, we suspect this variation to be minimal.

Table 2
Mean ratings obtained for the images for the selected sentences from the Mechanical Turk analysis [Means(SD)].

Statement	Intent	Mean(SD)
Elephants carried the supplies	Compassion	2.07(1.22)
	Innuendo	1.87(1.06)
	Irritation	1.80(1.42)
	Neutral	1.80(1.32)
Most oil paintings are done on canvas	Compassion	2.33(1.45)
	Innuendo	1.47(0.64)
	Irritation	1.73(1.16)
	Neutral	1.53(0.83)
The Pacific Ocean has fish	Compassion	1.87(1.25)
	Innuendo	2.00(1.00)
	Irritation	1.53(0.74)
	Neutral	1.33(0.72)
There are 60 min in 1 h	Compassion	2.00(0.93)
	Innuendo	2.73(1.53)
	Irritation	1.87(1.41)
	Neutral	1.27(0.59)
You have to take the ferry to get to the island	Compassion	2.73(1.49)
	Innuendo	2.00(1.07)
	Irritation	2.00(1.41)
	Neutral	1.47(0.92)

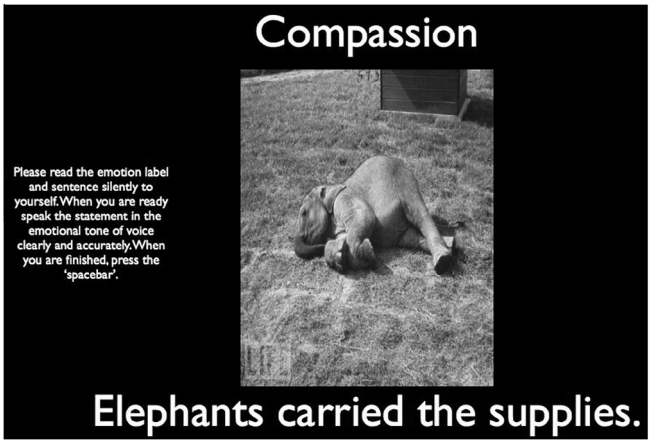

Fig. 2. Sample screen and affective inducing images that participants in the production task viewed.

which the affective statement could be produced. None of the talkers reported they had difficulty with this.

At the time of production, the talker was asked to pause, make note of the expression, view the image and read the statement silently. Once the talker felt comfortable with the affect label, statement and image, s/he was instructed to vocally produce that statement. The talker was asked to clearly produce each statement as naturally as possible based on his/her own interpretation of the expression. The instructions for production were designed as a means to elicit a more natural sounding affective production. Additionally, we wanted non-actors to produce each of these statements. Although having actors express these emotions has its advantages (Bänziger and Scherer, 2007), the current research was not designed to perfectly model *Compassion*, *Innuendo*, *Irritation*, and *Neutral* expressions, but rather to capitalize on the between and within talker variability in the vocal expression of these emotions using laypersons, because this might better represent what a listener experiences in the natural environment. Each statement was recorded at a 44.1 kHz, 16-bit CD quality, sampling rate.

6. Results and discussion

A multinomial logistic regression (Croissant, 2007) was used to determine how talkers' affective expressions were marked by a set of three acoustic characteristics (duration, F_0, and amplitude). The odds ratios are reported, as a measurement of association between the acoustics and the intended produced category. The odds ratios can be interpreted as the regression coefficients that are converted from log odds into odds ratios, and thus explaining the odds of the acoustic variable is associated with a particular affective intent.

Acoustic values were obtained via Praat, a synthesis and re-synthesis speech software package (Boersma and Weenink, 1992). Duration was sampled every 25 ms, providing time stamps throughout the entire voiced production (e.g., 0 ms, 25 ms, 50 ms, 75 ms, 100 ms...3000 ms) with the pitch and loudness measures collected at each step. Upon evaluation of each of these measures, visual inspection of F_0 resulted in outlier removal of any values above 500 Hz and below about 90 Hz (resulting in removal of about 6.5% of the data). F_0 values were then normalized across talkers, in order to control for biological aspects of pitch production, but were used to evaluate local levels of pitch differences produced as a function of the affective intent (e.g., size of vocal tract between males and females; Goldstein, 1980; Nordström, 1977). Norming involved taking the average F_0 for the female and male talkers (Female mean: 203 Hz, Male mean: 134 Hz). The mean F_0 values were then subtracted and the difference was added to each of the males F_0 values at each duration time step. Loudness, as a measure of intensity (in dB SPL) was also collected at each of the duration time stamps (see Table 1 for a list of the raw acoustic descriptives).

The multinomial analysis revealed that the categories based on Talker and Intention could be significantly differentiated by duration, F_0, and intensity (amplitude measure), (Talker: $\chi^2 = 2144.2$, $p < .001$; Intention: $\chi^2 = 465.25$, $p < .001$; Talker × Intention: $\chi^2 = 2144.2$, $p < .001$; see Appendix A for a detailed list of significant differences between Talker by Affect combinations; and see Fig. 3 for a 3-D representation of the acoustic space for each Talker and Talker intent). Since the purpose of the acoustic analysis was not to make generalizations about the acoustic properties of the affective categories of interest, the results from the regression model are provided in Appendix A

and only a general interpretation of the results are discussed.

Overall, the findings from the acoustic analysis suggest that while there may be some global affective markers in which our sample of talkers based their productions, much like those described for in the current literature (e.g., Banse and Scherer, 1996; Bachorowski, 1999; Hammerschmidt and Jurgen, 2007; Hawk et al., 2009; Scherer, 2003), as well as individualistic factors that influenced the final production. In that, the results suggested that there were some productions that did not significantly differ, as well as significant differences between and within talkers. This final outcome may not represent a clear-cut acoustic category for all talkers and talkers may produce a continuum of acoustics that may fall within a particular type of context. Therefore, even beyond the biological make-up of the individual, cultural learning and dialect might significantly influence the nature of an affective production and how listeners perceive such cues (Hawk et al., 2009; Ishii and Kitayama, 2002; Kitayama and Ishii, 2003; Kitayama et al., 2006). Therefore, the next natural question was, "Will listeners still be able to use the cue variation, exhibited by our talkers to match affective cues systematically, regardless of the variability between talkers, for each of the different affect categories?"

7. Perception and action study

A number of studies suggest that listeners have the ability to detect differences in affective prosody (e.g., Nygaard and Lunders, 2002; Nygaard and Queen, 2008; Scherer, 2003; Scherer and Oshinsky, 1977). Studies evaluating the accuracy and the categorization of affective intent have varied between studies, but overall, the suggestion is that listeners can identify affective categories with above chance performance when the affect cues are represented as global markers of a specific category (e.g., Banse and Scherer, 1996; Bachorowski, 1999; Hammerschmidt and Jurgen, 2007; Hawk et al., 2009; Scherer, 2003). However, many of these studies only evaluate the ability of listeners to decode intent behind single-word utterances and nonsense sentences, which could possibly miss the interaction between sentential meaning and the interpretation of intent beyond the literal meaning (for example see Leionenen et al., 1997; Bachorowski, 1999). Also, many of these studies evaluate affect categorization after processing has already occurred, which misses the dynamic process at the time of perception. The purpose of the current study is to determine how well listeners actively and dynamically decode affective intent, which has clear prosodic variation, beyond explicit sentential meaning.

At the time of perception, we sampled Wii-mote (i.e., similar to sampling computer-mouse x, y coordinates with the Wii-mote, a Nintendo Wii console controller) movements in a semi-continuous manner to evaluate how listeners processed the affective speech during the online processing of intent. For example, arm movement measures provide variables that assess online processing of responses, including where the cursor is on the x and y axis of the computer screen. This provides information about the time course of processing. For example, if the affective cues were processed post-perceptually, we would expect that listeners would wait until the statement is finished before they make their choice (for a review see Massaro and Cohen, 2000). Alternatively, if affect is processed early on in the perceptual system (almost as quickly as the linguistic input), we should see the x and y coordinates moving in the direction of the *correct* answer before the end of the statement. Additionally, an increase x-flips and y-flips, distance and response time might indicate processing difficulties (see Fig. 4 for a hypothetical example).

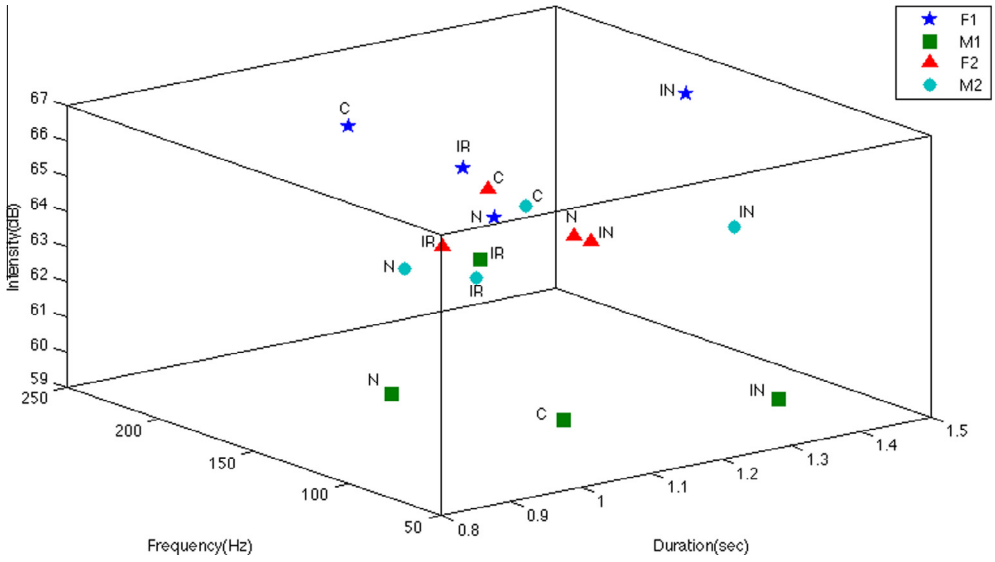

Fig. 3. 3-D representation of the acoustic space for the Talker by Intention (C = Compassion, IR = Irritation, IN = Innuendo, N = Neutral; F1 = Female 1, F2 = Female 2, M1 = Male 1, M2 = Male 2).

x and *y*-Flips represent flipping arm direction along the *x*- or *y*-axes (e.g., zig-zag or up-down pattern). These measures represent indecision on the part of the perceiver, where they are changing directions of their arm movements towards or away from the alternative response options (Dale et al., 2008). Distance refers to the total area of movement within the screen in pixels, which could also represent hesitation to choose a response option (Spivey et al., 2005). Response time was measured in milliseconds (ms), to represent how long it took participants to make their response. Each of these measures provides information about how listeners perceive the environment during moment-to-moment updates of processing the spoken stimulus.

Participants should have the ability to correctly categorize intent based on prosodic changes between and within talkers. Traditional evidence would argue that listeners are focusing on global cues to affective categories. However, if listeners differentially process talker variability, this might indicate that listeners are not only accessing global characteristics, but also integrating the speaker specific, *localized* cues. This will be demonstrated via bodily actions marked by acoustic cue changes in the signal, which should mirror online cognitive processing of such information. In sum, evaluating action dynamics during the online processing may provide evidence towards the mechanisms behind the perceptual processing of affective intent from various talkers.

8. Method

8.1. Participants

Participants included 24 undergraduate student volunteers from the University of Memphis. Participants provided self-report that they were native speakers of American English with no diagnosed vision or hearing impairments (mean age = 20.13 years; 19 females).

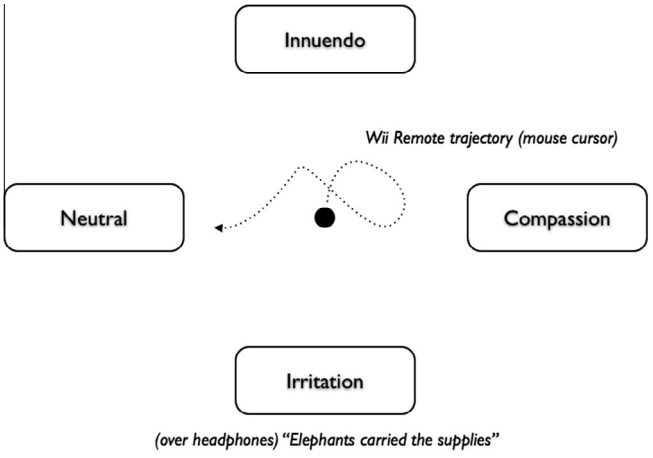

Fig. 4. (Hypothetical) example of an experimental trial for the statement "Elephants carried the supplies." The example here shows how the Wii trajectory measures of distance (number of pixels covered), *x* (zig-zag) and *y*-flips (up/down) may look for an innuendo statement.

8.2. Materials

The experiment took place in a private laboratory room. An Epson LCD projector was placed on a 30-in. high table. This projected an Apple Mac mini's display onto the wall at the end of the laboratory room (12 ft × 5 ft). The projection screen was approximately 5.5 feet in width (29.1° visual angle). The participant was positioned behind the LCD projector table, approximately 8 ft. away from the projection. A Nintendo Wii-remote was used as a wireless, arm-extended pointing device (i.e., very much like a wireless mouse that allows for less constrained arm movements) by having it communicate with a computer equipped with the Bluetooth transfer protocol (see Dale et al., 2008, for methodological details). At the base of the projection screen was a Nyko infrared emitter. Like the Wii console's sensor, this provided the Wii-remote a frame of reference for computing cursor position and a Macintosh framework called DarwiinRemote (© 2006, Hiroaki Kimura) accomplished the interfacing (see Fig. 5 for an example of the experimental set-up). Since the experiment was not performed in a sound attenuated chamber, the auditory stimuli were played over noise cancelling headset in order to reduce the amount of ambient noise that could potentially distract participants during the experiment (Razer™ Barracuda, gaming headset with the microphone removed). A MATLAB PsychToolbox-3 program controlled stimulus presentation and participant response collection (see Fig. 5 for an example of the virtual button display).

8.3. Stimuli

Stimuli included the 5 statements by 4 affect categories by 4 talkers productions from the acoustic analysis discussed above, resulting in 80 statements per talker. These statements were equated for average RMS amplitude to control for shifts in loudness between stimulus productions. This was a measure to globally equate the overall amplitude envelope across the entire stimulus (similar to turning the volume up for the stimuli produced at an overall lower amplitude). This method preserved the localized fluctuations in amplitude within the signal, as a means to set a comfortable listening level for all stimuli, because some talkers were closer to the microphone during production than the others.

9. Procedure

Participation included a categorization task that was designed to assess how affective prosody influenced participants' perception of intent. Participants were not given explicit instructions on how to determine intent. During the experiment, participants were presented with a sound file of one of the talkers by intent statements discussed

Fig. 5. An example of an experimental set-up with the virtual button display projected on the wall.

above in the acoustic analysis section. As the sound file was playing, participants were instructed to make their response as soon as they knew the *answer* (even if the sound file hand not finished playing). To make a selection, the participant would use the Wii-cursor to point and click on one of the virtual display buttons (see Fig. 5 for an example of the virtual button display).

The categorization task was composed of 3 blocks of experimental trials.[3] A fourth block included a set of questions about the perceived demographics of the talkers. Questions in this block included: (1) "What was the gender of the talker?", (2) "What was the race/ethnicity of the talker?", (3) Does the talker talklike you talk?" (identifiable) and (4) "How expressive did you feel the talker was?" (see Appendix B for the results from these questions). As a note, questions regarding race/ethnicity and gender were collected to determine if our listeners could correctly identify these characteristics of the talkers (as indicated by the literature; the listeners did well at these judgements: see Appendix B). Alternatively, expressivity was measured to determine if the listeners were sensitive to the expressiveness of the talker, which could influence the judgements of intent (expressivity was measured using a 5 point Likert scale: 1 = very expressive to 5 = not at all; mean talker score: 2.4). Additionally, we asked the listeners to judge how identifiable the talker was by asking if they could identify acoustically with the talker (i.e., "Does the talker talk like you?", this was also measured on a 5 point Likert scale: 1 = not at all to 5 = very much, mean talker score: 3.25). The blocks included randomly presented statements from each talker from the production study, but the virtual button locations changed per block (i.e., top, bottom, left and right respectively; Block 1: Neutral, Compassion, Innuendo, Irritation; Block 2: Irritation, Innuendo, Neutral, Compassion; Block 3: Compassion, Innuendo, Neutral, Irritation). Button locations changed between blocks to control for arm trajectory biases towards a particular portion of the screen (McKinstry et al., 2008). Participants were also permitted to take a brief rest break between blocks to reduce fatigue effects. During the span of the experiment, participants received a total of 480 trials [160 trials per block; 5 statements × 4 expressions (*Compassion*, *Innuendo*, *Irritation*, and *Neutral*) × 6 stimulus repetitions × 4 speakers (Female 1, Female 2, Male 1, Male 2)]. Stimuli were played over a headset and responses were made via a Wii-remote click on a virtual button display, projected onto a wall, which corresponded to the perceived target expression. The intent response categories were divided into 4 regions to allow a larger region of measure for arm movements while the participant made their responses and the button locations changed per block (e.g., the *Compassion* button was not always at the bottom of the screen).

Arm movement measures were semi-continuous measures collected during online processing of responses and included x-flips, $x_{100–400}$ ms, y-flips, $y_{100–400}$ ms, distance and response time. As stated above, x-flips represent flipping arm direction along the x-axis (e.g., zig-zag pattern). $x_{100–400}$ ms, a semi-continuous measure, refers to the cursor position along the x-axis during the first 100, 200, 300 and 400 ms of stimulus onset, to represent when and if participants started moving towards a the talker's intended response option. y-flips and $y_{100–400}$ ms are similar to x-flips and $x_{100–400}$ ms, but the data came from the y-axis. As stated previously, distance refers to the total area of movement within the screen in pixels. Response time was measured in milliseconds, to represent how long it took participants to make their response.

Each of these measures represents varying forms of complexity that could represent indecision or certainty (i.e., distance, x- and y-flips), as well as the time course of perceptual processing [i.e., response time, $x_{100–400}$ and $y_{100–400}$). Specifically, hesitation during a response action may result in a larger number for x-flips, y-flips and distance, representing more indecision. Shorter response times may indicate that listeners are readily and actively processing the acoustic information, especially if response time is shorter than the overall production of the statement. If there is evidence of movement on the x and y coordinates within the first 400 ms of dynamic statement, this suggests that affect is being processed relatively early in the stimulus. Each of these variables together will provide information about the nature of the perceptual processing of the affective statements.

The predictions for the perception and action portion of the experiment will address two points. First, time course of processing intent should happen early within the signal. It has been suggested that affect should not be processed as quickly as linguistic information, however if listeners are

[3] Note: a 4th block of experimental trials was not used to reduce the length of the experiment, in order to reduce fatigue effects from standing for 1.15 h.

able to accurately decode intent before the end of the statement this might suggest that affect is integrated early and immediately along with and potentially before the ending of the linguistic signal, as opposed to post-perceptual processing (de Gelder et al., 2000; Massaro, 1998; Massaro and Cohen, 2000). This would translate into rapid statistically significant changes in our arm-movement measures—showing that participants are starting to make systematic responses early on in sentence stimuli. Additionally, we predict that individual differences in talker cues may be responsible for differential processing exhibited by arm trajectory measures from listeners. Specifically, there should be evidence that if a listener is less confident in a speaker's intent, it should show up in their bodily movements (e.g., more x and y flips, longer response times, more distance covered [in pixels]).

10. Results and discussion

Two mixed effect logistic regression models were used to evaluate if the Talker, Affective Intent, the Wii variables [response time, distance, x-flip, y-flip, $x_{100–400}$ and $y_{100–400}$] and the Talker Acoustics (F_0, amplitude, and duration) predict the proportion of accuracy (i.e., the proportion of target intent that was selected by the listener) for (1) All Talkers (Model 1), and (2) Female 2 removed from the model[4] (Model 2), which include action dynamics of all correct and incorrect trials. This was done primarily to see the overall action dynamics and acoustic contribution as a function of accuracy. As a note, Talker and Talker intent were modeled as random slopes with Subject added as the random intercept. Mixed effects logistic regressions were used because the proportion of accuracy was estimated via the mixed effects logistic regression based on the categorical variable of accuracy ($1 = correct$ intent and $0 = incorrect$ intent). Jaeger (2008) explains that running an ANOVA on categorical data may sometimes produce spurious effects. Therefore, a mixed logistic regression was used to model the *Talker, Talker Intent* and *Subject* as a random effects, while appropriately manipulating the categorical outcome data as a discrete outcome. Finally, an additional multinomial logistic regression model (Model 3) was used to evaluate how the Wii trajectory and acoustic measures characterized the *correct* responses as a function of talker and talker intent, in attempts to see the "dynamic signature" of responding correctly.

10.1. Proportion target intent (Model 1; All Talkers)

The proportion of accuracy in selecting the corresponding talker intent was evaluated based on Talker, Intended Affective Category, the Wii trajectory variables [reaction time, distance, x-flip, y-flip, $x_{100–400}$ and $y_{100–400}$], and the Talkers' acoustics (F_0, amplitude, duration) as fixed effects, (see Appendix C.1 for results; see Fig. 6 for the means and standard errors for accuracy rates for Talker and Talker Intent). The results from this model indicated that the listeners differentially responded to each of the talkers with varying levels of accuracy, with significantly higher accuracy to Female 1, and significantly lower accuracy to Female 2 (*all* $p < .001$). The results also indicated that listeners were better able to categorize *Innuendo* ($b = 1.874$, $z = 7.201$, $p < .001$) and *Irritation* ($b = 1.818$, $z = 7.461$, $p < .001$) significantly better than *Compassion*.

Of the twelve Wii trajectory measures, three significantly predicted accuracy, which included response time ($b = -1.677$, $z = -5.755$, $p < .001$), distance ($b = 1.398$, $z = 2.739$, $p < .01$), and a marginal effect of y_{300} ($b = -3.548$, $z = -1.670$, $p = .09$). These results indicate that accuracy is indicative of faster response times, more distance covered (i.e., sampling the response space more) and marginally faster responses towards the correct intent within 300 ms along the y axis (y_{300}). The model also indicated that duration of the statement was a significant predictor of accuracy ($b = -6.494$, $z = -4.313$, $p < .001$). This indicates that listeners who were able to more quickly process the statement may have been better able to select the intended response. However, since Female 2 received significantly lower accuracy ratings than the other talkers, the next model removed F2 from the analysis to determine if the effects may have been affected by the listeners' difficulty judging her intention. Upon evaluation of listener responses to Female 2's statements, revealed that her Compassion responses were not significantly predicted by any of the Wii or Acoustic measures. This might indicate insufficient amount of variability in the data, because most listeners on average miscategorized these statements. Interestingly, Female 2 received an expressivity score of 2.75 and an identifiable score of 3.68. This suggests the listeners did view her as a moderately expressive talker (at least based on the likert scale and not relative to the other talkers), but also highly identified with her (based on the likert scale). This might indicate that the listeners did not find Female 2's affective expressions problematic at all. However, based on the participant responses to Compassion (primarily), their action dynamics tell a different story. Anecdotally speaking, the participants may not have been fully aware of the difficulty they were experiencing with this talker, which may have been related to the lack of feedback about their "correctness" of the interpretation, resulting in average talker ratings for Female 2.

10.2. Proportion target intent (Model 2; Female 2 removed)

The current model was identical to Model 1, with the exception of removing Female 2's statements from the data analysis. The results from Model 2 indicated that the listeners differentially responded to each of the talkers with

[4] Upon initial evaluation of the listener data, accuracy was relatively low for F2, to determine if the overall Wii trajectory effects and the acoustic predictors from Model 1 were due to the inaccuracy in responding to F2, F2 was removed from Model 2.

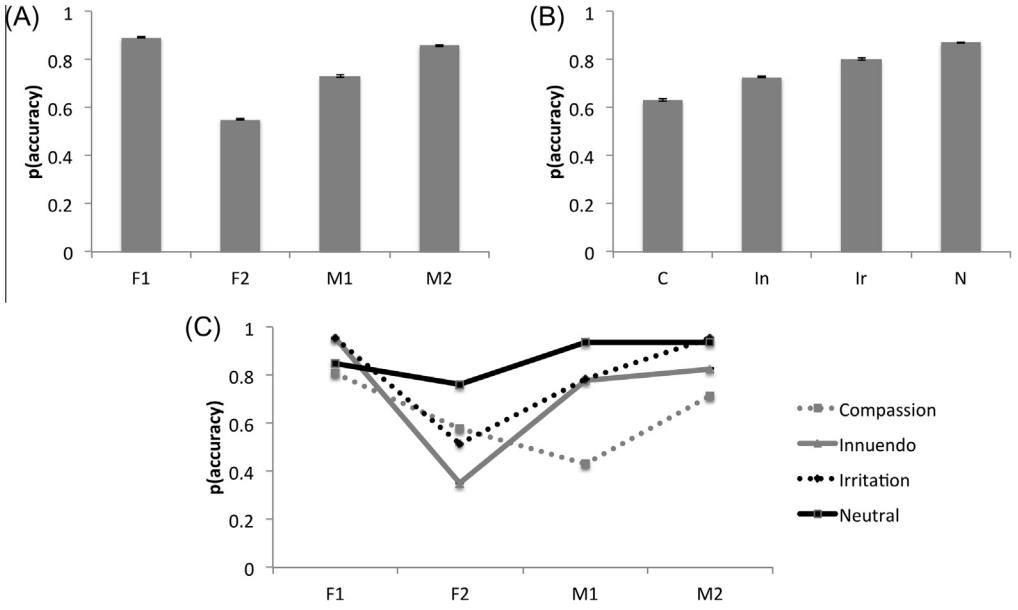

Fig. 6. Means and standard errors for the proportion of accurate responses: (A) Talker, (B) Affect and (C) Affect × Talker.

varying levels of accuracy and significantly higher accuracy to Female 1 (*all* $p < .001$; see Appendix C.2 for the *estimates*, *z* and *p-values*). The results also indicated that listeners were better able to categorize *Innuendo* ($b = 1.945$, $z = 6.465$, $p < .001$) and *Irritation* ($b = 1.888$, $z = 7.061$, $p < .001$) significantly better than *Compassion*. Of the twelve Wii trajectory measures, only two significantly predicted accuracy when Female 2 was removed, which included response time ($b = -2.081$, $z = -3.750$, $p < .001$) and distance ($b = 1.989$, $z = 2.885$, $p < .005$), while the marginal effect of y_{300} from Model 1 was no longer significant ($b = -4.236$, $z = -1.579$, $p = .11$). These results further indicate that accuracy is indicative of faster response times and more distance covered (i.e., sampling the response space more).

However, by removing Female 2 from the model, each of the acoustic variables predicted accuracy (though F_0 was marginal). This may indicate that listeners were sensitive to the amplitude variation ($b = -1.051$, $z = -2.843$, $p < .005$) and marginally sensitive to pitch variation ($b = -4.641$, $z = -1.913$, $p = .06$) in the talker's signals, meaning that the more variation the talker uses the more likely a listener will correctly identify the intended affective intention. Additionally, producing the statement faster may have also aided in decoding intent, as a means to preserve the acoustic variability while holding a potential match in working memory.

The current models consider all responses (*correct* and *incorrect*), which may have clouded the action dynamics of how a listener responds when they correctly identify the target intention, because we also included the *incorrect* intent categorizations. Therefore, the next model only considers the action dynamics of only the *correct* responses.

10.3. Multinomial logistic regression (Model 3; all talkers, correct responses only)

In the third model, we used a multinomial logistic regression (Croissant, 2007) to evaluate the relative likelihood that the *correct* selection of a speaker's intent was related the Wii trajectory and acoustic measures (see Section 5 for a full description of the type of analyses). The results from the model suggest that the selected predictors were a good fit ($\chi^2 = 33.308$, $p < .001$; see Fig. 7 for a sample of one participant's responses all Talkers' intents).

Since there were a large number of significant parameters in the model, only a small number of effects will be summarized here (please see Appendix C.3 for the estimates, *t* and *p*-values for the significant predictors). The results indicated that response time, distance, *x*flip, *y*flip, and *y*300 varied between Talkers and Affective Intent (all *p's* at least $p < .05$). For example, Female 1 received the highest rate of accuracy as compared to all other talkers. Listener's responded with faster response times, covered less distance, produced fewer *x*-flips (but more *y*-flips), and responded slower towards the *correct* response option to Female 1's *Irritation* relative to Female 2's *Compassion*. This suggests that more confident listeners waited a bit longer to make a decision (slower y_{300s}), but showed less hesitation (marked by faster response times, less distance covered and fewer *x*-flips). Though this is merely a descriptive account of the action dynamics of the listeners' responses, it does suggest that the listeners were more confident in their response to Female 1, relative to Female 2.

Alternatively, when comparing responses to the other talkers more variability was found. For example, listeners took longer to respond (response time), covered more distance and *x*flips, and began moving towards the correct

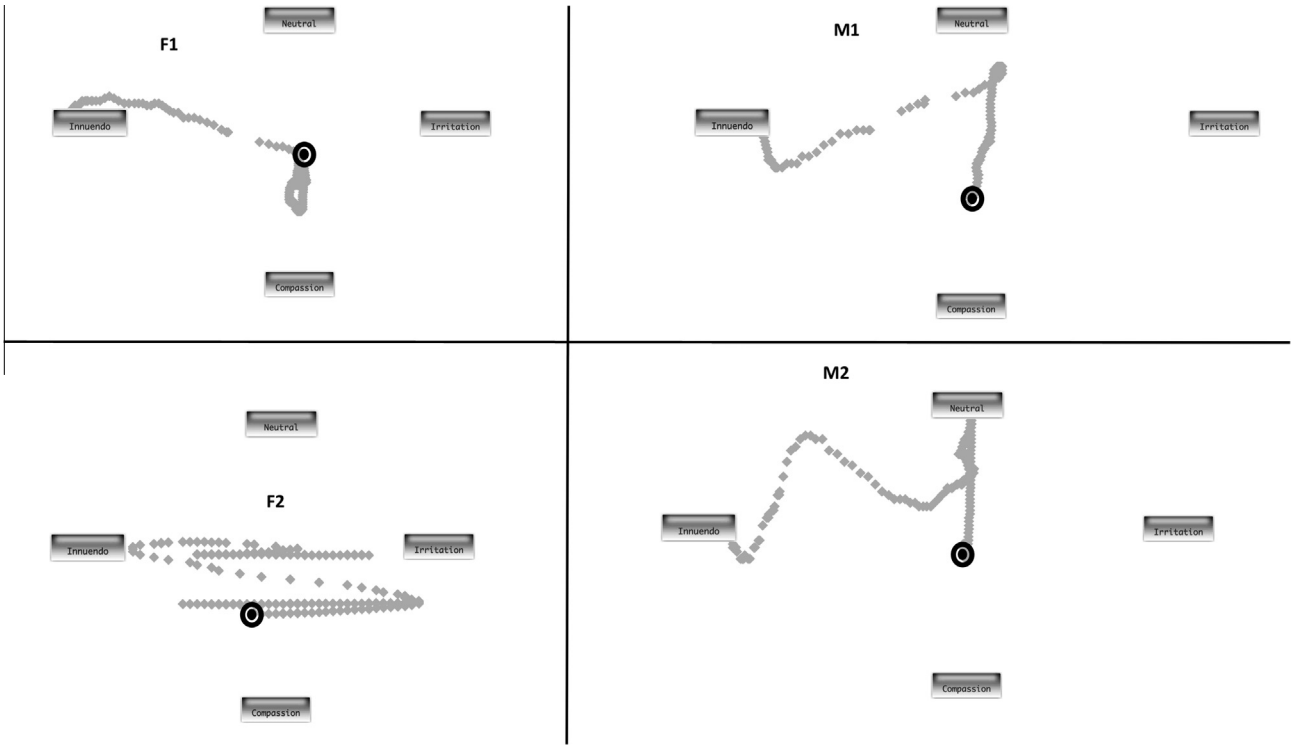

Fig. 7. Sample Wii trajectories from participant 15, statement "Elephants carried the supplies," with an Innuendo intonation, for each of the four talkers.

answer faster for Female 2's *Innuendo* relative to Female 1's *Compassion*. This is interesting because even though Female 2 was harder to *comprehend* relative to Female 1, the listener's action dynamics reveal something very interesting. Though it took them longer to make a decision, they "*knew*" what the correct answer was early in the signal, but there may not have been enough acoustic variability in the Female 2s signal to allow the listener to confidently choose the *correct* intent (as indicative of more distance and *x*flips). Furthermore, the likelihood of *correctly* choosing the intended affective category was also dependent on the acoustics (duration, amplitude, and mean F_0, all *p's* at least $p < .001$; see Appendix C.3 for *estimates*, *t* and *p-values*). The results from this analysis are not intended to provide concrete characteristics about each individual talker, rather they are meant to highlight the fact that listeners address talker variability in a flexible manner, which is indicated in the dynamics of the perception action system.

11. Summary and conclusions

It is clear from the following analyses that the listeners in the current study were able to successfully categorize the intent for almost all the talkers. What makes our findings novel and interesting is that how the participants *perceive* the intentions of the talker seems to be modulated by their perception action system. That is, during emotional/affect perception, even though the listeners were accurate (i.e., they can do the task), the state of their cognitive/psychological system reveals that their confidence in making the decision is sometimes shaky. These findings provide interesting implications towards the way we conceptual pragmatic interpretations of affective language.

However, differences in accuracy may speak directly to how listeners weight individual talker variability, as it relates to the talker-specific effects. This goes beyond the differences between talkers due to biological contributions to speech production (i.e., vocal tract characteristics). Rather, this speaks directly to the notion that listeners are sensitive to the differences in the talkers' cognitive realization of the affective intent. It could be that listeners do have a *prototypical* representation of particular affective categories, but the cognitive system needs to be able to flexibly handle both between (i.e., different talkers speak differently) and within talker variability (changing prosody should cue the listener that further interpretation of the signal is necessary). Therefore, listeners may try to force a talker's affect production into their own representation of that affective category, and if it does not fit, the action system may reflect indecision and hesitation during responding (as indicative of a higher instance of slower response times and increased distance, and *x* and *y* flips in some talkers and intents relative to others; see Model 3).

This notion is of particular importance especially because the visual modality has been suggested to be the most dominant for perception (e.g., Colavita, 1979; Gibson, 1933; Pick et al., 1969; Posner et al., 1976), but it seems intuitive that the auditory modality has its advantages in regards to communication (Hawk et al., 2009). Social beings rely on others for survival, which requires them to be able to respond felicitously in varying communicative settings. There are many ways in which individuals

communicate to decrease social distance, and decoding intent may be one of them. The current focus of this paper is to determine the how listeners handle talker variability during the interpretation of intent. Our analyses showed that both between and within talker variability differentially impacted a listener's perception and action mechanisms related to the interpretation of intent.

The results from the acoustic analysis strengthen claims that there necessarily needs to be overlapping acoustic correlates, but also supports that talker variability exists between and within talkers. In fact, it could be that the images used to elicit the affect production from each talker may have differentially influenced the talker's ability to produce the category the researchers intended. However, the results still provide strong evidence that listeners have a keen ability to detect these differences and use them to decode intent. It is important to call attention to talker variability because of the potential effect it has on listener perception. Simply put, we need to recognize what people are saying and who is saying it. The success of social interactions could be threatened if a person fails to understand or misinterprets something during communication (e.g., see Dijker, 1987; Gilovich et al., 1998; Kraut and Johnson, 1979; Miller and McFarland, 1987). Failure to consider the role of talker variability during the interpretation of intent forces us to make broad assumptions about affect in general (i.e., assuming all talker intent should be manifested equivalently). If interlocutors produced and perceived affect systematically as rises and falls in prosody, across all talkers then affect would be a useless cue to intent (i.e., merely noise in the linguistic signal). The evidence from the perception and action experiment supports the claims that affect cues are not merely noise, but a rich paralinguistic cue that influences pragmatic interpretation.

Our perception results suggest that listeners have the ability to adjust to variable speech signals at high levels of pragmatic interpretation and may represent a wider range of interactions with different individuals (Legge et al., 1984; Nygaard and Pisoni, 1998; Nygaard et al., 1994; Sheffert et al., 2002). However, these results do not speak to the nature of perceptual processing. As shown in Egidi and Nusbaum (2012), brain-imaging techniques have shown that neural mechanisms differentially respond to incongruent affective information. In our Wii trajectory analysis, we also have a birds-eye view into the processing mechanisms related to perceiving intent. We found that varying levels of confidence, certainty, indecision and hesitation during the online processing of these affective expressions when the responses were correct, but also when the responses were incorrect depending on the context (e.g., see Appendix C.3 for differences between Talker × Intention combinations). The overarching conclusions here are that affective expressions and talkers may not be created equally in terms of processing mechanisms. These online measures tell us a little more about how listeners do this. That is, depending on the salience of the speakers' intent, a listener may have an easier or more difficult time processing these statements. This is an important aspect to be considered, because a listener may seem to be very good at determining intent (high accuracy levels), but it does not mean that they are able to do this in a consistent and confident way. Additionally, when perceiving talkers that produce affect that aligns less with one's own representation of an affective category, successful interpretation is likely to fail.

Listeners may have arrived at their decisions differently depending on the expression and the ability of the talker. These differences could be directly related to how well the speakers were able to "act out" their intentions. However, as Scherer (2003) points out, affect is generally "acted out". So, the difference between success levels between talkers may be directly related to the speaker's ability to "act-out" their intentions. This is an interesting notion, because we should not naturally assume that everyone could produce affect or their intentions identically or even successfully (e.g., we all know someone who is difficult to interpret). Interestingly, the results here indicate that listeners are in fact able to handle such variability in an "actors" ability to express these emotionally-laden expressions. This speaks directly to the idea that affect perception is a dynamic process that may require listeners to adapt to the context (in this example, the speaker is the context).

Overall, this experiment suggests that the perception of vocally produced affect cues is not simple, and should be evaluated along a dynamic continuum (relative to a discrete affective category). The results from the categorization task mirror the results from previous studies (e.g., Nygaard and Lunders, 2002; Nygaard and Queen, 2008; Scherer, 2003; Scherer and Oshinsky, 1977), but evaluating the effects of categorical perception during dynamic online processing revealed that even when a listener is able to decode intent, they still may have a difficult time doing so. The realization that the talker brings a great deal of variability to the interaction may help elucidate how the listener handles various cues to discern intent. These studies provide valuable insight into how affective prosody influences the interpretation of intent that may have relevant social implications.

Though the evidence here suggests that affect, as it relates to intent, may be produced and perceived more continuously, there were a number of limitations to these studies. Specifically, this task was virtually devoid of context and participants were forced to choose from a previously defined set of categories. Our results are consistent with previous findings that support the notion that listeners are adept at interpreting intent, even under sparse contextual settings. For example, evidence from Scherer and Oshinsky (1977) has suggested that when global cues of affective expressions are synthetically imposed onto natural speech, listeners may still be able to correctly categorize the affect expression. Therefore, processing of affective information may happen very differently given a richer context for listeners to rely on. Participants may have also perceived

some of the expressions as other affective categories not provided, which we did not directly measure in the current experiment (e.g., especially with F2's Compassion statements).

The results reported here also highlight the importance of individual differences in both generation and perception of emotionally-laden language in dyadic interactions. A compelling example is when someone says, "I'm fine." For the talker, changes in intonations can drastically change the intent of this phrase. For the listener who has never interacted with the talker, the phrase may be interpreted literally, conveying positive/neutral valence. However, imagine the talker is interacting with a romantic partner, and has learned her partner's subtle prosodic cues over the course of their relationship. Instead of being taken literally, the listener knows to interpret that phrase and intonation as meaning that the talker is not fine. In other words, future research should examine how the talker and listener effects reported in the current study are moderated by relational familiarity.

Finally, we only evaluated three acoustic variables (F_0, amplitude, and duration). We chose to do this because we only wanted a preliminary analysis of the acoustic measures that have been shown to significantly contribute to vocal affect. Future studies necessarily need a more in-depth acoustic analysis to uncover the contributing acoustic measures that differentiate the affective intents evaluated in the current study. We are currently working to collect more production data and attempting to extract other relevant acoustic correlates, as a means to model the contribution of affective cues to the interpretation of intent. Additionally, future studies should evaluate richer contextual environments, social variables and alternative interpretations during the online processing of affective speech when it succeeds and fails. We are also currently considering the role of listener adaptation regarding, if and how listeners adapt their perceptions to more closely match those of a more difficult to understand talker's representation, through learning about those talkers' characteristics.

Acknowledgements

Special thanks goes to all undergraduate researchers Amy Roche and Ronni Jupson for all their hard work and dedication. We especially thank Jeremy Jamieson for reviewing the document and providing feedback, as well as Martijn Goudbeek for his invaluable contributions as a reviewer and the careful detail he spent reviewing our manuscript. Preparation of the manuscript was supported by a grant from the National Science Foundation to Rick Dale (NSF HSD-0826825).

Appendix A

Odds ratios (OR), t and p-values for each of the acoustic measures for Talker by Intent. The OR may be interpreted as whether the acoustic measure is predictive of the intended category. For example, the odds that F1 has produced an innuendo expression increase as the duration for her statement increases. Alternatively, you could interpret the OR in the following manner: When duration is longer, the intended statement is 1.4 times more likely to be innuendo than something else.

Measure	Talker	Intent	OR	t-Value
Duration	Female 1	Compassion	2.22	6.43^{***}
		Innuendo	3.37	10.41^{***}
		Irritation	2.08	6.21^{***}
		Neutral	1.89	5.30^{***}
	Female 2	Compassion	2.48	7.69^{***}
		Innuendo	2.41	7.54^{***}
		Irritation	1.87	5.22^{***}
		Neutral	2.52	7.94^{***}
	Male 1	Compassion	1.87	5.30^{***}
		Innuendo	3.03	9.54^{***}
		Irritation	1.34	2.31^{***}
	Male 2	Compassion	1.63	4.04^{*}
		Innuendo	3.00	9.15^{***}
		Irritation	1.31	2.11^{*}
F_0	Female 1	Compassion	1.02	20.89^{***}
		Innuendo	1.01	8.55^{***}
		Irritation	1.01	13.88^{***}
		Neutral	1.01	9.52^{***}
	Female 2	Compassion	1.02	9.41^{***}
		Innuendo	1.01	13.15^{***}
		Irritation	1.01	11.90^{***}
		Neutral	1.01	16.47^{***}
	Male 1	Compassion	1.01	6.08^{***}
	Male 2	Compassion	1.00	3.23^{**}
		Innuendo	0.99	-2.89^{**}
Intensity	Female 1	Compassion	0.96	-4.94^{***}
		Irritation	0.96	-4.73^{***}
		Neutral	0.97	-4.30^{***}
	Female 2	Compassion	0.95	-6.61^{***}
		Irritation	0.95	-5.35^{***}
		Neutral	0.95	-6.64^{***}
	Male 1	Compassion	0.98	-6.44^{*}
	Male 2	Compassion	0.96	-2.08^{***}
		Innuendo	0.98	-2.78^{**}
		Neutral	0.97	-3.71^{***}

Appendix B

Rating scores for the questions asked in Block 4 of the perception experiment (1 = very to 5 = not at all).

		Expressiveness	Identifiable	Ethnicity	Gender
Talker	Female 1	2.13(1.12)	3.21(1.18)	0.83(0.38)	0.10(0.00)
	Female 2	2.75(1.19)	3.68(0.95)	0.92(0.34)	0.10(0.00)
	Male 1	1.96(1.11)	2.47(1.39)	0.86(0.28)	0.98(0.20)
	Male 2	2.67(1.17)	3.58(1.02)	0.79(0.42)	0.10(0.00)

Appendix C

Estimates, t and p-values for each of the predictors in Models 1 (C.1), 2 (C.2), and 3 (C.3).

C.1. Appendix

Model 1 (*All Talkers*) estimates, z-scores, and p-values (Female 1 and *Compassion* as the reference categories).

Variable category			Estimate	z-Score
Talker		Female 2	−1.369	−8.093***
		Male 1	−2.293	−8.760***
		Male 2	−9.266	−3.598***
Affect		Innuendo	1.874	7.201***
		Irritation	1.818	7.461***
Wii measure		Response time	−1.677	−5.755***
		Distance	1.398	2.739**
		Y_{300}	−3.548	−1.670
Acoustic measure		Duration	−6.494	−4.313***
Talker × Affect	Female 2	Innuendo	−2.837	−10.903
		Irritation	−2.251	−8.613***
		Neutral	7.371	3.421***
	Male 2	Innuendo	−1.037	−3.788***
		Neutral	1.648	6.150***
	Male 1	Neutral	2.925	10.741***

C.2. Appendix

Model 2 (*Talker Female 2 removed*) estimates, z-scores, and p-values (Female 1 and *Compassion* as the reference categories).

Variable category			Estimate	z-Value
Talker		Male 1	−2.489	−7.449***
		Male 2	−1.195	−3.812***
Affect		Innuendo	1.945	6.465***
		Irritation	1.888	7.061***
Wii measure		Response time	−2.081	−5.549***
		Distance	1.989	2.885**
Acoustic measure		F0	−4.641	−1.913
		Amplitude	−1.051	−2.843**
		Duration	−4.335	−2.112*
Talker × Affect	Male 2	Innuendo	−1.112	−3.733***
		Irritation	5.812	1.738
		Neutral	1.890	6.516***
	Male 1	Neutral	3.182	10.536***

C.3. Appendix

Model 3 (Multinomial Logistic Regression, correct responses only) estimates, t-scores, and p-values.

Talker	Affect	Variable	Estimate	t-Value
Female 1	Irritation	Response time	−0.405	−3.179**
		Distance	−0.001	−3.109**
		xflip	−0.075	−3.297***
		yflip	0.039	2.318*
		F0	0.052	6.552***
		Amplitude	3.897	15.540***
		Duration	7.694	7.443***
	Innuendo	y300	0.051	2.130*
		F0	−2.865	−11.707***
		Amplitude	119.200	12.042***
		Duration	222.930	11.137***
	Neutral	y300	0.025	2.082*
		F0	−0.183	−20.275***
		Amplitude	−1.648	−11.229***
		Duration	−10.286	−12.745***
	Compassion	F0	2.652	3.793***
		Amplitude	27.244	3.699***
		Duration	−407.160	−3.668***
Female 2	Irritation	y300	0.026	2.145*
		F0	−0.119	−14.406***
		Amplitude	−1.753	−12.403***
		Duration	−8.623	−10.778***
	Innuendo	Response time	0.346	2.631**
		yflip	−0.040	−1.862
		F0	−0.037	−4.216***
		Amplitude	3.805	14.351***
		Duration	10.799	9.951***
	Neutral	F0	−1.075	−14.647***
		Amplitude	−14.609	−14.659***
		Duration	−12.414	−11.273***
Male 1	Irritation	F0	−10.629	−12.671***
		Amplitude	−0.993	−3.306***
		Duration	−294.530	−13.170***
	Innuendo	yflip	−0.071	−2.228*
		F0	−10.695	−12.750***
		Duration	−279.930	−12.518***
	Neutral	yflip	−0.052	−1.655
		F0	−10.666	−12.716***
		Amplitude	−1.532	−5.024***
		Duration	−303.350	−13.563***
	Compassion	xflip	0.078	2.262*
		yflip	−0.060	−2.026*
		F0	−10.584	−12.616***
		Amplitude	−1.256	−4.219***
		Duration	−291.080	−13.015***
Male 2	Irritation	F0	−10.617	−12.658***
		Duration	−288.590	−12.908***
		Amplitude	−2.806	−9.111***

Appendix C.3 (*continued*).

Talker	Affect	Variable	Estimate	*t*-Value
	Innuendo	xflip	0.071	2.024*
		yflip	−0.071	−2.343*
		F0	−10.674	−12.726***
		Amplitude	−2.937	−9.284***
		Duration	−278.370	−12.449***
	Neutral	yflip	−0.063	−2.044*
		F0	−10.700	−12.756***
		Amplitude	−2.559	−8.296***
		Duration	−296.610	−13.263***
	Compassion	xflip	0.055	1.721
		yflip	−0.072	−2.565*
		F0	−10.459	−12.478***
		Amplitude	−5.000	−14.121***
		Duration	−274.490	−12.285***

References

Attardo, S., Eisterhold, J., Hay, J., Poggi, I., 2005. Multimodal markers of irony and sarcasm. Humor 16 (2), 243–260.

Austin, J., 1962. How to do Things with Words. University Press, Oxford.

Bachorowski, J., 1999. Vocal expression and perception of emotion. Curr. Direct. Psychol. Sci. 8 (2), 53–57.

Bachorowski, J., Owren, M., 1999. Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. J. Acoust. Soc. Am. 106, 1054–1063.

Banse, R., Scherer, K., 1996. Acoustic profile in vocal emotion expression. J. Pers. Soc. Psychol. 70 (3), 614–636.

Bänziger, T., Mortillaro, M., Scherer, K., 2012. Introducing the geneva multimodal expression corpus for experimental research on emotion perception. Emo. 12 (5), 1161–1179.

Bänziger, T., Scherer, K.R., 2007. Using actor portrayals to systematically study multimodal emotion expression: the GEMEP corpus. Affective Computing and Intelligent Interaction. Springer, Berlin Heidelberg, pp. 476–487.

Bell, D.M., 1997. Innuendo. J. Pragmatics 27 (1), 35–59.

Bodenhausen, G., Moreno, K., 2000. How do I feel about them? The role of affective reactions in intergroup perception. In: Bless, H., Forgas, J.P. (Eds.), The Message within: The Role of Subjective Experience in Social Cognition and Behavior. Psychology Press, Philadelphia, pp. 283–303.

Boersma, P., Weenink, D., 1992. Praat: Doing Phonetics by Computer (Version 4.3.14) [Computer Software and Manual]. <www.praat.org>.

Brainard, D., 1997. The psychophysical toolbox. Spat. Vis. 10, 433–436.

Cheang, H.S., Pell, M.D., 2008. The sound of sarcasm. Speech Commun. 50 (5), 366–381.

Colavita, F., 1979. Human sensory dominance. Percept. Psychophys. 16, 409–412.

Croissant, Y., 2007. Mlogit: Multinomial Logit Model. R Package Version 0.2-3.

Crowder, R., Morton, J., 1969. Precategorical acoustic storage. Percept. Psychophys. 5, 365–373.

Dale, R., Kehoe, C., Spivey, M., 2007. Graded motor response in the time course of categorizing atypical exemplars. Mem. Cognit. 33, 15–28.

Dale, R., Roche, J., Snyder, K., McCall, R., 2008. Exploring action dynamics as an index of paired-associate learning. PLoS One 3 (3), e1728.

de Gelder, B., Pourtois, G., Vroomen, J., Bashoud-Levi, A.C., 2000. Covert processing of faces in posopagnosia is restricted to facial expressions: evidence from cross-modal bias. Brain Cognit. 44 (3), 425–444.

Dijker, A., 1987. Emotional reactions to ethnic minorities. Eur. J. Soc. Psychol. 17, 305–325.

Egidi, G., Gerrig, R.J., 2009. How valence affects language processing: negativity bias and mood congruence in narrative comprehension. Mem. Cognit. 37, 547–555.

Egidi, G., Nusbaum, H.C., 2012. Emotional language processing: how mood affects integration processes during discourse comprehension. Brain Lang. 122, 199–210.

Ekman, P., 1992. Are there basic emotions?. Psychol. Rev. 99 (3) 550–553.

Ekman, P., Friesen, W., Ellsworth, P., 1972. Emotion in the Human Face: Guidelines for Research and an Integration of Findings. Pergamon Press, New York.

Farmer, T., Cargill, S., Hindy, N., Dale, R., Spivey, M., 2007. Tracking the continuity of language comprehension: computer-mouse trajectories suggest parallel syntactic processing. Cognit. Sci. 31 (5), 889–909.

Fischer, A.H., Manstead, A.S.R., 2008. Social functions of emotion. In: Lewis, M., Haviland-Jones, J., Barrett, L.F. (Eds.), Handbook of Emotions, third ed. Guilford Press, New York.

Forster, K., 1979. Levels of processing and the structure of the language processor. In: Cooper, W., Walker, E. (Eds.), Sentence Processing: Psycholinguistic Studies Presented to Merril Carrett. Erlbaum, Hillsdale, NG, pp. 27–85.

Freeman, J., Ambady, N., 2010. Mousetracker: Software for studying real-time mental processing using a computer mouse-tracking method. Behav. Res. Method 42, 226–241.

Freeman, J., Dale, R., Farmer, T., 2011. Hand in motion reveals mind in motion. Front. Psychol. 2, 59.

Fridlund, A.J., 1994. Human Facial Expression: An Evolutionary View. Academic Press, New York.

Galloway, C., 1974. Non-verbal: the language of sensitivity. Theor. Pract. 13 (5), 380–383.

Gibson, J., 1933. Adaption, after-effect, and contrast in the perception of curved lines. J. Exp. Psychol. 16, 1–31.

Gibson, E., Piantidosi, S., Fedorenko, K., 2011. Using mechanical turk to obtain and analyze acceptability judgments. Lang. Linguist. Compass 5 (8), 509–524.

Gilovich, T., Savitsky, K., Medvec, V., 1998. The illusion of transparency: biased assessments of others' ability to read one's emotional states. J. Pers. Soc. Psychol. 75, 332–346.

Goldstein, U., 1980. An Articulatory Model for the Vocal Tracts of Growing Children. Ph.D. Thesis. MIT, MA.

Grundy, P., 2008. Doing Pragmatics, third ed. Taylor and Francis, New York, NY.

Guerrero, L., Floyd, K., 2006. Nonverbal expressions of emotion. In: Nonverbal Communication in Close Relationships. Lawerence Erlbaum Associates, Mahwah, NJ, pp. 107–132.

Halberstadt, J., Niendenthal, P., Kushner, J., 1995. Resolution of lexical ambiguity by emotional state. Psychol. Sci. 6 (5), 278–282.

Hammerschmidt, K., Jurgen, U., 2007. Acoustic correlates of affective prosody. J. Voice 21 (5), 531–540.

Hawk, S., van Kleef, G., Fischer, A., van der Schalk, J., 2009. "Worth a thousand words": absolute and relative decoding of nonlinguistic affect vocalizations. Emotion 9 (3), 293–305.

Ishii, K., Kitayama, S., 2002. Spontaneous attention to word content versus emotional tone differences among three cultures. Psychol. Sci. 14 (1), 39–46.

Jaeger, T.F., 2008. Categorical data analysis: away from anovas (transformation or not) and towards Logit Mixed Models. J. Mem. Lang. 59 (4), 434–446.

Juslin, P., Laukka, P., 2001. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. Emotion 4, 381–412.

Juslin, P., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: different channels, same code? Psychol. Bull. 129, 770–814.

Kaiser, L., 1962. Communication of affects by single vowels. Synthese 14 (4), 300–319.

Keltner, D., Haidt, J., 1999. Social functions of emotions at four levels of analysis. Cogn. Emot. 13, 505–521.
Kitayama, S., Ishii, K., 2003. Word and voice: spontaneous attention to emotional utterances in two languages. Cogn. Emot. 16 (1), 29–59.
Kitayama, S., Mesquita, B., Karasawa, M., 2006. Cultural affordances and emotional experience: socially engaging and disengaging emotions in Japan and the United States. J. Pers. Soc. Psychol. 91 (5), 890–903.
Kraut, R., Johnson, R., 1979. Social and emotional messages of smiling: an ethological approach. J. Pers. Soc. Psychol. 27 (9), 1530–1553.
Ladd, D., Scherer, K.R., Silverman, K.E.A., 1986. An integrated approach to studying intonation and attitude. In: Johns-Lweis, C., Croom (Eds.), Intonation and Discourse. Helm, London.
Lang, P.J., Bradley, M.M., Cuthbert, B.N., 2008. International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual. Technical Report A-8, University of Florida, Gainesville, FL.
Legge, G.E., Grosmann, C., Pieper, C.M., 1984. Learning unfamiliar voices. J. Exp. Psychol. Learn. Mem. Cogn. 10, 298–303.
Leionenen, L., Hiltunen, T., Linnankoski, I., Laakso, M., 1997. Expression of emotional–motivational connotations with a one-word utterance. J. Acoust. Soc. Am. 102 (3), 1853–1863.
Liscombe, J., Venditti, J., Hirschberg, J., 2003. Classifying Subject Ratings of Emotional Speech using Acoustic Features. Eurospeech, Geneva.
MacMillan, N., Goldberg, R., Braida, L., 1988. Resolution for speech sounds: basic sensitivity and context memory on vowel and consonant continua. J. Acoust. Soc. Am. 84, 1262–1280.
Majid, A., 2012. Current emotion research in language sciences. Emot. Rev. 4 (4), 432–443.
Martinez-Castilla, P., Peppe, S., 2008. Intonation features of the expression of emotions in Spanish: preliminary study for a prosody assessment procedure. Clin. Linguist. Phonet. 22 (4–5), 363–370.
Massaro, D.W., 1998. Perceiving Talking Faces: From Speech Perception to Behavioral Principle. MIT Press, Cambridge, Massachusetts.
Massaro, D., Cohen, M., 2000. Fuzzy logic model of bimodal emotion perception: comment on "The perception of emotions by ear and by eye" by de Gelder and Vroomen. Cogn. Emot. 14 (3), 313–320.
McKinstry, C., Dale, R., Spivey, M.J., 2008. Action dynamics reveal parallel competition in decision making. Psychol. Sci. 19 (1), 22–24.
Miller, D., McFarland, C., 1987. Plural ignorance: when similarity is interpreted as dissimilarity. J. Pers. Soc. Psychol. 53 (2), 298–305.
Morton, J., Trehub, S., 2001. Children's understanding of emotion in speech. Child Dev. 72 (3), 834–843.
Mozziconacci, S., 2001. Modeling emotion and attitude in speech by means of perceptually based parameter values. User Model. User-Adap. Inter. 11, 297–326.
Mullennix, J.W., Pisoni, D.B., 1990. Stimulus variability and processing dependencies in speech perception. Percept. Psychophys. 47 (4), 379–390.
Mullennix, J.W., Pisoni, D.B., Martin, C.S., 1989. Some effects of talker variability on spoken word recognition. J. Acoust. Soc. Am. 85 (1), 365–378.
Mullennix, J., Bihon, T., Bricklemyer, J., Gaston, J., Keener, J., 2002. Effects of variation in emotional tone of voice on speech perception. Lang. Speech 45 (3), 255–283.
Newman, R., Clouse, S., Burnham, 2001. The perceptual consequences of within-talker variability in fricative production. J. Acoust. Soc. Am. 109, 1181–1196.
Nordström, P.E., 1977. Female and infant vocal tracts simulated from male area functions. J. Phonet. 4, 81–92.
Nygaard, L., Lunders, E., 2002. Resolution of lexical ambiguity by emotional tone of voice. Mem. Cognit. 30 (4), 583–593.
Nygaard, L., Pisoni, D., 1998. Talker-specific learning in speech perception. Percept. Psychophys. 60, 355–376.
Nygaard, L., Queen, J., 2008. Communicating emotion: linking affective prosody and word meaning. J. Exp. Psychol. Hum. Percept. Perform. 34 (4), 1017–1030.
Nygaard, L., Sommers, M., Pisoni, D., 1994. Speech perception as a talker-contingent process. Psychol. Sci. 5, 42–46.
Pick, H., Warren, D., Hay, J., 1969. Sensory conflict in judgements of spatial direction. Percept. Psychophys. 6, 203–305.
Pisoni, D.B., 1992. Long-term memory in speech perception: some new findings on talker variability, speaking rate and perceptual learning. Speech Commun. 13, 109–125.
Posner, M., Nissen, M., Klein, R., 1976. Visual dominance: an information-processing account of its origins and significance. Psychol. Rev. 83, 157–171.
Rockwell, P., 2000. Lower, slower, louder: vocal cues of sarcasm. J. Psycholinguist. Res. 29 (5), 483–495.
Russ, J., Gur, R., Bilker, W., 2008. Validation of affective and neutral sentence content for prosodic testing. Behav. Res. Method 40 (4), 935–939.
Ryalls, J., Zipprer, A., Baldauff, P., 1997. A preliminary investigation of the effects of gender and race on voice onset time. J. Speech Lang. Hear. Res. 40, 642–645.
Sacharin, V., Schlegel, K., Scherer, K.R., 2012. Geneva Emotion Wheel Rating Study (Report). University of Geneva, Swiss Center for Affective Sciences, Geneva, Switzerland.
Scherer, K.R., 1980. The functions of nonverbal signs in conversation. In: St. Clair, R.N., Giles, H. (Eds.), The Social and Psychological Contexts of Language. Hillsdale, NJ: Erlbaum, pp. 225–244.
Scherer, K.R., 1986. Vocal affect expression: a review and model for future research. Psychol. Bull. 99, 143–165.
Scherer, K.R., 1988. On the symbolic functions of vocal affect expression. J. Lang. Soc. Psychol. 7, 79–100.
Scherer, K.R., 1994. Affect bursts. In: van Goozen, S.H.M., van de Poll, N.E., Sergeant, J.A. (Eds.), Emotions: Essays on Emotion Theory. Erlbaum, Hillsdale, NJ, pp. 161–193.
Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech Commun. 40, 227–256.
Scherer, K.R., Banziger, T., 2004. The role of intonation in emotional expressions. Speech Commun. 46 (3–4), 252–267.
Scherer, K.R., Ceschi, G., 1997. Lost luggage: a field study of emotion-antecedent appraisal. Motivat. Emot. 21 (3), 211–235.
Scherer, K., Oshinsky, J., 1977. Cue utilization in emotion attribution from auditory stimuli. Motivat. Emot. 1, 331–346.
Scherer, K.R., Ladd, R., Silverman, K.E.A., 1984. Vocal cues to speaker affect: testing two models. J. Acoust. Soc. Am. 76 (5), 1348–1356.
Scherer, K.R., Banse, R., Wallbott, H., 2010. Emotion inferences vocal expression correlate across languages and cultures. J. Cross Cult. Psychol. 32 (1), 76–92.
Schirmer, A., Kotz, S., Friederici, A., 2005. On the role of attention for the processing of emotions in speech: sex differences revisited. Cognit. Brain Res. 24 (3), 442–452.
Seibert, P.S., Ellis, H.S., 1991. A convenient self-referencing mood induction procedure. Bull. Psychonom. Soc. 29 (12), 1–124.
Sheffert, S., Pisoni, D., Fellowes, J., Remez, R., 2002. Learning to recognize talkers from natural, sinewave and reversed speech. J. Exp. Psychol. Hum. Percept. Perform. 28 (6), 1447–1469.
Snow, R., O'Connor, B., Jurafsky, D., Ng, A., 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii.
Sorokin, A., Forsythe, D., 2008. Utility data annotation with Amazon mechanical Turk. In: Proceedings of First IEEE Workshop on Internet Vision at CVPR.
Spivey, M., 2007. The Continuity of Mind. Oxford University Press, Oxford, UK.
Spivey, M., Grosjean, M., Knoblich, G., 2005. Continuous attraction toward phonological competitors. Proc. Natl. Acad. Sci. USA 102 (29), 10292–10298.
Swerts, M., Hirschberg, J., 2010. Prosodic predictors of upcoming positive or negative content in spoken messages. J. Acoust. Soc. Am. 128 (3), 1337–1345.
van Kleef, G.A., De Dreu, C.K.W., Manstead, A.S.R., 2004. The interpersonal effects of anger and happiness in negotiations. J. Pers. Soc. Psychol. 86, 57–76.