Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 65 (2014) 36–49

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Emotion in the voice influences the way we scan emotional faces

Simon Rigoulot *, Marc D. Pell

McGill University, Faculty of Medicine, School of Communication Sciences and Disorders, 1266 Avenue des Pins Ouest, Montréal, QC H3G 1A8, Canada
McGill Centre for Research on Brain, Language, and Music, Canada

Received 30 November 2013; received in revised form 6 May 2014; accepted 28 May 2014
Available online 6 June 2014

Abstract

Previous eye-tracking studies have found that listening to emotionally-inflected utterances guides visual behavior towards an emotionally congruent face (e.g., Rigoulot and Pell, 2012). Here, we investigated in more detail whether emotional speech prosody influences how participants scan and fixate specific features of an emotional face that is congruent or incongruent with the prosody. Twenty-one participants viewed individual faces expressing fear, sadness, disgust, or happiness while listening to an emotionally-inflected pseudo-utterance spoken in a congruent or incongruent prosody. Participants judged whether the emotional meaning of the face and voice were the same or different (match/mismatch). Results confirm that there were significant effects of prosody congruency on eye movements when participants scanned a face, although these varied by emotion type; a matching prosody promoted more frequent looks to the *upper* part of fear and sad facial expressions, whereas visual attention to upper and lower regions of happy (and to some extent disgust) faces was more evenly distributed. These data suggest ways that vocal emotion cues guide how humans process facial expressions in a way that could facilitate recognition of salient visual cues, to arrive at a holistic impression of intended meanings during interpersonal events.

Keywords: Speech; Prosody; Face; Eye-tracking; Emotion; Cross-modal

1. Introduction

Speech prosody refers to the extra-linguistic variations in speech (changes in pitch, tempo and loudness) that, among other functions, mark the pragmatic value of an utterance to the listener (Pell, 1999a,b), provide information about individual speaker characteristics (age, gender), and encode various intentions and beliefs of the speaker in the context of the utterance (Rigoulot et al., in press). During conversations, speech prosody is typically associated with other social cues like facial expressions or body movements; among these stimuli, faces appear to be privileged in many ways. For example, as early as 1967, Yarbus showed that eye fixations are more likely to be directed towards faces than towards any other part of a visual scene. Humans have the ability to quickly detect and analyze faces (Palermo and Rhodes, 2007) and possess an extensive mental inventory of 'known' faces (Bruce et al., 1992). Moreover, like speech prosody, faces are a critical source of information about the *emotional* state of another person.

Given their joint relevance to communication, person perception, and behavior more generally, interactions between speech prosody and facial cues are being intensively studied (Cvejic et al., 2010; Pell, 2005; Swerts and Krahmer, 2008). For example, Swerts and Krahmer (2008) recorded videos of speakers uttering a sentence with prominence (emphasis via prosody) on the first, middle or final word. They extracted the auditory and visual channels of these videos and presented them together in conditions that were congruent (e.g., visual and auditory channels both conveyed prominence on one of the three words) or incongruent (prominence produced in visual and auditory

* Corresponding author. Tel.: +1 5143984400x00010; fax: +1 5143988123.
E-mail address: simon.rigoulot@mail.mcgill.ca (S. Rigoulot).
URL: http://www.mcgill.ca/pell_lab (S. Rigoulot).

http://dx.doi.org/10.1016/j.specom.2014.05.006
0167-6393/

channels did not align). Participants had to indicate which word was the most prominent. The authors found that participants were faster to respond when the materials were congruent than incongruent, suggesting that visual cues can hinder the auditory processing of prominence. Interestingly, in a second experiment, the authors investigated the role of different regions of the face in these effects by blackening the upper or the lower part of the face. Their results show that facial cues located in the upper part of the face are stronger to bias the perception of speech prominence than those located in the lower part.

Along similar lines, a growing literature shows that *emotional* information encoded in the face and voice interacts in systematic ways (de Gelder and Vroomen, 2000; Pell, 2005) and that the perception of emotional meanings in the voice influences how listeners direct their attention to faces (Paulmann et al., 2012; Rigoulot and Pell, 2012). To better understand the effects of speech on face processing, this study investigated whether emotional prosody influences how listeners scan *specific regions* of a face that provide salient visual cues about the shared emotional meanings of the two stimuli through the analysis of eye gaze measures.

1.1. On the processing of facial expressions of emotion

To produce facial expressions of emotion, humans voluntarily or involuntarily contract different facial muscle groups, especially those involving the eyes, mouth, brows, nose, and cheeks (Ekman et al., 2002). Darwin was the first to suggest that this activity results in different spatial configurations that provide distinctive visual information corresponding to the participant's underlying emotion state (e.g., Darwin, 1872); for example, fear is characterized by raised eyebrows and the mouth tends to open and stretch horizontally (Facial Action Coding System, FACS; Ekman and Friesen, 1976; Ekman et al., 2002). The recognition of discrete emotional facial expressions could rely on the correct analysis of facial cues involving different parts of the face, as demonstrated by several studies (Bassili, 1979; Calder et al., 2000; Calvo and Nummenmaa, 2011). For instance, Calder and colleagues presented the top- or bottom-half of pictures displaying fearful, happy, disgusted, sad and surprised expressions and then analyzed error rates and reaction times of participants. They reported that anger, fear, and sadness were readily identified from the top section of the face, whereas happiness and disgust were readily identified from the bottom half of the face. This result suggests that the recognition of emotional facial expressions depends on *specific parts of faces* and varies by emotion type, with the upper part of the face providing more salient information for recognizing fearful, angry and sad faces, and the lower part of the face providing stronger cues for recognizing happy and disgusted faces.

Other researchers have studied the importance of features located in the upper (eyes, brows) and lower (mouth) part of the face during emotion recognition. Data suggest that the eye region is more important than other parts of the face for perceiving expressions of fear (Adolphs et al., 2005) and sadness (Eisenbarth and Alpers, 2011), although the role of other features, including those located in the lower part of the face (mouth in particular) is not to be excluded (see Blais et al., 2012; Beaudry et al., 2014). Relevant cues for detecting expressions of happiness and disgust seem to be more salient in the lower part of the face (mouth in particular; Gosselin and Schyns, 2001; Jack et al., 2009; Schyns et al., 2002). Work by Calvo and Marrero (2009) and Calvo and Nummenmaa (2008) argues that the mouth plays a unique role for the rapid detection of happy expressions (see also Beaudry et al., 2014 for the role of mouth in recognition of happiness); however, it is noteworthy that real versus posed smiles can be distinguished by looking only at the eyes (Ekman et al., 1990; Messinger et al., 2012 with children), suggesting that the importance of the lower (mouth) regions for recognizing happiness is not absolute. Similarly, facial expressions of disgust have been associated with increased fixations on the lower part of the face (mouth and lower part of the nose, Jack et al., 2009). Cultural differences in how individuals attend to different face regions during emotional processing have also been reported (Jack et al., 2009; Yuki et al., 2007; Tanaka et al., 2010). For example, Yuki et al. (2007) investigated whether facial cues are rated similarly by American and Japanese participants when presented chimeric emotional faces (emoticons) with different combinations of happy/sad/neutral eyes associated with happy/sad/neutral mouths; they found that the eye region biased perception to a greater extent in the Japanese group, suggesting an influence of cultural background in the way people use facial cues to process emotional facial expressions. Altogether, these findings reinforce the hypothesis that during face processing, recognition of discrete emotional expressions is guided by analysis of different face regions, even if the exact nature of these relationships and their cultural specificity remain unclear.

A particularly useful approach for investigating how specific face regions promote emotion recognition is by recording eye movements under different experimental conditions (Adolphs et al., 2005; Bate et al., 2009; Becker and Detweiler-Bedell, 2009; Green et al., 2003; Hunnius et al., 2011; Malcolm et al., 2008; Vassallo et al., 2009; Wong et al., 2005). When scanning different emotional expressions, distinct strategies or patterns have been described; in particular, participants demonstrated more frequent and longer fixations to the primary features of the face (mouth, eyes, and nose) when looking at threatening faces, such as anger and fear, than at other expression types (happy, sad and surprised, see Green et al., 2003; Bate et al., 2009). This result was interpreted as a "vigilant" scanning pattern, necessary for the efficient detection of a stimulus with potentially negative outcomes. However, the opposite pattern of results (i.e., less frequent and shorter fixations) has also been described (e.g., Hunnius