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

et al., 2011), consistent with an avoidant looking behavior when threat-related faces are encountered (cf. Becker and Detweiler-Bedell, 2009). Discrepant gaze patterns observed when threatening faces are presented have been attributed to the differential effects of task instructions, which are widely known to influence eye movements (Yarbus, 1967). Nonetheless, these studies again suggest that recognition of emotional facial expressions relies on certain features of the face, perhaps in an emotion-specific manner, although more data are needed to clarify the nature of these effects, and of primary interest here, to explore how visual scanning patterns are conjointly influenced by emotional *speech*.

1.2. Effects of emotional prosody on face processing

As noted earlier, emotional facial expressions encountered in social interactions are often accompanied by emotional speech cues; fluctuations in a speaker's pitch, the intensity of their voice, changes in voice quality, and variations in speech rate all supply core information about the speaker's emotional state as speech unfolds (Banse and Scherer, 1996; Pell and Kotz, 2011; Pell et al., 2009). Since facial and vocal cues are encountered simultaneously across two sensory modalities, vision and audition, the rapid detection and analysis of these combined signals are essential to correctly interpret their significance and to guide an appropriate behavioral response. To shed light on these processes, studies have investigated how emotional information conveyed by auditory and visual channels are integrated and how they interact (e.g., De Gelder and Vroomen, 2000; Paulmann and Pell, 2010). For example, de Gelder and Vroomen (2000) demonstrated that when participants were asked to identify facial expressions that had been 'morphed' between two emotion categories, decisions were biased in the direction of a simultaneously presented emotional prosody that matched one of the two emotional meanings; conversely, judgments of prosody as the target stimulus were biased by a concurrent facial expression in a similar manner.

More recently, eye-tracking studies have shown that when listeners are exposed to emotional sentences, the emotional meaning of the prosody influences the way that participants look at related versus unrelated emotional faces (Paulmann et al., 2012; Rigoulot and Pell, 2012). For example, Rigoulot and Pell (2012) presented an array of four emotional faces (fearful, angry, happy and neutral) to participants while they heard an emotional pseudo-utterance (e.g., *The fector egzullin tuh boshent*) expressing one of the four target emotions; measures of first fixation, and the frequency and duration of total fixations to each face, were analyzed in three different time windows which either coincided or followed presentation of the emotional utterance. In all three time windows, they found that participants looked longer and more frequently at emotional faces that matched the prosody, even long after the auditory stimulus had been heard, suggesting that the emotional meaning of vocal cues influences the way that humans visually scan and attend to emotional faces (see also Paulmann et al., 2012 for similar conclusions when English utterances were presented, and related work by Tanaka et al., 2010). These findings motivate additional research that investigates how emotional prosody influences eye movements and visual analysis of a concurrent emotional face.

1.3. The present investigation

In light of evidence that emotional speech cues influence how listeners gaze at faces, coupled with data suggesting that specific regions of a face are more salient for detecting certain facial expressions of emotion, an important question to address is whether emotional prosody systematically guides *where* listeners fixate on a face according to the emotional meaning of the vocal stimulus. While recent studies have explored the lateralization of visual attention to a face while hearing neutral and emotional speech (Thompson et al., 2009) and how emotional faces influence processing of audiovisual speech (Gordon and Hibberts, 2011), there are currently no data that inform how visual attention to *specific regions of a face* may be guided by the emotional meaning of speech prosody.

The present experiment was designed to test whether listening to an emotional utterance influences visual scanning patterns to an emotional face. We hypothesized that, in addition to causing listeners to gaze longer/more frequently at a facial expression within an array that is congruent with the prosody (Paulmann et al., 2012; Rigoulot and Pell, 2012), emotional meanings of prosody would reinforce gaze behavior towards critical visual cues (i.e., face regions) associated with the prosody. To test this hypothesis, we presented individual emotional faces to participants who simultaneously listened to an emotionally related or unrelated pseudo-utterance, while tracking the frequency and duration of their fixations to different regions of a target face. Participants judged whether the emotional expression of the voice and face matched. Based on the literature (e.g., Adolphs et al., 2005; Bassili, 1979; Calder and Young, 2005; Eisenbarth and Alpers, 2011), we concentrated on four discrete emotion types—fear, sadness, disgust, and happiness—for which there is some evidence that the location of informative facial features necessary to identify these emotional expressions differs. Given strong suggestions that the eyes area is of particular relevance for sadness (e.g., Calvo and Nummenmaa, 2008), and to achieve a balanced design with a similar number of emotions for which recognition has been linked to the eyes/upper half of the face (fear, sadness) versus mouth/lower-half of the face (happiness, disgust), expressions of anger were not included in the study.

Following Rigoulot and Pell (2012), we chose to analyze three gaze measures: location of first fixations, frequency of (total) fixations, and duration of (total) fixations. First fixations were chosen because some studies found that even

the first saccade can be guided by global information about a scene (Rousselet et al., 2005) or the emotional expression of a face (Becker and Detweiler-Bedell, 2009; Eisenbarth and Alpers, 2011). We predicted that the location of first fixations would be located to the eyes (Hall et al., 2010) and *not* influenced by prosody since we did not find matching effects for first fixations in our previous study (Rigoulot and Pell, 2012). We also analyzed the number and/or duration of fixations in two different time windows, one concurrent with the utterance ([0–2500 ms]) and one after the offset of the utterance ([2500–5000 ms]), based on Rigoulot and Pell's (2012) finding that these measures reveal an emotional matching effect, both during and after the display of auditory stimuli. We predicted that the frequency and duration of looks would increase towards emotionally relevant regions of the face that marked *congruent* rather than incongruent information with the prosody (eyes and brows of fearful and sad faces, mouth and nose of happy and disgusted faces).

2. Methods

2.1. Participants

The participants were 24 native English speakers (12 men/12 women, mean age: 20.6 ± 3.6 years) who were recruited through campus advertisements. All were right-handed with normal hearing and normal/corrected-to-normal vision, as determined through self-report. Informed written consent was obtained from each participant prior to their involvement in the study, which was reviewed and ethically approved by the Faculty of Medicine Institutional Review Board at McGill University (Montréal, Canada).

2.2. Stimuli

Materials consisted of short emotionally inflected utterances and photographs of faces with different emotional expressions. All prosodic and facial stimuli were selected from two existing databases of exemplars that have been validated and successfully used in previous work (Pell et al., 2009 for sentences; Tottenham et al., 2009 for faces).

2.2.1. Auditory stimuli

Given our interest in how prosody (and not semantic information) influences gaze behavior to facial cues, the auditory stimuli presented were emotionally-inflected pseudo-utterances that contain grammatical features of English, but no lexical-semantic information that could be used by listeners to understand emotions (e.g., *Someone migged the pazing* spoken in an angry voice; see Scherer et al., 1991; Pell and Baum, 1997 for earlier examples). As described elsewhere (Pell et al., 2009), the same pseudo-utterances were produced by two male and two female speakers (amateur actors) to portray a range of vocal emotions; these utterances were digitally recorded (single channel) in a sound-attenuated booth, split to present the mono signal to both ears, saved as individual audio files, and then perceptually validated by a group of 24 native listeners in a seven forced-choice emotion recognition task.

Based on the data of Pell et al. (2009), for this study we selected a subset of 96 pseudo-utterances produced by one male and one female speaker that reliably conveyed fear, disgust, happiness, or sadness to listeners (24 pseudo-utterances per emotion, half produced by the female and half produced by the male speaker). As noted above, these four emotions were selected due to purported differences in how humans scan corresponding facial expressions of these emotions (Adolphs et al., 2005; Bassili, 1979; Calder et al., 2000). We carefully selected these sentences by ensuring that the emotional meaning of the prosody for all items was recognized by at least 80% of participants in the validation study (Pell et al., 2009) and that there was minimal confusion with other emotions. Statistically, there was no significant difference in mean target recognition for the selected stimuli across the four emotional prosody types, $F(3, 33) = 1.606$; $p = 0.256$). Given that speech rate is a critical parameter for encoding emotional distinctions in the vocal channel (Juslin and Laukka, 2003; Pell et al., 2009), the mean duration of pseudo-utterances conveying each of the four emotional meanings varied to some degree, ranging from around 1.5 to 2.4 s (for fear and disgust, respectively). The major perceptual and acoustic properties of auditory stimuli selected for the study are furnished in Table 1 by emotion type.

2.2.2. Visual stimuli

All facial expressions used in the study were color photographs (506×650) selected from the NimStim database (Tottenham et al., 2009). We selected 96 emotional faces posed by twenty-four actors (12 female, 12 male), each actor posing four distinct expressions depicting fear, sadness, disgust, and happiness. Individual items were chosen on the basis of the percentage of recognition of the emotional expression for each face (Tottenham et al., 2009) and based on data collected from 16 young participants who did not take part in the main study. The pilot study was run to ensure that our selection of stimuli was also appropriate for Canadian English participants.[1]

2.3. Experimental design/procedures

Participants were invited to take part in a study of "communication and emotion". They were seated in a quiet, dimly lit room at a 75 cm distance from the computer screen. Stimuli were presented on a View Sonic P95f monitor with Intel Pentium 4 computer.

[1] Sixteen Canadian English participants (8 men and 8 women, mean age: 24.9 ± 3.9 years) were asked to identify the emotional expression of the selected faces. Their accuracy was high (80% of correct responses) which reinforced our selection of faces from NimStim database.

Table 1
Major perceptual and physical parameters of the emotional pseudo-utterances presented in the experiment.

	Emotion			
	Fear	Sad	Disgust	Happy
% Recognition	84 ± 6	85 ± 4	83 ± 6	82 ± 7
Pitch mean (Hz)	238 ± 22	133 ± 10	135 ± 8	154 ± 12
Pitch range (Hz)	135 ± 43	86 ± 24	96 ± 28	78 ± 11
Duration (ms)	1470 ± 178	1691 ± 291	2403 ± 167	1730 ± 295

Eye-movements were recorded with an Eye Link II eye tracking system (head mounted video-based; SR Research, Mississauga, Ontario, Canada) connected to an Intel Core2Duo computer (2.79 GHz). The sampling rate of the eye tracker was 500 Hz. Experiment Builder software (SR Research) was used for stimulus presentation. The eye tracker was calibrated at the onset of testing and whenever needed during administration of the experiment. The calibration was accepted if the average error was less than 0.5° in pupil-tracking mode.

To construct trials, individual auditory and visual stimuli were paired, ensuring that the sex of the speaker always matched the sex of the face, although one facial expression could be paired with sentences uttered by different speakers and one sentence could be paired with different actors posing the facial expression. Each of the 96 pseudo-utterances appeared four times in the experiment, paired with an exemplar of each of the four emotional facial expressions (fear, sadness, disgust, happiness) posed by different actors. This yielded a total of 384 trials in the experiment. Each trial began with a centrally located visual marker that participants were asked to fixate, allowing for drift-correction of the eye-tracker. When the participant's eye was fixated on the circle, the experimenter initiated the trial. A random delay of 100–300 ms (ms) was inserted at the beginning of each trial to prevent anticipatory saccades. Then, a single face appeared on a grey background for 5000 ms while a pseudo-utterance was presented to both ears at a comfortable listening level over headphones (the onset of the auditory and facial stimuli was precisely synchronized; see Fig. 1). Headphones were used to ensure a high quality emotional signal while eliminating potential noise and distractors during the study.

Participants were informed that they would hear a "nonsensical" sentence as each face appeared and that their task was to judge whether the emotional meaning of the speaker's voice matched the face; this ensured that participants paid attention to both the auditory and visual stimulus during each trial. Participants pressed labeled YES/NO keys on a two-button response box (the position of yes and no response buttons was counterbalanced across participants). If they did not answer in the 5000 ms of display of the face, the next trial was triggered and the response of the participant was considered as incorrect. The auditory stimuli averaged approximately 1800 ms in duration, whereas the face was always presented for 5000 ms. At the end of each trial, a blank screen appeared for 1000 ms and the next trial was triggered. Participants completed eleven practice trials before each recording session to acquaint them with the experimental procedures and to expose them to features of the stimuli. The experiment lasted approximately 1 h, after which participants were debriefed about the purposes of the study and paid for their participation ($20 CAD). Given the high number of stimuli and the fatigue that could have been induced, a mandatory break was inserted in the middle of the experiment and participants were also free to ask for additional breaks whenever they needed (the eye-tracker was recalibrated after each break).

2.4. Data analyses

Data for 21 participants (11 women and 10 men; mean age: 20.9 ± 3.5 years old) were considered in all statistical analyses; data for three participants (2 males, 1 female) could not be analyzed due to a technical issue. We analyzed the accuracy of participants' behavioral responses by running a 4 × 2 multivariate ANOVA (MANOVA) with repeated measures of emotional expression of the face (fear, disgust, happy, sad) and the matching status of the prosody (matching, mismatching). The multivariate approach to repeated measurements with one measure is similar to a univariate ANOVA but avoids problems with sphericity (Vasey and Thayer, 1987).

For eye-tracking measures, we analyzed the location of first fixations following trial onset, and the frequency and duration of fixations to specific regions of interest (ROI) on the face in two successive temporal windows: from the onset to 2500 ms after the display of the auditory/facial stimulus [0–2500 ms]; and from 2500 ms to 5000 ms after the onset of the stimuli [2500–5000 ms]. The first time window was chosen to investigate the *concurrent* influences of prosody on the scanning of the face, as all sentences were fully played during the first 2500 ms. Gaze behavior in the second time window allowed inferences about the *remote* effects of the prosody on fixation patterns, after the speech stimulus had been fully played, given our previous finding that prosody can have long-lasting effects on face processing (Rigoulot and Pell, 2012).

To analyze the eye-tracking data, we defined four rectangular areas of each face that were constant for each actor/identity across emotional expressions, but defined individually for each actor due to the physical particularities of each face selected from the NimStim database.

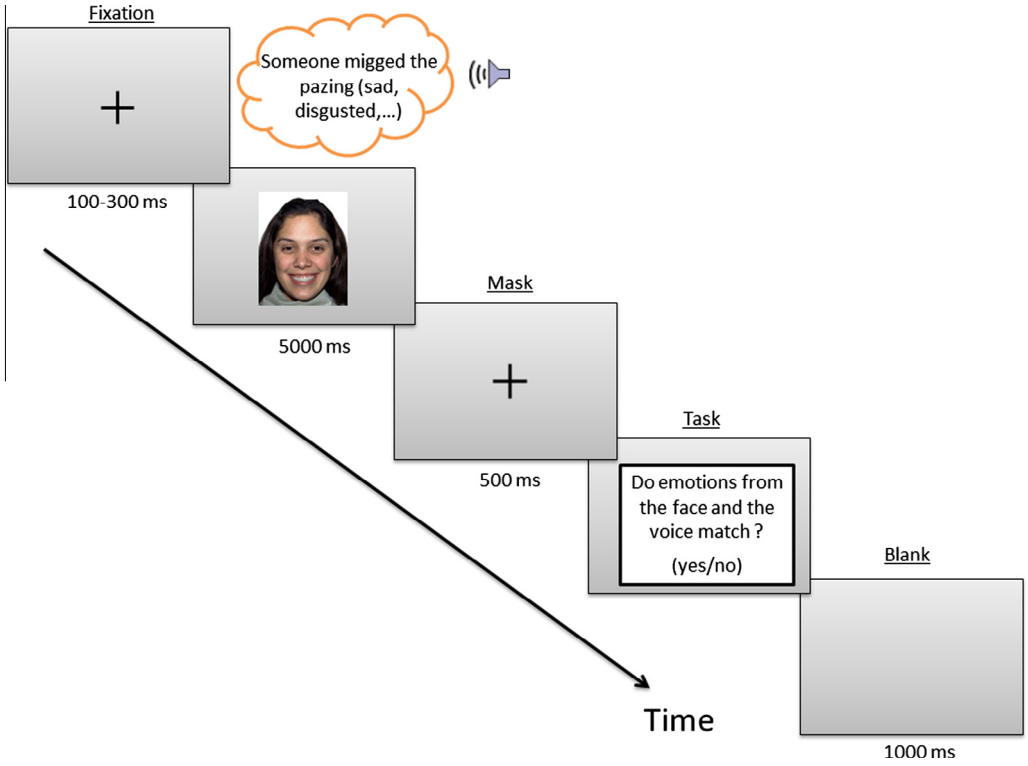

Fig. 1. Illustration of a trial sequence.

These four areas surrounded the mouth, the right eye, the left eye, the brows (see Fig. 2 for an example). Prior to statistical analyses, data for the right and left eye were averaged to create a single ROI labeled "eyes". In addition, we defined a fifth ROI that followed the boundaries of the nose.[2] The frequency and duration of fixations to all seven target cells were automatically generated by Data Viewer for each trial. The frequency of fixation is the number of times the participant stopped or had a saccade in any face ROI. The duration of the fixation was the time spent on an ROI. First fixation was identified for each trial as the earliest fixation on any part of interest of the face. Whenever participants looked at two different locations in the same target cell in a row, this was counted as two different fixations (with different durations).

Gaze measures (first fixations, total fixations, and mean duration of fixations) were analyzed in separate $4 \times 4 \times 2$ MANOVAs with repeated measures of emotional expression type (fear, sadness, disgust, happiness), ROI (brows, eyes, nose, mouth), and matching status of the prosody (match, mismatch). Prior to these analyses, the frequency or duration of looks to mismatching faces was averaged for each participant since there were three mismatching and one matching face in each trial. Analyses of eye movements were performed only on trials that yielded a correct behavioral response to ensure that scanning patterns referred to instances where participants recognized the (matching or mismatching) target emotion of both stimuli. Post hoc comparisons (Tukey's HSD, $p < .05$) were conducted when a significant main or interactive effect was observed.

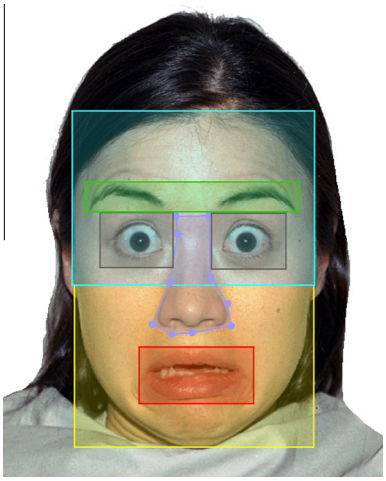

Fig. 2. To define the regions of interest, each face was split into seven areas, blue: upper part of the face; green: brows; black: left and right eyes; purple: nose; yellow: lower part of the face; red: mouth). The frequency and the duration of fixations were recorded as a function of their location within the face. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

[2] For the purpose of the more global analysis shown in Footnote 4, we defined two larger areas including previous features, the upper part of the face (above the tip of the nose) and the lower part of the face (below the tip of the nose).

3. Results

3.1. Behavioral performance

Overall, the ability of participants to judge the matching status of the emotional voice and conjoined face was good (80.3%). The statistical analysis (repeated measure MANOVA) showed two main effects (emotional facial expression: $F(3, 18) = 20.842$; $p < 0.001$; matching status of the voice: $F(1, 20) = 15.787$; $p < 0.001$) and a significant interaction between these factors ($F(3, 18) = 7.380$; $p < 0.001$; see Fig. 3). Post hoc analyses (Tukey's HSD) revealed that for each type of emotional face except disgust, the accuracy of the participants was higher when the emotional content of the voice matched rather than mismatched the emotional expression of the face (fear: $p = 0.009$; sadness: $p = 0.001$; disgust: $p = 0.667$; happy: $p = 0.003$).

3.2. Eye gaze measures by facial region of interest

Table 2 supplies data showing the location of first fixations to each cell in the visual array, as well as the mean frequency and duration of total fixations to each cell calculated in each of two temporal windows ([0–2500 ms], [2500–5000 ms]), according to the facial region of interest (ROI), facial expression type, and emotional prosody type.

3.2.1. Analysis of first fixations

The 4 (Face type) × 4 (face ROI) × 2 (prosody matching status) MANOVA on the location of first fixations revealed a main effect of ROI ($F(3, 18) = 85.554$; $p < 0.001$) and a significant interaction of facial expression type and ROI ($F(9, 12) = 14.025$; $p < 0.001$). Further analyses revealed that first fixations were more frequent to the eyes and nose, than to the mouth and the brows, for all emotional faces except disgust; whereas participants initially fixated on the eyes more frequently than the mouth for fear, sadness, and happiness ($p < 0.001$), there was no difference in the number of first fixations to the eyes versus mouth for expressions of disgust ($p = 0.826$; see Table 2). The interaction of face and prosody matching status was also significant for this analysis ($F(3, 18) = 3.997$; $p < 0.05$). Post hoc analyses revealed that when prosody matched the face, first fixations to fearful faces were significantly more numerous than to sad faces ($p = 0.029$). No other main or interactive effects involving prosody were observed for first fixations.[3]

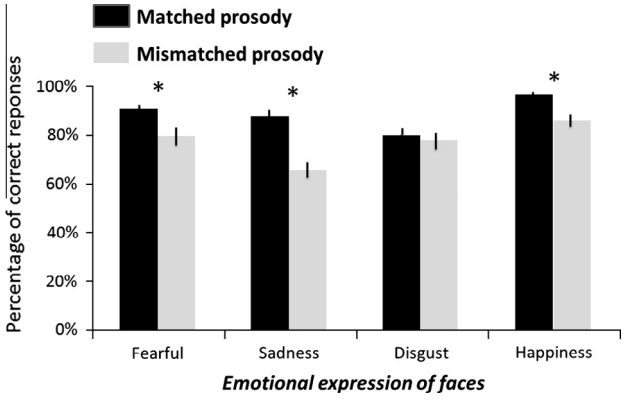

Fig. 3. Ability of participants to judge the matching status of prosody-face pairs, mean percent correct (for matching and mismatching trials as a function of facial expression type, error bars refer to SEM).

3.2.2. Analysis of early time window [0–2500 ms]

3.2.2.1. Frequency of total fixations. The 4 × 4 × 2 MANOVA revealed significant main effects of emotional face type ($F(3, 18) = 22.550$; $p < 0.001$), ROI ($F(3, 18) = 53.206$; $p < 0.001$), and prosody matching status ($F(1, 20) = 24.741$; $p < 0.001$). Generally speaking, participants looked most frequently at the nose region during the task than at the eyes and mouth regions (which also differed marginally, $p = 0.056$; eyes received more looks than the mouth). Total looks to the brow region were significantly fewer than to any other ROI, $ps < 0.041$). In terms of emotion, participants tended to look at happy and fearful faces significantly more often than at the other expressions ($ps < 0.001$). Interestingly, faces that matched the emotional meaning of concurrent prosodic cues were looked at more frequently overall than when prosody conflicted with the facial expression.

The main effects are informed by significant two-way interactions of face type × ROI, ($F(9, 12) = 11.084$; $p < 0.001$) and Face type x matching status, ($F(3, 18) = 12.141$; $p = 0.001$), and a three-way interaction of Face type, ROI and prosody matching status ($F(9, 12) = 2.992$; $p = 0.040$). Step-down analyses compared how the number of looks at each face type varied according to the ROI in the face as a function of prosody matching status (see Fig. 3). Since looks to the brow region were always fewest and were never significantly influenced by the prosody matching status according to the step-down analyses for each facial expression, these data are omitted below for expository purposes. The 4 × 2 step-down analysis on each face type revealed a marginally significant interaction of ROI and prosody matching status for fear ($F(3, 18) = 3.013$; $p = 0.057$) and sadness ($F(3, 18) = 2.960$; $p = 0.054$), and a significant interaction for happiness ($F(3, 18) = 3.944$; $p = 0.025$).

When participants saw a fearful face, they always looked more frequently at the nose/eye regions than at the mouth irrespective of prosody status; however, hearing a congruent fearful voice caused them to look significantly more often at the nose/eyes than when the prosody conveyed a conflicting meaning ($ps < 0.001$). There was

[3] An analysis of the *duration* of first fixations could not be reliably performed for these data since there were relatively few first fixations outside the eyes. For example, the mean number of first fixations to the brows across emotional faces was 0.97 and more than 50% of participants did not look at the brows for their first fixation. A lack of data points therefore precluded this analysis.

Table 2
Frequency of first fixations, frequency of total fixations, and mean fixation duration (in millisecond), measured during the display of the face according to facial expression type, emotional prosody type, and facial region of interest (congruent prosody-face conditions are marked by shaded cells).

		Measure																			
		First Fixation (frequency)				Fixations (frequency) [0 – 2500 ms]				Fixations (duration) [0 – 2500 ms]				Fixations (frequency) [2500 – 5000 ms]				Fixations (duration) [2500 – 5000 ms]			
Facial Features	*Facial Emotion*	Emotional Prosody																			
		Fear	Sad	Disgust	Happy	Fear	Sad	Disgust	Happy	Fear	Sad	Disgust	Happy	Fear	Sad	Disgust	Happy	Fear	Sad	Disgust	Happy
Brows	*Fear*	18	15	13	15	125	109	85	121	324	290	361	299	190	144	130	204	395	364	369	396
	Sadness	17	23	31	34	236	278	75	242	272	315	335	302	245	271	97	262	397	372	336	380
	Disgust	36	37	24	28	242	171	230	241	340	324	358	371	259	210	265	245	430	399	368	429
	Happiness	20	24	32	24	164	198	194	210	302	316	326	310	195	230	167	272	368	402	411	373
Eyes	*Fear*	177	186	190	197	810	661	607	768	333	325	358	348	830	659	677	777	393	410	404	406
	Sadness	163	156	154	144	608	684	254	594	350	345	327	355	574	683	263	621	409	404	414	405
	Disgust	101	94	110	86	441	314	433	413	350	349	329	340	455	292	416	429	407	410	393	417
	Happiness	142	156	151	157	545	668	555	735	369	350	358	351	547	648	645	716	434	436	435	430
Mouth	*Fear*	75	68	76	59	410	290	309	336	301	320	326	322	369	310	300	267	387	380	382	407
	Sadness	52	59	48	54	216	292	108	276	315	337	315	313	200	285	108	231	389	407	355	397
	Disgust	121	112	101	106	417	362	427	450	332	357	344	339	370	342	449	378	426	456	408	444
	Happiness	86	79	90	81	331	407	385	498	364	341	344	347	325	359	326	394	430	417	423	430
Nose	*Fear*	157	162	133	129	1128	933	863	1080	330	342	331	346	691	548	582	664	403	442	403	414
	Sadness	187	171	198	163	897	998	399	1022	333	334	348	345	544	679	303	624	421	434	408	428
	Disgust	170	172	174	176	922	787	858	947	352	349	350	360	588	510	651	603	431	452	409	432
	Happiness	168	158	158	136	895	1043	856	1009	354	344	353	338	518	678	567	638	450	433	462	431

no influence of the prosody status on looks to the mouth region ($ps > 0.226$), implying that congruent prosodic information only increased fixations to the upper (eyes and nose) region of the face for fear. Similarly, for sad faces participants made significantly more looks to the eyes and nose region than to the mouth (eyes–nose > mouth, $ps < 0.001$), but looks *increased* to the eyes and nose ($ps < 0.001$) but not to the mouth ($ps > 0.157$) when a matching sad prosody was heard. This again signifies that a matching, sad prosody promoted greater looks to upper (eye/nose) regions of sad faces than when prosody conveyed a different emotional meaning.

For happy faces, looks were also more frequent to the nose and eyes overall than to the mouth region, consistent with observations for fear and sad faces. However, when accompanied by happy prosody, fixations to the eyes *and* mouth both increased significantly in number than when prosody conveyed a different meaning. There was no evidence that prosody status influenced looks to the nose ($ps > 0.139$). This pattern suggests that happy prosody promoted greater looks to both upper and lower regions of the face when matching happy faces were encountered (see Fig. 4). For disgust, the step-down analysis only yielded a main effect of ROI ($F(3, 18) = 37.206$; $p < 0.001$), explained by more frequent fixations of the nose than other ROIs when disgust faces were scanned ($ps < 0.001$). There was no main or interactive effect of prosody matching status for disgust (F's < 1.180, $ps > 0.345$).

3.2.2.2. Mean duration of fixations. The $4 \times 4 \times 2$ MANOVA performed on the mean duration of looks to each face revealed only a significant main effect of ROI ($F(3, 18) = 36.043$; $p < 0.001$). In general, mean fixations to the eyes were significantly longer in the first time window than to all other regions of the face ($ps < 0.001$). There was no evidence that prosody influenced gaze as a function of

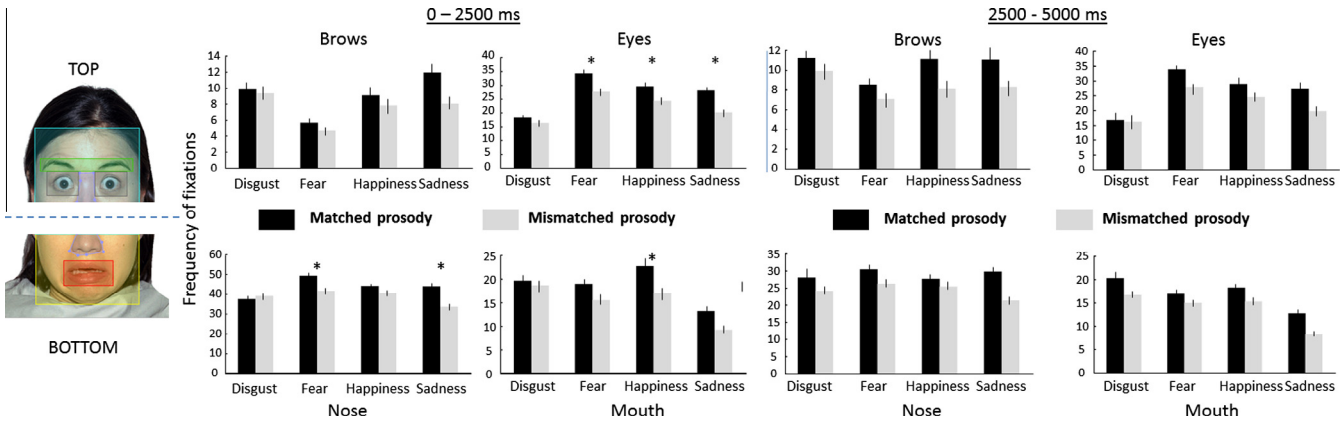

Fig. 4. Mean number of fixations per trial as a function of the ROI of the face, brows, eyes, nose, mouth) for each facial expression type, according to the matching status of the prosody, *: $p < 0.05$). Please note that significant contrasts are only shown for the effects of interaction between facial expression and matching status of the prosody.

the other variables, although the main effect of face type was a strong trend in the data ($F(3, 18) = 2.591$; $p = 0.085$), showing that fixations tended to be longer overall for happy faces.[4]

3.2.3. Analysis of late time window [2500–5000 ms]

3.2.3.1. Frequency of total fixations. The same 4 × 4 × 2 MANOVA performed on data in the "late" time window, following presentation of the emotional pseudo-utterance, yielded significant main effects of emotional face type ($F(3, 18) = 19.617$; $p < 0.001$), ROI ($F(3, 18) = 17.809$; $p < 0.001$), and the prosody matching status ($F(1, 20) = 33.653$; $p < 0.001$). With only slight differences from the first analysis, participants looked significantly more often overall in the second time window at: the nose and eyes region (than mouth or brows); at happy faces than all other facial expressions; and at faces that *matched* versus mismatched the emotion conveyed by preceding prosodic information. In addition, the analysis produced a significant two-way interaction of Face type x ROI ($F(9, 12) = 9.865$; $p < 0.001$). Post hoc exploration of the two-way (Face type × ROI) interaction showed that the nose and eyes were looked at significantly more often than the mouth ($ps < 0.001$) for all facial expressions except disgust; for disgust faces, the nose was fixated most frequently ($ps < 0.001$) but there was no difference in the frequency of looks at the eyes and the mouth ($p = 0.945$).

3.2.3.2. Mean duration of fixations. The 4 × 4 × 2 MANOVA performed on the mean duration of looks in the second temporal window revealed two significant main effects for ROI ($F(3, 18) = 85.725$; $p < 0.001$) and for the emotional facial expression ($F(3, 18) = 5.088$; $p = 0.010$). Again, fixations to the eyes were significantly longer on average than to all other face regions ($ps < 0.001$). The main effect of face type was explained by significantly longer fixations to happy faces on average than to sad or fearful faces ($ps < 0.036$), an effect that was only marginally significant based on analyses of the early time window.

4. Discussion

Previous studies have shown that the visual scanning of a face varies as a function of its emotional expression (Bate et al., 2009; Green et al., 2003). In light of evidence that emotional prosody modifies visual attention to a congruent versus incongruent emotional face (Paulmann et al., 2012; Rigoulot and Pell, 2012), this study used eye-tracking to test whether emotional meanings of speech prosody influence the way we scan *specific regions* of a related emotional face (Calder et al., 2000; Gosselin and Schyns, 2001). As anticipated, we found that gaze measures (first fixations, frequency of total fixations) were systematically modulated by the emotional relationship between vocal emotion cues in speech and the emotional meaning of a conjoined face. Eye movements to particular face regions (especially the eyes) were significantly influenced by both the emotional expression of the face and by its relationship to the prosody, although not entirely as predicted by previous work on this topic. These patterns are discussed in more detail below.

[4] As some previous studies have defined their regions of interest in different ways (e.g., Calder et al. (2000), we reanalyzed our data to broadly evaluate how gaze patterns varied to the upper versus lower half of faces (irrespective of specific features). A 2 × 4 × 2 MANOVA with repeated measures of ROI (upper, lower), emotional face type (fear, sadness, disgust, happiness), and prosody matching status (match, mismatch) was run separately on first fixations, total number of looks, and mean duration of total looks. Main and interactive effects involving the new ROI factor were then examined. Results showed that first fixations were preferentially directed to the upper rather than lower part of the face for all emotional expressions (Emotion × ROI, $F(3, 18) = 16.764$, $p < .001$), and that the mean duration of total fixations to the upper versus lower part of the face was marginally longer (ROI main effect, $F(1, 20) = 4.230$, $p = .053$). The analysis of total looks yielded a significant three-way interaction ($F(3, 18) = 3.993$, $p = .024$). The frequency of looks to the lower part of the face was significantly fewer for sad expressions when compared to all other emotion types; also, when prosody matched versus mismatched the face, looks were always more frequent to both the lower and upper parts of the face except in one condition (looks to the upper part of disgusted faces).

4.1. Effects of emotion congruency on behavioral judgements

Our task required participants to explicitly evaluate whether or not the emotional content of vocal and facial features matched for each trial. We found that participants were accurate at making this judgment, confirming that they focused their attention to emotional features of both stimuli during the task. As anticipated, behavioral performance was significantly better in most cases when emotional information *matched* across sensory modalities than when it mismatched; these results reinforce that information extracted from the auditory and visual channels interacts significantly during the processing of emotion (e.g., Brosch et al., 2009; Campanella and Belin, 2007; De Gelder and Vroomen, 2000; Dolan et al., 2001; Pourtois et al., 2005). Moreover, behavioral performance is often enhanced by congruent bimodal (e.g., audio-visual) stimuli (Collignon et al., 2008; Paulmann and Pell, 2011).

Exceptionally, when disgust faces were accompanied by a disgust voice, no enhancement of emotional judgments was observed in the match versus mismatch condition, and participants were generally less accurate when judging disgust trials compared to the other face types overall (see Fig. 2).[5] A recent cross-modal priming study that presented vocal and facial expressions of three negative emotions (anger, disgust, sadness) noted similar idiosyncrasies for disgust in how speech cues influenced the accuracy and latency of decisions about disgust faces (Jaywant and Pell, 2012). These patterns could relate to increased difficulties recognizing disgust in the vocal channel (Paulmann and Pell, 2011; Pell and Kotz, 2011), despite the fact that we presented stimuli that were all known to be highly representative of the target emotion. There could also be strong attentional biases to disgust stimuli that influenced our behavioral data (Charash and McKay, 2002; Cisler et al., 2009). However, since behavioral decisions were generally accurate for all emotion types (including disgust), and gaze measures were restricted to correctly identified trials, these distinctions should not impede an understanding of how prosody modulates eye movements to different regions of a (disgusting) face.

4.2. Effects of facial expression type on gaze measures

Gaze measures provide an *implicit* index of how participants look at emotional faces, with or without accompanying speech cues. Irrespective of the quality of the speech stimulus, we noted that the first fixation of participants was directed more frequently towards the eyes/nose (upper) part of the face than to the mouth (lower) part of the face. In addition, participants dwelled significantly longer on the eyes, overall, than on any other region of the face in each temporal window analyzed, emphasizing that the eyes are a highly salient part of the face that provide critical information when engaged in emotional face processing. This result is in line with previous work showing that the eyes convey distinctive information about the emotional expression of a face (Matsumoto, 1989) and that adults tend to direct first fixations toward the eye region when judging the meaning of facial expressions (e.g., Hall et al., 2010). The fact that we also observed a high number of fixations to the *nose* when scanning emotional faces, while initially surprising, is not unusual in the literature (e.g., Bate et al., 2009) and could further emphasize the importance of the nose region in (emotional) face processing (see Jack et al., 2009; Neath and Itier, 2014). Given that foveal vision covers around two degrees of the visual field, and extrafoveal (peripheral) vision is known to be sensitive to emotional cues (Bayle et al., 2009; Calvo and Lang, 2005; Rigoulot et al., 2011, 2012), it is likely that many nose fixations included informative features in adjacent regions of the face, especially the eyes, as a strategy for processing and integrating several facial features at one time.

The strong tendency to focus initially on the eyes/upper face did not hold true for disgust; uniquely for this emotion, the relative number of first fixations to the mouth and to the eyes did not significantly differ. In addition, our data reveal increased fixations to the mouth region when participants were presented disgust faces (and to a lesser extent happy faces) than when other face types were presented, and there was a strong trend for participants to simultaneously look *longer* at the mouth region overall, but only for happy faces (review Table 2). These patterns imply that there was more equal distribution of looks to the mouth versus eyes in the case of disgust (and to some extent happiness). This result is in line with Calder et al.'s (2000) data arguing that the lower part of the face is of relatively greater importance when categorizing faces of disgust (see also Gosselin and Schyns, 2001; Schyns et al., 2002). Thus, while it is true that first fixations are directed mainly to the eye regions of a face—irrespective of accompanying speech cues—it seems that the eyes may not provide sufficient cues to correctly identify disgust (and possibly happy) faces. This may have led participants to selectively shift their gaze towards more informative regions located in the lower part of these facial expressions, yielding a higher proportion of first fixations and total looks to the mouth.

Interestingly, our data for the "late" time window also show that mean gaze durations for fearful and sad faces were shorter on average than for happy faces, an effect reported in previous work (Paulmann et al., 2012; Rigoulot and Pell, 2012). As we uncovered no evidence at all that prosody influenced the *duration* of fixations to particular faces regions, these gaze duration patterns likely

[5] Interestingly, when we split out all the possible combinations of an emotional face with an emotional prosody (4 facial expressions × 4 emotional prosodies), we found that the accuracy of the participants was good overall, except for three combinations of face and voice. Participants tended to perform poorly when sad or fearful faces were paired with disgust prosody, and when disgust faces were paired with sad prosody, suggesting that high error rates were often associated in some way with disgust stimuli (see discussion).

refer to basic visual attentional processes (approach/withdrawal tendencies) that humans have developed in response to rewarding, appetitive, or pleasant stimuli; emotional faces associated with potentially negative outcomes tend to trigger avoidance behaviors/shorter fixations, whereas pleasant stimuli linked to positive social outcomes, such as happy faces, promote sustained attention with increased dwell times (Becker and Detweiler-Bedell, 2009; Hunnius et al., 2011; Stins et al., 2011). These seemingly automatic responses governing *how long* participants fixated on different face types were unaffected by processing emotional prosody at the same time, although this was not true when the frequency of fixations was inspected.

4.3. Effects of emotional prosody on gaze measures

In addition to basic differences in how participants scanned emotional faces, our data confirm that listening to emotional speech influences the visual analysis of faces in systematic ways. Participants looked more frequently at faces that communicated the same rather than a different emotional meaning as the voice, although there was no corresponding change in the mean duration of looks according to prosody status. The fact that emotional prosody modulated gaze behavior according to the underlying meaning of each stimulus (i.e., congruent versus incongruent) serves to replicate and extend conclusions of recent eye-tracking studies using related stimuli and methods (Paulmann et al., 2012; Rigoulot and Pell, 2012). Our current findings are also in line with broader studies using the visual world paradigm (Cooper, 1974; Tanenhaus et al., 1995) that highlight various congruency effects on visual scanning patterns, for example, showing that participants tend to fixate a picture in an array that is semantically related versus unrelated to a spoken word (Dahan et al., 2001; Huettig and Altmann, 2005; Yee and Sedivy, 2006; Yee et al., 2009). Here, the mechanism by which voice-face congruency modulates the number of looks to a target face is presumed to reflect unconscious processes that index shared conceptual knowledge about emotions, that are activated simultaneously from prosody and facial displays in our task (see Niedenthal, 2007; Pell et al., 2011 for related commentary).

Of main theoretical interest in this study, the meaning of prosody had a significant impact on how participants looked at specific face regions for different expression types when the frequency of total fixations was analyzed (significant three-way interaction of face type × ROI × prosody matching status). Although it was noted that looks to the eyes and nose region were more frequent than to the mouth region for all emotions except disgust, we found that participants looked significantly *more* at the eyes when listening to a matching fear voice (versus mismatching prosody). Similarly, increased looks to the nose/eyes were selectively observed when sad faces were presented with a matching sad prosody. These tendencies were significant in the early time window when both stimuli were simultaneously presented, and remained a strong trend in the late time window after speech cues were no longer available. These patterns suggest that congruent prosodic information reinforces tendencies to scan the upper part of the face for fear and sad expressions, in line with the idea that cues to recognize these emotions are more informative in upper face regions (Calder et al., 2000).

For happiness and disgust, the effects of prosody on gaze patterns did not fall neatly in line with claims that critical information for recognizing these two emotions is distributed in the lower part of the face (Gosselin and Schyns, 2001; Schyns et al., 2002). For happiness, we found that participants uniquely distributed their attention more evenly to upper and lower regions of happy faces, and that matching prosody produced increased fixations of *both* the mouth and the eye regions; this suggests that a congruent happy prosody promoted greater attention to both upper and lower parts of happy faces, not just the lower part of the face. For disgust, the effect of prosody was negligible with no clear evidence that gaze was selectively biased towards lower regions of the face, at least in the first temporal window of analysis. For happy facial expressions, it can therefore be said that congruent prosodic information amplifies eye movements to different parts of a face in a more general manner than previously characterized (Gosselin and Schyns, 2001; Schyns et al., 2002). This explanation is consistent with the idea that the eyes are typically the most important region for recognizing facial expressions of emotion, but in the case of happiness at least, secondary analyses of the mouth region may be necessary to identify these expressions. This idea fits with recent data reported by Eisenbarth and Alpers (2011), who found that the ratio of fixation durations to the upper versus lower part of emotional faces was smallest (i.e., most equally distributed) for happy faces, as well as with recent data showing that the mouth region of happy faces (i.e., a smile) biases emotion recognition to a greater extent than the mouth region of other emotional expressions when the eye and mouth regions of facial expressions were manipulated and paired in various ways (Calvo and Fernández-Martín, 2013). Important to this study, our analyses show that these emotion-specific scanning patterns for happy faces can be triggered or *enhanced* when listening to congruent speech prosody conveying happiness.

4.4. On the time course of the effects of emotional prosody

In terms of the time course, effects of speech prosody on gaze appear to be emotion specific in the first time window when prosodic and facial cues overlapped, promoting specific fixation patterns that may help to efficiently process and recognize specific face targets. However, the specificity of these effects was mitigated in the second time window where congruent prosody continued to influence fixation patterns over incongruent prosody for all face types, but not to specific face regions. These results confirm data obtained by Rigoulot and Pell (2012) showing that pros-

ody-face congruency effects persist for a certain time, even after the auditory stimulus is no longer present.

In our previous investigation, the influence of matching/mismatching prosody on fixations to one of four emotional faces in an array was analyzed in three separate time windows; an initial temporal window when the prosodic and face stimuli were simultaneously present, and two subsequent time windows that allowed the remote effects of prosody to be explored while participants passively viewed the face array. In the late temporal windows, Rigoulot and Pell (2012) found that the frequency and duration of looks at fearful faces (but not anger, happy or neutral faces) were influenced by a matching (fearful) prosody even after the auditory stimulus was not available; these results were interpreted in light of the importance of events related to fear and threat which may be kept in mind longer than other types of information. Here, the influence of prosody congruency in the late time window extended to all four emotional expressions (fear, sadness, happiness, disgust). This apparent discrepancy is probably related to two main factors that differed between studies: the attentional focus of the participants and the absence of competition between emotional faces during each trial. In this study, attentional resources of our participants were explicitly focused on the emotional *relationship* of the face and the voice, whereas in Rigoulot and Pell (2012) participants passively performed trials, meaning that the evaluation of the emotional tone of the voice was made implicitly. The fact that we directed participants' attention here to the underlying emotional features of the speech and face stimuli, coupled with the absence of competition between emotional faces in this study (we presented only one target face versus four faces per trial in Rigoulot and Pell, 2012), may well explain why the effect of prosody-face congruency was observed more broadly, affecting all four emotion types in the late time window in this study. Clearly, more studies are needed to fully understand the mechanisms of these congruency effects and how task instructions and the degree of competition among different exemplars serve to modulate them over time.

4.5. Limitations of the study

Two main factors limit what conclusions can be drawn from this study. The first one is directly related to the selection of our stimuli; our facial stimuli consisted of static displays of emotional facial expressions, whereas our auditory stimuli were semantically-anomalous pseudo-utterances (to rule out semantic effects on emotion processing). These stimuli are limited in their ecological validity when compared to audio-visual or video stimuli, for example. Thus, although our study constitutes an important step in the understanding of cross-modal effects in the domain of emotion processing, future studies using more ecological stimuli are clearly needed. A second methodological issue raised by our study relates to the *duration* of the emotional utterances we presented, which naturally varies as a function of emotion type (Pell and Kotz, 2011), and how this variability may have influenced visual scanning patterns over time. Fearful utterances are usually spoken fast and have a short duration, whereas utterances conveying disgust tend to be significantly longer, which was true of our stimulus set (although cf. Rigoulot and Pell, 2012 who controlled utterance duration across emotions). Given our goal of presenting utterances that were not truncated in any manner, it is unclear how differences in the duration of utterances across the four emotion categories affected particular gaze measures, or the timing of eye movements, in the first time window where this variation occurred; more studies using shorter time analysis windows or which control stimulus exposure in various ways are needed. Nonetheless, since our first time window exceeded the duration of all vocal stimuli (2500 ms), we can confidently argue that the patterns we observed reflect the effects of processing emotional meanings of prosody on how listeners scan a face, even if the precise evolution of these effects in the first time window cannot be determined with precision.

5. Conclusion

During social communication, speech prosody and facial cues interact and any incongruence between channels can hinder the processing of what is said (Swerts and Krahmer, 2008). Similar interactions can be observed when facial and vocal expressions convey emotion, which motivated the current study of how speech prosody affects visual attention and scanning of a face. Based on measures of eye movements, our new data partially support arguments about the importance of upper versus lower regions of the face in the recognition of discrete emotional expressions (e.g., Calder et al., 2000) *and* our prediction that congruent prosodic information selectively reinforces gaze patterns to these critical face regions. Based on our results, it seems that emotional cues in the voice promote and amplify a visual scanning pattern that is best suited to render a decision about the congruency of stimuli in each sensory modality, although this pattern is more complex than previously envisioned and largely affects the *frequency* of fixations to different parts of an emotional face. Further studies are needed to demonstrate how emotional cues in speech could facilitate visual attention to salient visual markers of the face that permit a holistic impression to be formed about the emotion communicated *between* sensory modalities (the explicit requirement of our task). Our study uncovers further ways that speech prosody is likely to guide eye movements and visual attention to other socially-relevant cues in the immediate environment, a skill that is paramount for understanding speaker intentions during interpersonal communication.

Acknowledgements

We are grateful to Catherine Knowles and Hope Valeriote for running the experiment. This research was

funded by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (to MDP).

References

Adolphs, R., Gosselin, F., Buchanan, T.W., Tranel, D., Schyns, P., Damasio, A.R., 2005. A mechanism for impaired fear recognition after amygdala damage. Nature 433 (7021), 68–72.

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

Bassili, J.N., 1979. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. J. Pers. Soc. Psychol. 37 (11), 2049–2058.

Bate, S., Haslam, C., Hodgson, T.L., 2009. Angry faces are special too: evidence from the visual scanpath. Neuropsychology 23 (5), 658–667.

Bayle, D.J., Henaff, M.-A., Krolak-Salmon, P., 2009. Unconsciously perceived fear in peripheral vision alerts the limbic system: a MEG study. PLoS ONE 4 (12), e8207.

Beaudry, O., Roy-Charland, A., Perron, M., Cormier, I., Tapp, R., 2014. Featural processing in recognition of emotional facial expressions. Cogn. Emot. 28 (3), 416–432.

Becker, M.W., Detweiler-Bedell, B., 2009. Early detection and avoidance of threatening faces during passive viewing. Quart. J. Exp. Psychol. 62 (7), 1257–1264.

Blais, C., Roy, C., Fiset, D., Arguin, M., Gosselin, F., 2012. The eyes are not the window to basic emotions. Neuropsychologia 50, 2830–2838.

Brosch, T., Grandjean, D., Sander, D., Scherer, K.R., 2009. Cross-modal emotional attention: emotional voices modulate early stages of visual processing. J. Cogn. Neurosci. 21 (9), 1670–1679.

Bruce, V., Burton, A.M., Craw, I., 1992. Modelling face recognition. Philos. Trans. Roy. Soc. Lond. B Biol. Sci. 335 (1273), 121–127 (Discussion 127).

Calder, A.J., Young, A.W., 2005. Understanding the recognition of facial identity and facial expression. Nat. Rev. Neurosci. 6, 641–651.

Calder, A.J., Young, A.W., Keane, J., Dean, M., 2000. Configural information in facial expression perception. J. Exp. Psychol. Hum. Percept. Perform. 26 (2), 527–551.

Calvo, M.G., Fernández-Martín, A., 2013. Can the eyes reveal a person's emotions? Biasing role of the mouth expression. Motiv. Emot. 37 (1), 202–211.

Calvo, M.G., Lang, P.J., 2005. Parafoveal semantic processing of emotional visual scenes. JEP: Hum. Percept. Perform. 31 (3), 502–519.

Calvo, M.G., Marrero, H., 2009. Visual search of emotional faces: the role of affective content and featural distinctiveness. Cogn. Emot. 23 (4), 782–806.

Calvo, M.G., Nummenmaa, L., 2008. Detection of emotional faces: salient physical features guide effective visual search. J. Exp. Psycho.: Gen. 137 (3), 471–494.

Calvo, M.G., Nummenmaa, L., 2011. Time course of discrimination between emotional facial expressions: the role of visual saliency. Vis. Res. 51 (15), 1751–1759.

Campanella, S., Belin, P., 2007. Integrating face and voice in person perception. Trends Cogn. Sci. 11 (12), 535–543.

Charash, M., McKay, D., 2002. Attention bias for disgust. J. Anxiety Disord. 16 (5), 529–541.

Cisler, J.M., Olatunji, B.O., Lohr, J.M., Williams, N.L., 2009. Attentional bias differences between fear and disgust: implications for the role of disgust in disgust-related anxiety disorders. Cogn. Emot. 23 (4), 675–687.

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., et al., 2008. Audio-visual integration of emotion expression. Brain Res. 1242, 126–135.

Cooper, R.M., 1974. The control of eye fixation by the meaning of spoken language. A new methodology for the real-time investigation of speech perception, memory, and language processing. Cogn. Psychol. 6, 84–107.

Cvejic, E., Kim, J., Davis, C., 2010. Prosody off the top of the head: prosodic contrasts can be discriminated by head motion. Speech Commun. 52 (6), 555–564.

Dahan, D., Magnuson, J.S., Tanenhaus, M.K., 2001. Time course of frequency effects in spoken-word recognition: evidence from eye movements. Cogn. Psychol. 42 (4), 317–367.

Darwin, C., 1872. The Expression of the Emotions in Man and Animals, third ed. Philosophical Library, New York (London: Harper Collins, New York: Oxford University Press, 1998).

De Gelder, B., Vroomen, J., 2000. The perception of emotions by ear and by eye. Cogn. Emot. 14 (3), 289–311.

Dolan, R.J., Morris, J.S., de Gelder, B., 2001. Crossmodal binding of fear in voice and face. Proc. Natl. Acad. Sci. USA 98 (17), 10006–10010.

Eisenbarth, H., Alpers, G.W., 2011. Happy mouth and sad eyes: scanning emotional facial expressions. Emotion 11 (4), 860–865.

Ekman, P., Friesen, W.V., 1976. Pictures of Facial Affect. Consulting Psychologists Press, Palo Alto, CA.

Ekman, P., Davidson, R.J., Friesen, W., 1990. The Duchenne smile: emotional expression and brain physiology: II. J. Pers. Soc. Psychol. 58, 343–353.

Ekman, P., Friesen, W.V., Hager, J., 2002. The Facial Action Coding System. London, UK.

Gordon, M.S., Hibberts, M., 2011. Audiovisual speech from emotionally expressive and lateralized faces. Quart. J. Exp. Psychol. 64 (4), 730–750.

Gosselin, F., Schyns, P.G., 2001. Bubbles: a technique to reveal the use of information in recognition tasks. Vis. Res. 41 (17), 2261–2271.

Green, M.J., Williams, L.M., Davidson, D., 2003. In the face of danger: specific viewing strategies for facial expressions of threat? Cogn. Emot. 17 (5), 779–786.

Hall, J.K., Hutton, S.B., Morgan, M.J., 2010. Sex differences in scanning faces: does attention to the eyes explain female superiority in facial expression recognition? Cogn. Emot. 24 (4), 629–637.

Huettig, F., Altmann, G.T.M., 2005. Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. Cognition 96 (1), B23–B32.

Hunnius, S., de Wit, T.C.J., Vrins, S., von Hofsten, C., 2011. Facing threat: infants' and adults' visual scanning of faces with neutral, happy, sad, angry, and fearful emotional expressions. Cogn. Emot. 25 (2), 193–205.

Jack, R.E., Blais, C., Scheepers, C., Schyns, P.G., Caldara, R., 2009. Cultural confusions show that facial expressions are not universal. Curr. Biol. 19, 1543–1548.

Jaywant, A., Pell, M.D., 2012. Categorical processing of negative emotions from speech prosody. Speech Commun. 54 (1), 1–10.

Juslin, P.N., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: different channels, same code? Psychol. Bull. 129 (5), 770–814.

Malcolm, G.L., Lanyon, L.J., Fugard, A.J.B., Barton, J.J.S., 2008. Scan patterns during the processing of facial expression versus identity: an exploration of task-driven and stimulus-driven effects. J. Vis. 8 (8) (Art. 2).

Matsumoto, D., 1989. Face, culture, and judgments of anger and fear: do the eyes have it? J. Nonverbal Behav. 13 (3), 171–188.

Messinger, D.S., Mattson, W.I., Mahoor, M.H., Cohn, J.F., 2012. The eyes have it: making positive expressions more positive and negative expressions more negative. Emotion 12 (3), 430–436.

Neath, K., Itier, R.J., 2014. Facial expression discrimination varies with presentation time but not with fixation on features: a backward masking study using eye-tracking. Cogn. Emot. 28 (1), 115–131.

Niedenthal, P.M., 2007. Embodying emotion. Science 316 (5827), 1002–1005.

Palermo, R., Rhodes, G., 2007. Are you always on my mind? A review of how face perception and attention interact. Neuropsychologia 45 (1), 75–92.

Paulmann, S., Pell, M.D., 2010. Contextual influences of emotional speech prosody on face processing: how much is enough? Cogn., Affect. Behav. Neurosci. 10 (2), 230–242.

Paulmann, S., Pell, M.D., 2011. Is there an advantage for recognizing multi-modal emotional stimuli? Motiv. Emot. 35, 192–201.

Paulmann, S., Titone, D., Pell, M.D., 2012. How emotional prosody guides your way: evidence from eye movements. Speech Commun. 54, 92–107.

Pell, M.D., 1999a. The temporal organization of affective and non-affective speech in patients with right-hemisphere infarcts. Cortex 35 (4), 455–477.
Pell, M.D., 1999b. Fundamental frequency encoding of linguistic and emotional prosody by right hemisphere-damaged speakers. Brain Lang. 69 (2), 161–192.
Pell, M.D., 2005. Prosody-face interactions in emotional processing as revealed by the facial affect decision task. J. Nonverbal Behav. 29 (4), 193–215.
Pell, M.D., Baum, S.R., 1997. Unilateral brain damage, prosodic comprehension deficits, and the acoustic cues to prosody. Brain Lang. 57 (2), 195–214.
Pell, M.D., Kotz, S.A., 2011. On the time course of vocal emotion recognition. PLoS ONE 6 (11).
Pell, M.D., Paulmann, S., Dara, C., Alasseri, A., Kotz, S.A., 2009. Factors in the recognition of vocally expressed emotions: a comparison of four languages. J. Phonet. 37 (4), 417–435.
Pell, M.D., Jaywant, A., Monetta, L., Kotz, S.A., 2011. Emotional speech processing: disentangling the effects of prosody and semantic cues. Cogn. Emot. 25 (5), 834–853.
Pourtois, G., de Gelder, B., Bol, A., Crommelinck, M., 2005. Perception of facial expressions and voices and of their combination in the human brain. Cortex 41 (1), 49–59.
Rigoulot, S., Pell, M.D., 2012. Seeing emotion with your ears: emotional prosody implicitly guides visual attention to faces. PLoS ONE 7 (1), e30740.
Rigoulot, S., D'Hondt, F., Defoort-Dhellemmes, S., Despretz, P., Honoré, J., Sequeira, H., 2011. Fearful faces impact in peripheral vision: behavioral and neural evidence. Neuropsychologia 49, 2013–2021.
Rigoulot, S., D'Hondt, F., Honoré, J., Sequeira, H., 2012. Implicit emotional processing in peripheral vision: behavioral and neural evidence. Neuropsychologia 50, 2887–2896.
Rigoulot, S., Fish, K., Pell, M.D., in press. Neural correlates of inferring speaker sincerity from white lies: an event-related potential source localization study. Brain Res..
Rousselet, G., Joubert, O., Fabre-Thorpe, M., 2005. How long to get to the "gist" of real-world natural scenes? Vis. Cogn. 12 (6), 852–877.
Scherer, K.R., Banse, R., Wallbott, H.G., Goldbeck, T., 1991. Vocal cues in emotion encoding and decoding. Motiv. Emot. 15 (2), 123–148.
Schyns, P.G., Bonnar, L., Gosselin, F., 2002. Show me the features! Understanding recognition from the use of visual information. Psychol. Sci. 13 (5), 402–409.
Stins, J.F., Roelofs, K., Villan, J., Koojiman, K., Hagenaars, M.A., Beek, P.J., 2011. Walk to me when I smile, step back when I'm angry: emotional faces modulate whole-body approach-avoidance behaviors. Exp. Brain Res. 212, 603–611.
Swerts, M., Krahmer, E., 2008. Facial expression and prosodic prominence: effects of modality and facial area. J. Phonet. 36, 219–238.
Tanaka, A., Koizumi, A., Imai, H., Hiramatsu, S., Hiramoto, E., de Gelder, B., 2010. I feel your voice: cultural differences in the multisensory perception of emotion. Psychol. Sci. 21 (9), 1259–1262.
Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C., 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268, 1632–1634.
Thompson, L.A., Malloy, D.M., LeBlanc, K.L., 2009. Lateralization of visuospatial attention across face regions varies with emotional prosody. Brain Cogn. 69 (1), 108–115.
Tottenham, N., Tanaka, J.W., Leon, A.C., McCarry, T., Nurse, M., Hare, T.A., et al., 2009. The NimStim set of facial expressions: judgments from untrained research participants. Psychiat. Res. 168 (3), 242–249.
Vasey, M.W., Thayer, J.F., 1987. The continuing problem of false positives in repeated measures ANOVA in psychophysiology: a multivariate solution. Psychophysiology 24 (4), 479–486.
Vassallo, S., Cooper, S.L., Douglas, J.M., 2009. Visual scanning in the recognition of facial affect: is there an observer sex difference? J. Vis. 9 (3) (Art. 11).
Wong, B., Cronin-Golomb, A., Neargarder, S., 2005. Patterns of visual scanning as predictors of emotion identification in normal aging. Neuropsychology 19 (6), 739–749.
Yarbus, A., 1967. Eye Movements and Vision. Plenum Press, New York.
Yee, E., Sedivy, J.C., 2006. Eye movements to pictures reveal transient semantic activation during spoken word recognition. J. Exp. Psychol. Learn. Mem. Cogn. 32 (1), 1–14.
Yee, E., Overton, E., Thompson-Schill, S.L., 2009. Looking for meaning: eye movements are sensitive to overlapping semantic features, not association. Psychon. Bull. Rev. 16 (5), 869–874.
Yuki, M., Maddux, W.W., Masuda, T., 2007. Are the windows to the soul the same in the east and west? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. J. Exp. Soc. Psychol. 43, 303–311.