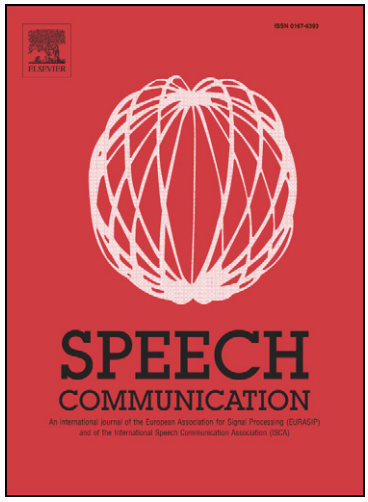
How emotional prosody guides your way: evidence from eye movements

Silke Paulmann, Debra Titone, Marc D. Pell



How emotional prosody guides your way: evidence from eye movements

Silke Paulmann, Debra Titone, Marc D. Pell

PII: S0167-6393(11)00111-7
DOI: 10.1016/j.specom.2011.07.004
Reference: SPECOM 2009

To appear in: *Speech Communication*

Received Date: 25 December 2010
Revised Date: 8 July 2011
Accepted Date: 8 July 2011

Please cite this article as: Paulmann, S., Titone, D., Pell, M.D., How emotional prosody guides your way: evidence from eye movements, *Speech Communication* (2011), doi: 10.1016/j.specom.2011.07.004

How emotional prosody guides your way: evidence from eye movements

Silke Paulmann[1] , Debra Titone[2,4], & Marc D. Pell[3,4]

[1] Department of Psychology, University of Essex, Colchester, United Kingdom

[2] Department of Psychology, McGill University, Montréal, Canada

[3] School of Communication Sciences and Disorders, McGill University, Montréal, Canada

[4] McGill Centre for Research on Language, Mind and Brain, Montréal, Canada

Corresponding author:

Silke Paulmann

University of Essex

Department of Psychology

Wivenhoe Park

Colchester

C04 3SQ

United Kingdom

phone: +44- (0)1206-873422

fax: +44- (0)1206-873801

email: paulmann@essex.ac.uk

1. Introduction

Over the last decades, numerous studies have documented how the brain processes spoken language. Interestingly, this research has focused primarily on how the neuro-cognitive system operates on *what* was said (semantic meaning), rather than *how* something was said (prosody), despite the fact that prosodic information can be equally important for interpreting the meaning of speech. For instance, imagine that you and your partner enter a room where your four-year old child is sitting surrounded by a pile of toys, and you say, "Now look at this mess". Depending on whether you use an angry tone of voice or a pleasant, admiring voice when you speak, your spouse is likely to interpret your utterance and react in distinct ways (e.g., by surveying the mess and preparing an appropriate response, or by stopping to admire your child among the toys). This example emphasizes that emotional prosody plays a central role in pragmatic language comprehension (Wilson & Wharton, 2006), but so far there has been little research whether emotional prosodic cues are instrumental for guiding basic eye movements during (social) information processing.

The purpose of the present study was to address if listeners make immediate use of emotional prosodic cues during a visual search task. Specifically, we explore whether listeners detect congruent information between auditory and visual input information, and if so, how that governs their eye movements. The paradigm adopted also allowed us to explore if emotional prosodic cues are instrumental when generating anticipatory predictions during on-line speech comprehension.

1.1 Emotional processing within and across modalities

1.1.1. Visual modality:

Much of the literature on emotional information processing has focused on stimuli presented in the visual modality (e.g., emotional scenes or faces). In particular, the visual search paradigm

has been used by researchers to understand how we detect and recognize emotional features of facial expressions. These investigations have repeatedly shown that emotionally expressive faces in an array are detected very rapidly (see Frischen, Eastwood, Smilek, 2008 for recent review). Interestingly, some research suggests that negative faces are detected systematically faster and/or more accurately than positive or neutral faces (Hansen & Hansen, 1988; Eastwood et al., 2001; Oehman et al., 2001; Horstmann, 2007). For instance, Eastwood and colleagues presented participants with emotional and neutral schematic faces and asked them to detect targets among distractors. They reported faster reaction times to detect angry faces among neutral distractors than to detect happy faces among neutral distractors (Eastwood, Smilek, Merikle, 2001). These data replicated results reported by Oehman and colleagues (2001) but ruled out a possible emotion and animacy confound present in the latter experiment. Thus, while these results argue for an advantage of detecting negative over positive emotional faces, other data imply that emotion-specific influences on visual search are related to contextual features in which a target face is presented, rather than reflecting true "processing advantages" for specific facial expressions of emotion (Carroll & Russell, 1996; Frischen et al., 2008). Irrespective of this debate, the combined results of these experiments argue that emotion-specific details of facial expressions are detected rapidly and possibly pre-attentively (e.g. Oehman et al., 2001) which generally reinforces the idea that rapid detection of emotional meaning from faces, and possibly other emotional stimuli, is necessary to anticipate upcoming events (c.f. Frischen et al., 2008).

Next to behavioral methodologies, neurophysiological measurements have also been applied. In particular, event-related brain potential (ERP) correlates of emotional and neutral face processing have been explored (e.g. Eimer & Holmes, 2002; Ashley, Vuilleumier, Swick, 2004; Batty & Taylor, 2003; Paulmann & Pell, 2009). These studies have identified different early ERP components that are sensitive to emotional versus neutral faces. For instance, the P200 component (a positive deflection with a maximum peak ~ 200 ms after face onset), has been shown to be

stronger for emotional in contrast to neutral faces (e.g. Eimer & Holmes, 2002; Paulmann & Pell, 2009), implying that emotional-relevant details are extracted rapidly during face processing. In addition, slightly later occurring negative ERPs (maximum peaks occurring ~240 ms after face onset) have also been reported to be differently modulated for different emotional expressions (e.g. Eimer, Holmes, McGlone, 2003). This negative deflection has been argued to reflect more detailed perceptual analysis during face processing (e.g. Eimer et al., 2003; Paulmann & Pell, 2009). The advantage of ERPs over behavioral methodologies lies in their excellent temporal resolution and it is crucial to note that peak latencies for these early components do not differ between neutral and emotional facial expression, suggesting a similar processing time-course for early processing stages of emotional and non-emotional faces.

While ERPs can establish the time-course of processing stages involved in face processing, eye tracking allows continuous monitoring of eye movements in response to emotional picture presentation, thereby furnishing insight as to how emotional cues are used during social information processing. Recently, Calvo & Nummenmaa (2008) explored eye movement patterns in response to emotional and neutral facial expressions applying a visual search task. Participants always viewed a face from one emotional category (happy, angry, sad, disgusted, surprised, and fear) among six neutral faces and both response latencies and eye movements were measured during target face detection. Results from their experiment 2 revealed that happy facial expressions were fixated and localized earlier than other emotional facial expressions due to their visual saliency (e.g. smile). Also, the authors report a correlation between first fixation and accuracy in detection, that is, the earlier a happy face was fixated and localized, the faster the face was detected among neutral distractors by participants as reflected in faster behavioral response latencies. Moreover, pictures showing surprised or disgusted facial expressions were also fixated and localized earlier than fearful, angry, or sad facial expressions, suggesting that eye movements to emotional facial expressions depend on saliency features. These findings were replicated in Calvo, Nummenmaa,

Avero (2008). Here, the authors also presented faces in an inverted fashion to confirm that physical distinctiveness or saliency (possibly revealing emotional content) helps participants find the correct face during visual search.

1.1.2. Auditory modality:

While the majority of research on emotional expression processing has focused on visual emotional expressions (e.g. faces), vocal emotional expressions have also received some attention over the past years. For instance, several studies have revealed that emotions can successfully be recognized from speech (e.g. Banse & Scherer, 1996; Paulmann, Pell & Kotz, 2008) even in the absence of emotional semantic cues (e.g. Scherer et al., 2001; Pell et al., 2009a,b). The majority of investigations on emotional prosody demonstrate high recognition rates for emotional stimuli (e.g. often four times higher than expected by chance; see Pittam & Scherer, 1993). Interestingly, and comparable to visual expressions, not all emotional categories are recognized equally well (e.g. anger and sadness are often easier to recognize than fear or pleasant surprise; Banse & Scherer, 1996; Johnstone & Scherer, 2000). Differences in recognition rates are likely due to their distinctiveness at the acoustic level, again comparable to distinctiveness of visual expressions (see above). Some behavioral studies have looked at the influence of emotional prosody on emotional word meaning processes. For instance, Nygaard and Lunders (2002) investigated how lexical ambiguity may be disambiguated by means of emotional prosody. They presented listeners with emotional homophones (e.g. dye/ die) spoken in a neutral or emotional tone of voice. Listeners were then asked to transcribe the homophones and results indicated that emotional meanings of homophones were more often transcribed when the homophone was spoken in an emotional tone of voice, suggesting that prosody can provide a context to disambiguate lexically ambiguous words. Similarly, Kitayama & Ishii (2002) report an influence of emotional prosody on emotional word meaning processes in an emotional Stroop task. In an emotional stroop task participants listen to

emotional words (e.g. “smile”--happiness) spoken in an emotionally congruent (e.g. happy) or incongruent (e.g. sad) tone of voice. They then have to categorize or identify either the emotional word meaning while ignoring emotional prosody, or have to identify the emotional prosody of the presented word while ignoring its emotional meaning. Differences between judging emotionally congruent and emotionally incongruent words are argued to reflect involuntary access to task-irrelevant feature (e.g. prosody) of the word. Influence of emotional prosody on judging the emotional meaning of a word is thus taken as an indicator that emotional prosodic cues are processed involuntarily.

Behavioral findings have been complemented by electrophysiological evidence that emotional speech prosody is processed in a rapid, highly automatic[1], and involuntary manner (Vroomen, Driver, de Gelder, 2001; Vroomen & de Gelder, 2000; Paulmann & Kotz, 2008; Schirmer Striano, Friederici, 2005; Sauter & Eimer, 2009). For instance, Schirmer and colleagues presented emotional and neutral syllables in a mismatch negativity (MMN) paradigm. The authors report differentiation between emotional and neutral syllables arguing for pre-attentive processing of emotional and neutral prosody. Other event-related brain potential (ERP) studies suggest that affective details such as arousal (Paulmann & Kotz, 2006), valence (Schirmer et al., 2005; Paulmann & Kotz, 2008; Sauter & Eimer, 2009), and discrete emotion attributes of vocal expressions (Paulmann & Pell, 2010) can be indexed within 200 ms after speech onset. For instance, we have shown that the P200 component is differentially modulated by vocal expressions of six different basic emotions (anger, disgust, fear, sad, happy, pleasant surprise) when compared to neutral expressions. While it can be argued that the P200 component primarily reflects emotional salience detection and thereby fails to directly provide evidence that emotional category *meaning* can be extracted this quickly, we recently showed that even brief exposure (e.g. 200 ms) to

[1] Here, we follow the definition of Calvo & Avero (2008) who define 'automatic' processing as „quickly, efficiently, unintentionally, and/or unconsciously“.

emotional prosodic cues is sufficient to access emotional category meaning from memory (Paulmann & Pell, 2010 and see further below for experiment details).

Interestingly, rapid extraction of emotional prosodic cues is confirmed by studies which suggest that emotional prosodic meanings are often activated *before* emotional semantic information is extracted from the utterance, which is thought to occur approximately 300-400 milliseconds after word onset (Bostanov & Kotchoubey, 2004; Schirmer, Friederici & Kotz, 2002, 2005). Given that sentence meaning interpretation often requires the listener to monitor information until the very end of the sentence (e.g., "look at this mess" vs. "look at this girl"), it is therefore possible that emotional prosody is used earlier than semantic cues to guide sentence interpretation, and/or that emotional prosody is more informative in particular speech contexts or temporal processing intervals as speech unfolds. Indeed, work by Schirmer et al. (2002, 2005) suggests that emotional prosody is used to contextually integrate a word in the same way as semantic information as reflected in N400 differences for visually presented emotional words that were preceded by congruent or incongruent presented auditory emotional sentences. Interestingly, some report *earlier* ERP responses to contextually incongruous vocal expressions than usually observed for contextually incongruous visually presented words (N300; Bostanov & Kotchoubey, 2004).

However, it should be noted that this possible timing advantage does not render emotional prosody to be "more important" during sentence or context interpretation; instead, it has been claimed that semantics can simply not be ignored in particular contexts (e.g. Besson, Magne, & Schön, 2002; Kotz & Paulmann, 2007; Paulmann & Kotz, 2008b; Paulmann et al., 2008; Pell, Jaywant, Monetta, & Kotz, In press). For instance, in a cross-splicing paradigm, we previously explored the integrative time-course of emotional prosody with neutral semantics and of emotional prosody with emotional semantics (Kotz & Paulmann, 2007; Paulmann & Kotz, 2008b). Results revealed a prosodic expectancy positivity in response to sentences that contained an emotional prosodic expectancy violation only, and a negative early N400-like ERP component in response to

sentences that contained an emotional prosodic and semantic expectancy violation. The two distinct ERP components not only suggest different underlying neural mechanisms for emotional prosody and emotional semantic processing but their latency differences also point to different time-course underlying these processes. Finally, given that combined expectancy violations elicited an N400-like response (and not a positivity), we hypothesized that emotional semantic processing may override emotional prosodic processing when both information types interact in time (Kotz & Paulmann, 2007; Paulmann & Kotz, 2008b).

Aside from indications that emotional prosody influences word processing (Bostanov & Kotchoubey, 2004; Nygaard and Lunders, 2002; Kitayama & Ishii, 2002) and sentence interpretation (Schirmer et al., 2002; 2005), there is *cross-modal* evidence that emotional prosody influences decisions about visual events. Reports demonstrate that facial expressions are processed advantageously when preceded by an emotionally-congruent rather than incongruent vocal stimulus (e.g. Carroll & Young, 2005; Paulmann & Pell, 2010; Pell, 2005a,b), and vice versa (de Gelder & Vroomen, 2000; Hietanen, Leppänen, Illi, & Surakka, 2004). Facial expressions of emotion also pre-attentively influence judgments of the emotional connotations of music (Thompson, Russo, & Quinto, 2008). These cross-modal emotional congruence effects in information processing have been linked to the activation of so-called emotion concepts, or emotion-related units in associative memory, which refer to discrete emotion states and are commonly activated by associated events in the auditory and visual modalities (Bower, 1981; Carroll & Young, 2005; Niedenthal & Halberstadt, 1995; Innes-Ker & Niedenthal, 2002; Pell, 2005a,b; Russell & Lemay, 2000). The underlying assumption is that as more information 'primes' an emotion concept, it becomes more accessible, leading to facilitation or preferential processing of emotionally congruent as opposed to incongruent stimuli. This claim fits with recent evidence that emotion recognition tends to be faster and more accurate in situations when multi-modal cues are available (Paulmann, Jessen, Kotz, 2009; Paulmann & Pell, 2011; see Scherer, 1989 for a related discussion).

It should be noted that there is recent evidence which suggests a differential effect of emotion on visual processing. Specifically, Zeelenberg & Bocanegra (2010) explored how auditorially and visually presented emotional and neutral words influence decisions in a two-alternative identification task. In their experiment 1, spoken words were presented to participants and then followed by a visually presented masked target word which was followed by visual presentation of two choice alternatives. Results revealed better identification of neutral target words when auditory cue words were emotional as opposed to neutral. However, in Experiment 2, which presented cue words in the visual and not the auditory modality, emotional cues hampered identification of the visual targets. Taken together, these results suggest differential effects of emotion on within- as opposed to cross-modality processing. This should be taken into account when interpreting effects of emotional prosody on visual search.

Finally, in her graduate thesis, McManus (2010) explored gaze fixation patterns during visual and auditory emotion processing. Participants were presented with short movie-clips showing either congruent or incongruent face voice information (e.g. sentence such as "Look in the box" spoken in a happy tone of voice while the speaker looks happy or angry). While the effect of prosody on visual search was not directly investigated, fixation patterns of participants revealed a strong preference to look at eye-regions of actors during presentation of emotionally congruent movies. Specifically, when listening to/looking at happy or angry movie clips, participants fixated to left eyes, while when listening to/looking at movies conveying fear, participants fixated more often on the right eye of actors. For incongruent movies, the pattern was less straight-forward and varied with task instructions (identify emotion of actor vs. determine if emotional prosody and emotional facial expression are congruent). This suggests that listeners detect congruent information between modalities which in turn influences their eye gaze behavior.

While these studies underline that emotional prosody can impact on visual processing, most of them do not tell us *how* emotional prosody impacts on visual search as speech is processed in

real time; one way to explore this question is to monitor the eye movements (gaze patterns) of participants as they listen to emotionally-inflected speech, to determine whether emotional features of prosody immediately influence gaze patterns in an emotionally-congruent manner.

1.2. Effects of emotional prosody on visual processing

So far, few studies have used the eye-tracking methodology to test whether emotional prosody is implicitly registered by listeners to guide their "actions" during on-line speech processing, i.e., their eye gaze and visual attention to related social cues. One rare study by Berman et al. (2010) explored whether young children can make use of emotional prosodic cues for referential mapping during on-line speech processing. The authors applied a variation of the so-called 'visual-world' paradigm that has been successfully employed in studies investigating the relationship between visual perception, action, and language (e.g. Dahan & Tanenhaus, 2005; Henderson & Ferreira, 2004; Spivey, Typer, Eberhard, & Tanenhaus, 2001), and in the on-line use of *linguistic* prosodic cues such as contrastive stress (e.g. Ito & Speer, 2008; Weber, Grice, Crocker, 2006). In this paradigm, participants' eye movements to visual cues during sentence comprehension are tracked, allowing researchers to make inferences about particular facets of on-line language processing (Tanenhaus, Spivey-Knowlton, Eberhard, Sedivy, 1995). For instance, Allopenna, Magnuson, & Tanenhaus (1998) presented listeners with a visual display of four objects (e.g. beaker, beetle, speaker, carriage) and played participants instructions such as "Pick up the beaker". Results revealed that upon onset of the word "beaker", participants were more likely to fixate on onset competitors (in this case "beetle") than on unrelated distractors (e.g. carriage), showing that eye movements can be used to make inferences about on-line speech comprehension processes.

Similarly, Berman et al. (2010) presented their child participants with an array of three images displaying two objects of the same category but that differed in their underlying emotional meaning (e.g. broken/intact doll) and one object of a different category (horse). Children then

listened to instructions such as “Look at the doll” spoken in a positive, negative, or neutral tone of voice. Results revealed that older children were more likely to look at images displaying objects that matched the tone of voice of the speaker as opposed to objects that mismatched the tone of voice, suggesting that emotional prosody can be used for referential mapping during on-line sentence comprehension. In addition, there is evidence from linguistic prosody that listeners can use linguistic prosodic cues during on-line speech processing. For instance, Ito & Speer (2008) presented participants with an ornament grid while listening to instructions such as “First hang the blue ball, then hang the green angel” to trim a Christmas tree. Their results confirmed that listeners' eye movements are guided by prosodic cues. For instance, more fixations occurred to a green ball (as opposed to green angel) on the grid if the instructions were spoken in such a way that the adjective in the second instruction (green) was stressed (as this led participants to believe that ball will again be the target object). Collectively, studies like these (strictly speaking, the two latter studies are not visual world but rather visual search studies) allow commenting on the temporal dynamics underlying language comprehension mechanisms such as word recognition, or linguistic and emotional prosody processing.

Thus, there is some evidence that both emotional and linguistic prosody is used by listeners for referential resolution (or more precise for referent selection). However, what is unclear from these data is the effect emotional prosodic cues can have on visual attention to related *social cues* such as facial expressions. That is, the question is whether we can find evidence by means of the eye tracking methodology that emotional prosodic cues are immediately used by listeners during a visual search task. Such a finding would help establish that eye-tracking can successfully be applied to investigate *emotional* prosodic processing in *real-time*. In short, we explored whether we can replicate behavioral and electrophysiological evidence in that emotional prosodic cues are a) rapidly extracted during sentence comprehension, and b) used to establish a meaningful emotional context.

Finally, our approach can also be used to establish if emotional prosodic cues can be used to guide listeners' eye movements.

1.3. The present investigation:

In the current study, we monitored participants' eye gaze to an array of facial expressions as they listened to emotionally-intoned sentences — simple commands, such as "Click on the happy face"—spoken in a prosody that was emotionally congruent or incongruent with the face specified in the instructions. We chose to present stimuli from four different emotional categories (anger, dear, sadness, happiness) and a neutral category based on previous research (see sections above). Specifically, it has been argued that these emotional categories can be considered to reflect "basic" emotions (e.g. Ekman, 1992) and they are reported to be recognized universally (e.g. Pell et al., 2009a), thereby emphasizing their validity in experimental use. To assess whether emotional information in the utterance influences eye movements, comparisons between emotional matching and mismatching trials were executed. This approach is similar to studies using priming and off-line behavioral methods (e.g., Pell, 2005a). Equally important, we examined patterns of eye gaze in two distinct time windows: in the initial part of the utterance when only emotional features of prosody could meaningfully guide participants' eye movements (all information preceded the emotional adjective, i.e., "Click on the/…"); and in the latter part of the utterance as emotional word meaning-related information about the target face is processed (e.g., "…/happy face"). By comparing the effects of emotional prosody in two different time windows, our data will shed light on how prosodic cues alone guide visual attention to a matching face (pre-emotional label window), and show whether these effects serve to modulate eye movements dictated by explicit semantic cues that are task-relevant during sentence comprehension (post-emotional label window).

Based on past electrophysiological findings (Schirmer & Kotz, 2006; Paulmann & Kotz, 2008; Paulmann & Pell, 2010; Pell & Skorup, 2008), we hypothesized that emotional prosodic

meanings would be evaluated implicitly and rapidly after speech onset. Building on results reported by Berman et al. (2010), detection of congruent information (i.e. happy prosody, happy facial expression) should result in longer and more frequent eye fixations to faces that match rather than mismatch the emotion of the prosody. These effects should be robustly detected in the pre-emotional label time-window that precedes the onset of the emotional adjective. Alternatively, based on studies that report facilitation effects during congruent emotion information processing (e.g. Zeelenberg & Bocanegra, 2010, Pell, 2005a,b), it could be argued that detection of emotional congruent information should result in shorter and fewer fixations to faces that match versus faces that mismatch the emotional tone of voice. No matter what the direction of effects, differences between emotionally congruent and emotionally incongruent information processing during the early time window will help establish that emotional prosody can be extracted rapidly and can immediately be used to influence listeners' eye gaze to related social cues (here facial expressions).

In addition, this early time-window can be used to establish if listeners make use of emotional prosodic cues when developing anticipatory predictions during on-line speech processing. Results reported by Berman and colleagues (2010) suggest that this is not necessarily the case, as pre-schoolers showed no effect of using emotional prosodic information to anticipate which object to fixate on (that is, there were no eye movement pattern differences between conditions before the onset of the noun in the instruction “look at the doll”). This would suggest that listeners can extract emotional prosodic information and can use this information during referent selection, but are not influenced by these cues otherwise. Two points are important to note though: first, participants in Berman et al.'s (2010) study were pre-schoolers and results may reflect their (in)sensitivity to emotional vocal cues; second, images displayed objects rather than socially relevant cues such as facial expressions. Thus, if participants in the present study show evidence of using emotional prosodic cues to anticipate which face they may have to click on next (i.e. happy

prosody could mean that happy face will have to be clicked on), this would provide evidence that anticipatory predictions can be developed during on-line speech processing in adults.

Once listeners encounter lexical-semantic information in the utterance which specifies the target emotion of the face (post-emotional label window), we expected that the influence of emotional prosody on eye movements would diminish since lexical-semantic cues are directly relevant to the task; also, in the face of conflicting cues from prosody and semantics, semantic information should dominate (Besson et al., 2002; Kotz & Paulmann, 2007). However, it is uncertain how quickly influences of prosody on eye movements should dissipate in the post-emotional label time window as the lexical-semantic meaning unfolds.

2. Materials and Methods

2.1. Participants

Twenty native speakers of English participated in the study (10 female; mean age = 21.2 years; mean education = 15.9 years). None of the participants reported any hearing impairments, and all had normal or corrected-to-normal vision. All participants gave informed consent before completing the study, which was ethically approved by the McGill Faculty of Medicine Institutional Review Board. All participants were compensated financially for their involvement.

2.2. Stimulus material

Auditory stimuli - The speech stimuli were simple commands that instructed participants to click on a specific facial expression within a visual array ("Click on the xx face", where "xx" was an emotion term). All utterances were produced by two different professional speakers (one female, one male), digitally recorded in a sound-attenuated booth using a high quality, AKG head-mounted microphone (16-bit, 44.1 kHz sampling rate). The experiment included five different emotion "categories": anger, fear, happiness, sadness, and neutral. Each speaker produced utterances instructing the listener to click on a target face representing each of the five emotion categories,

produced in each of the five tones of voice (5 target emotions x 5 prosodic emotions x 2 speakers = 50 auditory stimuli total). The precise semantic terms used in the instructions were to "Click on the angry/frightened/happy/sad/ or neutral face" (each spoken in an angry, frightened, happy, sad, and neutral prosody). Choice of adjectives was based on previous research (e.g. Pell et al., 2009a,b) Thus, for each emotional category the speaker produced one sentence where the semantic content and prosody matched (e.g. "Click on the happy face" spoken in a happy prosody), three sentences where the semantic content and prosody mismatched (e.g. "Click on the happy face" spoken in an angry, frightened, or sad prosody), and one sentence spoken in a neutral prosody. During the recording session, each speaker produced at least three versions of each of the 25 target utterances and these were entered into a pilot study involving 17 young listeners; these participants were asked to categorize the emotion conveyed by the prosody irrespective of the semantic instruction (5 alternative [anger, fear, sad, happy, neutral] forced-choice task).

Based on these pilot data, the item with the highest accuracy rate based on the prosody for each command was always selected for presentation in the eye-tracking experiment (per speaker). Emotional target accuracy rates for the chosen stimuli were high overall (Mean = 84%, where chance performance in the pilot study was 17%). The selected stimuli were then subjected to acoustic analyses using Praat speech analysis software to characterize their primary acoustic features; as shown in Table 1, acoustic properties of the stimuli were consistent with those reported for angry, frightened, happy, and sad utterances in previous studies of emotional prosody (e.g., Pell et al., 2009b). For example, frightened utterances were produced with the highest fundamental frequency/pitch and with limited pitch variation, whereas sad utterances were produced with the lowest pitch and little pitch variation. Angry and happy utterances both displayed moderate increases in pitch height (mean) although happy utterances exhibited increased pitch variation across the utterance. Frightened utterances (M = 1280 ms) tended to be shorter than angry, sad, and happy utterances (M = 1580 ms, 1540 ms, 1470 ms, respectively) which is also known to be one of

the natural properties for signaling this emotion (Pell et al., 2009b). Acoustical properties for each word (Figure 1a & b), as well as F0 listings for 50ms segments for the beginning of the utterances (Table 2) are also displayed. Example waveforms can be seen in Figure 2.

[Table 1 & 2, Figures 1a, b & 2 about here]

Visual stimuli - The target events were static, cropped facial expressions (black and white photographs; 170 x 220 pixels) posed by three female and three male actors, representing each of the five emotion categories. These stimuli have been used successfully in previous work (e.g., Pell, 2005a,b; Pell & Skorup, 2008) and were again selected based on their emotional properties as determined by a norming study (all selected exemplars were recognized at a consensus level > 85% target recognition, where chance = 12.5%). For each of the six actors, one exemplar was chosen for each of the 5 emotion categories (30 faces total). We focused on these five emotion categories because facial expressions of anger, fear, joy/happiness, and sadness are believed to be basic, universally-recognizable emotions (Ekman, Sorenson & Friesen, 1969), and there is evidence that all five emotion categories, including neutral, are recognized categorically (e.g., Young, Rowland, Calder, Etcoff, Seth, & Perrett, 1997). An example of emotional facial expressions posed by one of the actors is provided in Figure 3.

[Figure 3 about here]

2.3. Experimental Design

Trials were constructed by matching individual auditory stimuli with a visual circular array consisting of six facial expressions posed by the same actor (as shown in Figure 3). Each visual array consisted of one expression of each emotion and two neutral expressions. Since semantic cues in the instruction always dictated the target face for each trial, individual trials in the experiment were coded according to the relationship of the prosody with the face: there were two

experimental conditions (match or mismatch) and one filler condition (trials spoken with a neutral prosody). Note that 'match' and 'mismatch' trials were defined solely by stimuli expressing anger, fear, happiness, and sadness (not neutral). For example, in the match condition the speaker's prosody was emotionally congruent with the face specified in the instruction ("Click on the happy face" spoken in a happy prosody; also, anger-anger, fear-fear, sad-sad). In the mismatch condition, the speaker's prosody was incongruent with the face ("Click on the happy face" spoken in an angry, frightened, or sad prosody). In the neutral condition, the speaker's prosody was always neutral ("Click on the angry/frightened/happy/sad or neutral face" spoken in neutral prosody).

Since there were always three possible mismatches of the prosody to the target for each instruction, utterances in the match condition were paired three times with arrays posed by the three actors who were the same sex as the speaker. The same facial arrays presented in the match condition were then used in the corresponding mismatch condition for the same instruction, although note that these were distinct recordings spoken in the three incongruent prosodies. The position of target facial expressions in the array also was controlled; each match, mismatch, and neutral trial was repeated six times in the experiment, with the target facial expression appearing once in all six possible locations in the circular array. If one again considers the instruction "click on the happy face", this sentence was presented in the experiment a total of 18 times spoken in a matching prosody (3 repetitions of the match stimulus x 6 spatial locations), 18 times in a mismatching prosody (3 unique mismatch stimuli x 6 spatial locations), and six times in a neutral prosody, per speaker. This added to 144 matching and 144 mismatching trials (18 trials x 4 emotions x 2 speakers, per condition) and 132 neutral filler trials, or 420 trials in total which were fully randomized within a single experiment.

2.4. Procedure

Participants were tested in a quiet, dimly lit room. They were seated at a 75 cm distance from a computer screen by means of a chin rest. Eye-movements were recorded with an Eye Link II eye tracking system (head mounted video-based; SR Research) with 500 Hz sampling rate. 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. Participants were instructed to listen to the auditory stimulus and follow the instructions that they heard, responding with a computer mouse. The onset of eye tracking data collection was synchronized to the onset of the critical auditory instruction and was recorded until participants pressed the mouse button.

Each trial began with a centrally located visual marker which participants were asked to fixate. This allowed for drift-correction of the eye-tracker, i.e. only once the participant's eyes fixated the visual marker, the experimenter manually began presentation of the facial array, which appeared on a grey background positioned on a virtual circle equally distributed around the fixation point. The circular face array was presented for 2500 ms, after which the fixation cross returned to the screen for 300 ms (to ensure that participants' eye position was always in the center of the screen before the auditory instructions began). Following the fixation cross, the same face grid was presented again accompanied by the auditory instructions requiring participants to click on a specific facial expression. We presented the circular face array before presenting the array accompanied by the auditory instructions to allow participants to first perceptually analyze the images, and thus establish a perceptual map of the faces (see Dahan and Tanenhaus, 2005). After participants clicked on a facial expression in the array, the next trial was triggered. Participants completed ten practice trials before each recording session, which acquainted them with the experimental procedures and features of the stimuli. The experiment lasted approximately one hour and 15 minutes, which included a self-determined break at the mid-point of the session.

2.5. Data analyses

Fixations to target cells were automatically generated using Data Viewer (SR Research). Two dependent measures of interest were analyzed: gaze duration and number (frequency) of fixations to a target cell. Only correctly answered trials were entered into the statistical analyses (on average, 11.18% of all trials were rejected per participant). There were two main time windows identified within the spoken command which were analyzed separately: a "pre-emotional label" and a "post-emotional label" time window. To investigate the immediate effects of emotional prosody on visual search independent of emotionally-relevant semantic cues, analysis of the pre-emotional label time window considered gaze measures from the onset of auditory instruction until the onset of the emotional adjective in the sentence (as this is the point in time when participants actually know which face they have to click on). In this time window, there is no impact of semantic information, only prosody. The dependent variables of gaze duration and number of fixations were entered into separate 2 x 4 ANOVAs with repeated measures of *match* (match vs. mismatch of the prosody to face) and *prosody* (anger, fear, happiness, sadness). To focus the data on patterns of greatest theoretical interest, neutral prosody filler trials were excluded from the analyses. Analyses also excluded anticipatory eye movements (latency onset < 150 ms (see Matin, Shao, & Boff, 1993)).

In a second analysis (post-emotional label time window), the combined effects of prosody on explicit semantic cues referring to the target face were investigated using the same dependent measures; however, in this time interval, the independent variables needed to code for the relationship of *both* prosody and the semantic instruction as a function of which face was being fixated. The post-emotional label window was defined from the onset of the adjective until the end of the utterance (sentence offsets averaged 900 ms across items). Gaze duration and number of fixations occurring in this time window were entered into separate 4 x 4 repeated measure ANOVAs, with the within-subjects factors of *emotional prosody* (anger, fear, happiness, sadness) and *cues* (eye fixations matched the: semantic instruction (SEM); prosodic cues (PROS); both

semantic and prosodic cues (SEM + PROS); neither semantic nor prosodic cues (NONE). Neutral filler trials were again excluded, as were eye movements with a latency onset < 150 ms and > 900 ms after adjective onset.

3. Results

Behavioral responses were collected solely to ensure that participants listened carefully to the sentences and these data were not subjected to statistical analysis. As expected, behavioural performance in the experiment was very accurate (greater than 90% correct facial target "clicks" on average). Nonetheless, one of the 20 participants was excluded from further analysis due to poor behavioral performance (53% correct clicks), indicative of poor attention to the stimuli or failure to comply with task goals. As noted above, only correctly answered trials were entered into the statistical analyses for the eye tracking data. Mean gaze duration and frequency of looks to faces that matched the speaker's prosody (pre-emotional label time window) or specific combinations of prosodic and semantic cues present in the instruction (post-emotional label time window) are provided in Table 3, according to the emotional meaning of the prosody.

[Insert Table 3 about here]

3.1. Pre-emotional label Time Window:

Analyses in the pre-emotional label time window focused mainly on whether eye fixations matched the value of the speaker's prosody when listening to the instructions. Results for fixation durations revealed a significant main effect of *match*, $F_{\text{Duration}}(1, 18) = 9.17$; $p < .01$, but no such effect was found for number of looks. Overall, participants looked significantly longer at a facial expression that matched the emotional prosody of the speaker than faces that did not match the prosody (277

ms vs. 259 ms), as illustrated in Figure 4a&b. There was no effect of *match* or *prosody* on the frequency of saccades to faces in the pre-emotional label time.

[Figure 4a & b about here]

3.2. Post-emotional label Time Window:

Analyses in the post-emotional label time window (the period following the onset of the emotional adjective) considered the emotional relationship between a face that was fixated and different combinations of speech cues present in this time interval: semantics (SEM); prosody (PROS); both semantics + prosody (SEM+PROS); or none of these cues (NONE). Results indicated a main effect of *cues* which was highly significant for both gaze measures, $F_{\text{Duration}}(3, 54) = 107.34$; $p < .0001$; $F_{\text{Frequency}}(3, 54) = 65.75$, $p < .0001$. Duncan post-hoc comparisons ($p < .05$) among the four cue conditions showed that fixations were longest to a face that matched both the semantics and prosody (306 ms), which were longer than to faces that matched only the semantic cues (294 ms). Fixations in both cue conditions with semantic cues were longer than to a face that matched only the prosody (231 ms) or neither the prosody and semantics (226 ms), which did not differ significantly from each other. In terms of frequency, the mean number of looks was also significantly greater in the SEM + PROS condition (1.15 looks) and in the SEM condition (1.13 looks) than in the PROS and NONE conditions (1.06 vs. 1.04 looks, respectively). For duration only, the main effect of emotional prosody was also significant, $F_{\text{Duration}}(3, 54) = 4.08$; $p < .05$. Fixations were shorter on average when listening to a fearful tone of voice than when listening to any other tone of voice (238 ms vs. average of 250 ms). The main effects of *cues* on gaze duration and frequency in the post-emotional label time window are illustrated in Figure 5a&b.

The ANOVAs performed in the post-emotional label time window also yielded a significant interaction of *cues* x *emotional prosody* for both measures, $F_{\text{Duration}}(9, 162) = 4.79$, $p < .001$; $F_{\text{Frequency}}(9, 162) = 2.64$, $p < .05$**.** Duncan post-hoc tests ($p < .05$) for duration measurements

revealed qualitatively similar but slightly different patterns in how the cue conditions influenced eye movements for angry, frightened, and sad prosody when compared to happy prosody. When speakers sounded angry, frightened, or sad, fixations were longer in the SEM+PROS and SEM conditions (which also differed from each other), when compared to the PROS and NONE conditions (which did not differ from each other). When speakers sounded happy, fixations were longer in the SEM+PROS condition than in the SEM condition or the PROS and NONE conditions. A similar influence of emotional prosody on the effects of cues was found for fixation frequency data: for instructions spoken in an angry tone of voice, there were significantly more fixations in the SEM + PROS condition (1.16 looks) than in the SEM or PROS condition (both 1.1 looks) or in the NONE condition (1.04 looks). The latter also differed significantly from the SEM or PROS condition. For instructions spoken in a frightened or sad tone of voice, more fixations were found in the SEM + PROS as well as SEM condition (1.14/1.16 and 1.13/1.14 looks respectively) than in the PROS or NONE condition (both 1.1/1.0 looks). Finally, for happy instructions, fixations were higher in the SEM+PROS condition (1.17 looks) than in the SEM (1.12 looks) or the PROS and NONE conditions (both 1.0 looks).

In summary, it can be said that semantic cues in the instruction, whether spoken in a congruent or conflicting prosody (i.e., SEM+PROS and SEM conditions, respectively), always promoted more frequent and longer eye movements to a matching face as one would expect. Prosodic cues alone had little influence on gaze measures in the post-emotional label time interval.

[Figure 5a-b about here]

4. Discussion

This study used eye-tracking method to implicitly gauge how listeners use emotional prosodic cues in speech during instructed visual search. Next to assessing how and when emotional prosody can influence eye movements during instructed visual search, we explored whether emotional prosody

is used by listeners to develop anticipatory predictions during on-line speech comprehension. Our data argue that emotional prosody is involuntarily registered by listeners with immediate effects on eye gaze and visual attention, similar to what was recently shown in a study where angry prosody was presented to young listeners in a dot probe task (Brosch et al., 2008). Specifically, we found that young adults made systematically longer eye movements to faces that represent the same emotional meaning of the prosody than faces which represent another basic emotion, in a context when explicit semantic content of speech did not signal which face to look at (i.e., the 'pre-emotional label time window'). As expected, the influence of emotional prosody on eye movements diminished as soon as semantic information relevant to the task goals was encountered during speech processing ('post-emotional label time window'). Below, we discuss the implications of these patterns for an understanding of the on-line effects of emotional speech prosody on social cognition and behavior.

4.1. Pre-emotional label time window: Implicit effects of emotional prosody on eye gaze

Recent eye-tracking experiments show that early saccade programming is automatically influenced by the emotional content of stimuli such as visual scenes (Kissler & Keil, 2008; Nummenmaa, Hyönä, & Calvo, 2009), and separately, that gaze preference can be modulated by such factors as the mood of a participant in an affectively-congruent manner (e.g., Isaacowitz, Toner, Goren, & Wilson, 2008). Along similar lines, our results in the pre-emotional label time window demonstrate an immediate impact of emotional prosody on participants' eye movements, causing them to fixate longer on facial expressions that were emotionally-congruent with a speaker's prosody. These data firmly indicate that the emotional meaning of prosodic cues was analyzed and registered in memory by listeners before the semantic message of the instruction was encountered, despite the fact that prosodic cues in the pre-emotional label window often misled participants to the eventual target selection during the visual search task. The observation that

prosody guided saccades to an emotionally-*congruent* face furnishes new evidence that the meaning of emotional prosody is evaluated implicitly during on-line sentence processing, even in contexts when prosodic cues are not directly relevant to the task (Paulmann & Kotz, 2008; Paulmann & Pell, 2010; Pell & Skorup, 2008; Schirmer et al., 2005; Sauter & Eimer, 2009; Wambacq, Shea-Miller, Abubakr, 2004). In particular, the present findings from the pre-emotion label time window suggest that emotional prosodic information is used during social information processing to predict upcoming events. This is in contrast to recent findings from a study applying a similar paradigm in pre-schoolers (Berman et al., 2010). Although pre-schoolers seem to use emotional prosody during referent selection (e.g. when listening to "Look at the doll" spoken in a positive tone of voice, more fixations were directed towards an intact vs. a broken doll), the authors failed to find evidence suggesting that pre-schoolers used emotional prosodic information to predict which affect-compatible object they had to look at next. As suggested before, the discrepancy between these two findings could be due to the age difference in participants or due to differences in stimuli. Based on evidence provided by Berman et al. (2010), it seems likely that pre-schoolers' use of emotional prosodic information per se is only in its footsteps, i.e. any influence of emotional prosodic cue use before the noun onset cannot yet be expected to occur. Alternatively, it can be argued that emotional prosody cue use during processing of socially less relevant cues like objects does not exhibit the same influence that it does during socially more relevant facial expression processing. Future studies should try to explore this issue further.

Our findings are, however, in line with recent behavioral evidence which reported that faces accompanied by an emotionally congruent voice are recognized with higher accuracy and/or more quickly than faces accompanied by an incongruent voice, even when only one information channel is the focus of attention (de Gelder & Vroomen, 2000; Exp. 2 & 3; Massaro & Egan, 1996; Pell, 2005a,b). These effects imply that there are strong, and perhaps obligatory links between emotion processing mechanisms dedicated to related events in the auditory and visual modality (de

Gelder & Vroomen, 2000; Pell, 2005a). Here, we support and extend these claims by demonstrating that the emotional relationship between prosody and facial expressions influences eye gaze patterns during on-line speech processing, and very shortly after spoken language begins to unfold. Thus, similar to how linguistic features of prosody may be used by listeners to guide visual attention (Ito & Speer, 2008), emotional-prosodic attributes of speech are used not only to interpret the intended contextual meaning of an incoming message (Pell, 2006; Wilson & Wharton, 2006), but to guide concurrent behaviors such as eye movements. Presumably, these responses would tend to facilitate multi-modal emotion integration and social information processing in natural communication settings that are typically characterized by both auditory and visual cues. This means that saccade generation is probably influenced by the emotional content of a visual target stimulus (e.g., Kissler & Keil, 2008), *as well as* by existing emotional meanings already present in memory, such as those activated when concurrently processing emotional attributes of speech while scanning a visual array or scene.

The fact that gaze patterns were governed by the emotional congruency of the two events allows inferences about the nature of emotional representations indexed by the two stimuli (although note that our data were not designed to tell us whether emotional information from auditory and visual stimuli are actually integrated, see de Gelder & Vroomen, 2000 for a discussion). In line with previous suggestions, cross-modal emotional congruency effects are likely to be explained by the co-activation of emotion-related conceptual units in memory, which are associated with both vocal and facial expressions and triggered by prototypical cues about basic emotions in each communication channel (Borod et al., 2000; Bowers, Bower & Heilman, 1993; Bower, 1981; Hansen & Shantz, 1995). Previous data imply that the connections between prosody and related facial expressions index *emotion-specific* details about each event, and that these details constitute the mechanism of priming when participants encounter nonverbal displays of emotion (Pell, 2005a; Paulmann & Pell, 2010; Russell & Lemay, 2000). Consistent with this view, in our

pre-emotional label time window we found that implicit effects of prosody on gaze duration occurred quickly and in an emotion-specific manner, since our trials were defined by the emotional category of the displays, and our effects generalized to each of the four emotion types of interest. The observation that eye movements are guided by the emotional meaning conveyed by prosody is novel and should be investigated further; these congruence effects on visual behavior could be one of the ways that affect, mood, and emotional aspects of an individual's external and internal environment promote bias during information processing (see Fazio, 2001 for a discussion).

The idea that emotional prosody is evaluated very *quickly* after speech onset also fits with recent data. Several reports suggest that emotional prosody is meaningfully processed within the first 200 milliseconds (ms) after word or sentence onset (Paulmann & Kotz, 2008; Paulmann & Pell, 2010; Schirmer et al., 2005; Wambacq et al., 2004). Specific acoustic parameters that mark discrete emotions in speech and music (e.g., intensity, pitch timbre, roughness, etc.) may even be extracted within the first 100 ms of stimulus exposure (Schirmer & Kotz, 2006; Koelsch & Siebel, 2005). Qualitative inspection of the distribution of eye fixations observed in our pre-emotional label time window is consistent with the notion that emotional prosody is analyzed rapidly; we observed a general increase in fixations to emotional faces in the array, and an initial spike in the number of fixations recorded to faces that matched rather than mismatched the emotional prosody, in the region of 260-310 ms following speech onset. If one assumes that prosodic information in the auditory stimulus was acoustically analyzed and evaluated for meaning to some extent within the first 100-200 ms of speech presentation, followed by approximately 150 ms needed to initiate an eye movement (Matin, Shao, & Boff, 1993), this could explain why eye fixations to emotional faces (particularly matching faces) started to increase in the observed time interval. The same rationale applies to knowledge activated by emotional facial expressions: fixation durations were between 200 and 300 ms, and since participants looked longer at emotionally congruent than incongruent faces, it can be assumed that emotional meanings were extracted from the faces within this short

time frame. Certainly, comments on the specific time course for processing emotional prosody (or faces) cannot be made with precision from our analyses. Still, the claim that emotional meanings of prosody are robustly activated in the pre-emotional label time window, as inferred by systematic eye movements to a congruent facial expression in this time period, can be strongly advanced, as future studies seek to clarify the exact timing of these effects.

Given the rapid nature of effects, it needs to be explored which acoustic cues listeners can use to extract the emotionality of the vocal stimulus. We have previously suggested that, comparable to emotional facial expressions, emotionality is not construed from one or two single acoustic cues, but probably from an acoustic-configuration pattern that comprises a range of different acoustic cues (Paulmann & Kotz, 2008; Paulmann & Pell, 2010). Here, results from acoustic analyses suggest that F0 levels at the beginning of the utterances (see Tables 1 &2) are quite distinct, perhaps giving participants first clues about which emotional tone of voice the speaker used. In addition, word duration on the first word ("Click") differed for the different emotions (e.g. frightened tone of voice resulted in shorter word duration than angry, happy, or sad tone of voice). Finally, it can be speculated that voice quality characteristics (often referred to as creaky, harsh, breathy, or whispered tone of voice) are different between the different emotional tones and will help listeners discern which emotion they are listening to. While there are several recent studies exploring acoustical configuration patterns of emotional sentences (e.g. Banse & Scherer, 1996; Paulmann et al., 2008; Pell et al., 2009a,b) as well as of emotional non-verbal expressions (e.g. Sauter et al., 2010) it is yet to be determined which cues listeners use for a rapid detection of emotional salience and/or category membership.

4.2. Post-emotional label time window: Effects of prosody on explicit semantic cues indicating where to look

While our main objective was to clarify whether listeners implicitly use emotional prosody with corresponding effects on their eye movements, it was also of interest how prosodic attributes of a speaker's voice would affect the processing of emotional semantic cues that were directly relevant to the target selection. That is, we asked whether emotional prosody would continue to influence eye movements in the "post-emotional label time window", after onset of the emotional adjective which specified which face to click on.

If one looks at the post-emotional label time window (150 ms after adjective onset to the end of the sentence), it is clear that once lexical-semantic information about emotions is retrieved, the preceding influence of emotional prosody on eye movements is largely mitigated. Consistent with the task requirements, participants invariably looked longer and more frequently to faces that corresponded to the semantic instruction, irrespective of whether the speaker's prosody conveyed the same or a conflicting emotion (that is, in both our SEM+PROS and SEM conditions, saccades were always guided by the meaning of the semantic information). Faces which matched the semantic instruction were always associated with much longer/more frequent looks than faces which matched only the prosody or neither set of auditory cues (the latter two conditions rarely differed). The importance of semantic information in our results is in no way surprising, since our participants were explicitly instructed to visually locate a face that matched the semantic instruction in order to select and "click" on it. However, what we could not predict from existing data was whether the emotion of the prosody would somehow influence patterns of eye movements in the post-emotional label time window beyond those tied to the effects of the semantic information on visual search.

Our data show that the effects of prosody in the post-emotional label time window are very small. There were virtually no differences in fixation measures to faces that matched the semantically-incongruent prosody of the instruction (PROS condition) and faces that matched neither the semantic or prosodic meaning encoded by the utterance (NONE condition), arguing that

emotional prosodic cues can be ignored by listeners during instructed visual search. This idea is supported, not only by the fact that eye movements in the post-emotional label time window were not largely predicted by emotional prosody, but by the observation that our participants performed very well behaviorally (which reflects how well they attended only to the semantic instruction when responding). Previous work indicates that when auditory stimuli contain both prosodic and semantic information about emotion, the meaning of lexical-semantic information is dominant, even if the task focuses attention on emotional prosody (see Besson et al., 2002; Grimshaw, 1998; Kitayama & Ishii, 2002; Kotz & Paulmann, 2007; Pell et al., In press; Schirmer & Kotz, 2003). Still, there are indications that semantic information can also be successfully ignored at times in favor of prosodic meanings when these cues conflict (e.g., Bowers et al., 1987). Thus, while it is likely that semantic cues often dominate prosody when conflicting information is encountered in the two channels, the relevance of particular cues to the task goals is always likely to assume an important role in how both channels are processed and meaningfully compared (Massaro & Egan, 1996).

As well, it cannot be said that the effects of prosody in the post-emotional label time window were completely negligible. In some conditions, we observed a significant, moderating effect of prosody on fixations to a face that matched the semantic instruction; for example, we found that the duration of saccades to a face that matched *both* the prosody and the semantics was significantly greater than to faces that matched only the semantic cues (review Figure 5a). Still, our data raise the question: how can emotional prosody be ignored during speech processing? One answer may be that the emotional prosody is processed more "automatically" than lexical-semantic cues, and therefore is easier to ignore; for example, Wambacq et al. (2004) reported that emotional prosody is analyzed 200 milliseconds earlier in conditions when participants focus on semantic aspects of the utterance, when compared to prosodic aspects of the stimulus. Other data similarly imply that processing emotional prosody occurs independent of attentional control during speech

processing (Pell, 2005a; Vroomen, Driver, & de Gelder, 2001). Another consideration exemplified by our results is that a semantic context for interpreting emotions—even a highly simplified one as used here—takes time to build, whereas prosodic cues are present immediately from speech onset; thus, a listener is already likely to hold representations of what is communicated by emotional prosody before any related meanings are encountered in the semantic channel, which could allow attention to be allocated fully to incoming semantic cues. Finally, there is some evidence that the lexical-semantic channel is generally more helpful or reliable than prosody for recognizing emotions, and hence more "dominant" (consistent with the notion that certain emotions are easier to recognize in specific communication channels, e.g., Hess, Kappas, & Scherer, 1988). For instance, Grimshaw (1998) reported stronger semantic interference effects on prosody than vice versa using an emotional Stroop test; similarly, Schirmer & Kotz (2003) reported that the influence of prosody on semantic processing was weaker than the influence of semantics on prosodic processing after presenting stimuli which were emotionally congruent or incongruent between the two channels. The relationship between semantics and prosody (or other contextual cues) could also be informed by cultural conventions (e.g., Kitayama & Ishii, 2002). Certainly, more studies will be needed to properly assess how emotional cues conveyed by prosody and semantic information are processed, and how these influence gaze behavior at different points of time during spoken interactions.

5. Conclusion

Taken together, results of the current investigation add to a growing body of work that suggests that emotional prosody processing is a highly rapid and possibly involuntary process. The eye tracking technique was successfully employed to measure implicit processing of emotionally-intoned sentences during a visual search task. Results imply that emotional prosodic cues are used to guide and speed visual search behavior, although the task did not require participants to use these cues, and visual attention was systematically influenced by the emotional meanings activated by

prosodic cues in an emotion-congruent manner. Our demonstration that eye tracking can be successfully employed to index facets of emotional prosody processing opens up new possibilities to investigate the implicit, "real time" use of prosodic information by listeners in tasks which more closely resemble natural human interactions.

6. References

Allopenna, P. D, Magnuson, J.S. & Tanenhaus, M.K. (1998). Tracking the time course of spoken word recognition: evidence for continuous mapping models. *Journal of Memory and Language*, 38, 419-439.

Ashley V, Vuilleumier P, Swick D. (2004). Time-course and specificity of event-related potentials to emotional expressions. *Neuroreport*, 15, 211–215.

Banse, R., & Scherer, K. R. (1996). *Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology*, 70(3), 614–636.

Batty M, Taylor M.J. (2003). Early processing of the six basic facial emotional expressions. *Brain Research: Cognitive Brain Research*, 17, 613–620.

Berman, J.M.J, Chambers, C.G., Graham, S.A. (2010). Preschooler's appreciation of speaker vocal affect as a cue to referential intent. *Journal of Experimental Child Psychology*, 107 (2), 87-99.

Besson, M., Magne, C., Schön, D. (2002). Emotional prosody: sex differences in sensitivity to speech melody. *Trends in Cognitive Science,* 6, 405–407.

Borod, J.C., Pick, L.H., Hall, S., Sliwinski, M., Madigan, N., Obler, L.K., Welkowitz, J., Canino, E., Erhan, H.M., Goral, M., Morrison, C., Tabert, M. (2000). Relationships among Facial, Prosodic, and Lexical Channels of Emotional Perceptual Processing. *Cognition and Emotion*, 14, 193-211.

Bostanov, V., Kotchoubey, B. (2004). Recognition of affective prosody: continuous wavelet measures of event-related brain potentials to emotional exclamations. *Psychophysiology,* 41, 259-268.

Bower, G.H. (1981). Mood and memory. *American Psychologist*, 36, 129–148.

Bowers, D., Bower, R., Heilman, K. (1993). The nonverbal affect lexicon: Theoretical perspectives from neuropsychological studies of affect perception. *Neuropsychology*, 7, 433-444.

Bowers, D., Coslette, H.B., Bauer, R., Speedie, L., & Heilman, K.M. (1987). Comprehension of emotional prosody following unilateral hemispheric lesions: processing defect versus distraction defect. *Neuropsychologia,* 25, 317-328.

Brosch, T., Grandjean, D., Sander, D., Scherer, K.R. (2008). Cross-modal emotional attention: Emotional voices modulate early stages of visual processing. *Journal of Cognitive Neuroscience*, 21, 1670–1679.

Calvo, M.G. & Nummenmaa, L. (2008). Detection of emotional faces: salient physical features guide effective visual search. *Journal of Experimental Psychology: General*, 137,471-494.

Calvo, M.G., Nummenmaa, L., & Avero, P. (2008). Visual search of emotional faces: Eye-movement assessment of component processes. *Experimental Psychology*, 55, 359-370.

Carroll, J.M. & Russell, J.A. (1996). Do facial expressions signal specific emotions? Judging the face in context. *Journal of Personality and Social Psychology*, 70, 205 – 218.

Carroll, N.C., & Young, A.W. (2005). Priming of emotion recognition. *The Quarterly Journal of Experimental Psychology*, 58A, 1173-1197.

Dahan, D. & Tanenhaus, M.K. (2005). Looking at the rope when looking for the snake: conceptually mediated eye movements during spoken word recognition. *Psychonomic Bulletin & Review*, 12, 453 – 459.

DeGelder, B. & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, 14, 289-311.

Eastwood, J.D., Smilek, D. & Merikle, P.M. (2001). Differential attentional guidance by unattended faces expressing positive and negative emotion. *Perception & Psychophysics*, 63, 1004 – 1013.

Eimer M, Holmes A. (2002). An ERP study on the time course of emotional face processing. *Neuroreport*, 13, 427–431.

Eimer M., Holmes A., McGlone F.P. (2003). The role of spatial attention in the processing of facial expression: an ERP study of rapid brain responses to six basic emotions. *Cognitive, Affective, and Behavioral Neuroscience*, 3, 97–110.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pancultural elements in facial displays of emotion. *Science,* 164(3875), 86-88.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion*, 15, 115-141.

Frischen, A. Eastwood, J.D. Smilek, D. (2008). Visual search for faces with emotional expressions. *Psychological Bulletin*, 134, 662–676.

Grimshaw, G.M. (1998). Integration and interference in the cerebral hemispheres: Relations with hemispheric specialization, *Brain and Cognition,* 36, 108–127.

Hansen, C.H. & Hansen, R.D. (1988). Finding the face in the crowd: An anger superiority effect. *Journal of Personality & Social Psychology*, 54, 917–924.

Hansen, C.H. & Shantz, C.A. (1995). Emotion-specific priming: Congruence effects on affect and recognition across negative emotions. *Journal of Personality and Social Psychology*, 21, 548–557.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson and F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Hess, U. Kappas, A., & Scherer, K.R. (1988). Multichannel communication of emotion: Synthetic signal production. Hillsdale: NJ: Erlbaum.

Hietanen, J.K., Leppänen, J.M., Illi, M., & Surakka, V. (2004). Evidence for the integrationof audiovisual emotional information at the perceptual level of processing. *European Journal of Cognitive Psychology*, 16, 769–790.

Horstmann, G. (2007). Preattentive face processing: What do we learn from visual search experiments? *Visual Cognition*, 15, 799–833.

Innes-Ker, A., Niedenthal, P. (2002). Emotion concepts and emotional states in social judgment and categorization. *Journal of Personality and Social Psychology*, 83, 804–816.

Isaacowitz, D.M. Toner, K., Goren, D., & Wilson, H.R. (2008). Looking while unhappy: Mood congruent gaze in young adults, positive gaze in older adults. *Psychological Science,* 19, 843-853.

Ito, K. & Speer, S. R. (2008). Anticipatory effect of intonation: Eye movements during instructed visual search. *Journal of Memory and Language,* (58) 541-573.

Kissler J., Keil A. (2008). Look—don't look! How emotional pictures affect pro- and anti-saccades. *Experimental Brain Research*, 188, 215-222

Kitayama, S. & Ishii, K. (2002). Word and voice: Spontaneous attention to emotional utterances in two languages. *Cognition and Emotion*, 16, 29–59 .

Koelsch, S. & Siebel, W. (2005). Towards a neural basis of music perception, *Trends Cogn. Sci.,* 9, pp. 578–584.

Kotz, S. A., & Paulmann, S. (2007). When emotional prosody and semantics dance cheek to cheek: ERP evidence. *Brain Research*, 1151, 107–118.

Massaro, D.W. & Egan, P.B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin Review*, 3, 215 – 221.

Matin, E., Shao, K.C., Boff, K.R., (1993). Saccadic overhead. Information-processing time with and without saccades. *Perception & Psychophysics,* 53, 372–380.

McManus, Susan M., (2009). Gaze Fixation During the Perception of Visual and Auditory Affective Cues. *Psychology Theses.* Paper 70. http://digitalarchive.gsu.edu/psych_theses/70

Niedenthal, P. M. & Halberstadt, J.B. (1995). The Acquisition and structure of emotional response categories. *The Psychology of Learning and Motivation*, 33, 23–63.

Nummenmaa, L., Hyönä, J., & Calvo, M.G. (2009). Emotional scene content drives the saccade generation system reflexively. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 305-323.

Nygaard, L. C., & Lunders, E.R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition*, 30, 583-593.

Öhman, A. Flykt and F. Esteves (2001). Emotion drives attention: Detecting the snake in the grass, *Journal of Experimental Psychology: General,* 130, 466–478.

Paulmann, S. & Kotz, S.A. (2006). Valence, arousal, and task effects on the P200 in emotional prosody processing. *In Proceedings of Architectures and Mechanisms for Language Processing 2006* (AMLaP 2006), Nijmegen, The Netherlands, p. 37.

Paulmann, S. & Kotz, S.A. (2008a). Early emotional prosody perception based on different speaker voices. *NeuroReport,* 19(2), 209-213.

Paulmann, S. & Kotz, S.A. (2008b). An ERP Investigation on the Temporal Dynamics of Emotional Prosody and Emotional Semantics in Pseudo- and Lexical Sentence Context. *Brain and Language*, 105, 59-69.

Paulmann, S. & Pell, M.D. (2009). Decoding emotional faces depends on their representational value: ERP evidence. *Neuroreport*, 20, 1603-1608.

Paulmann, S. & Pell, M.D. (2010). Contextual influences of speech tone on emotional processing: how much is enough? *Cognitive, Affective, and Behavioral Neuroscience,* 10, 230-242.

Paulmann, S. & Pell, M.D. (2011). Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion*, 35 (2), 192-201.

Paulmann, S., Jessen S., & Kotz, S.A. (2009). Investigating the multi-modal nature of human communication: Insights from ERPs. *Journal of Psychophysiology*, 23, 63 – 76.

Paulmann, S., Pell, M.D., Kotz, S.A. (2008). How aging affects the recognition of emotional speech. *Brain and Language*, 104, 262-269.

Pell, M. D. (2005b). Prosody face interactions in emotional processing as revealed by the facial affect decision task. *Journal of Nonverbal Behavior*, 29 (4), 193-215.

Pell, M. D. & Skorup, V. (2008). Implicit processing of emotional prosody in a foreign versus native language. *Speech Communication*, 50 (6), 519-530.

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009a). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, 33(2), 107–120.

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S.A. (2009b). Factors in the recognition of vocally expressed emotions: a comparison of four languages. *Journal of Phonetics*, 37, 417-435.

Pell, M.D., Jaywant, A., Monetta, L., & Kotz, S.A. (In press). Emotional speech processing: disentangling the effects of prosody and semantic cues. *Cognition and Emotion*.

Pell, M.D. (2005a) Nonverbal emotion priming: Evidence from the 'Facial Affect Decision Task'. *Journal of Nonverbal Behavior*, 29 (1), 45-73.

Pell, M.D. (2006). Cerebral mechanisms for understanding emotional prosody in speech. *Brain and Language*, 96 (2), 221-234.

Pittam, J., & Scherer, K. R. (1993). Vocal expression and communication of emotion. In M. Lewis & J. M. Haviland (Eds.), Handbook of emotions (pp. 185-197). New York: Guilford Press.

Russell, J., Lemay, G., (2000). *Emotion Concepts.* In Lewis, M. & Haviland-Jones, M (Eds.), Handbook of Emotion, New York: Guilford Press.

Sauter D.A., Calder A.J., Eisner F., Scott S.K. (2010). Perceptual cues in non-verbal vocal expressions of emotion. *The Quarterly Journal of Experimental Psychology*, 63 (11), 2251-2272.

Sauter, D., & Eimer, M. (2010). Rapid detection of emotion from human vocalizations. *Journal of Cognitive Neuroscience*, 22, 474-481

Scherer, K. R., Banse, R., & Wallbott, H. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, 32(1), 76–92.

Scherer, K.R. (1989). Vocal measurement of emotion. New York: Academic Press.

Schirmer, A. & Kotz, S.A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, 10, 24-30.

Schirmer, A., Kotz, S.A., & Friederici, A.D. (2002). Sex differentiates the role of emotional prosody during word processing. *Cognitive Brain Research*, 14, 228-233.

Schirmer, A., Kotz, S.A., 2003. ERP evidence for a sex-specific Stroop effect in emotional speech. *Journal of Cognitive Neuroscience*. 15, 1135–1148.

Schirmer, A., Striano, T. & Friederici, A.D. (2005). Sex differences in the preattentive processing of vocal emotional expressions. *NeuroReport*, 16, 635-639.

Schirmer,A., Kotz,S.A., & Friederici, A.D. (2005).On the role of attention for the processing of emotions in speech: Sex differences revisited. *Cognitive Brain Research*, 24, 442-452.

Spivey, M.J., Tyler, M.J., Eberhard, K.M. & Tanenhaus, M.K. (2001). Linguistically mediated visual search. *Psychological Science*, 12, 282-286.

Tanenhaus M.K., Spivey-Knowlton M.J., Eberhard K.M., Sedivy J.E. (l995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268,1632-1634.

Thompson, W. F., Russo, F. A., & Quinto, L. (2008). Audio-visual integration of emotional cues in song. *Cognition and Emotion,* 22*,* 1457-1470.

Vroomen, J., & de Gelder, B. (2000). Sound enhances visual perception: cross-modal effects of auditory organization on vision. *J Exp Psychol Hum Percept Perform,* 26 (5), 1583-1590.

Vroomen, J., Driver, J., & de Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources? *Cognitive, Affective, & Behavioral Neuroscience*, 1, 382-387.

Wambacq, I.J., Shea-Miller, K.J., Abubakr, A. (2004). Non-voluntary and voluntary processing of emotional prosody: An event-related potentials study. *Neuroreport*, 15, 555 -559.

Weber, A., Grice, M., & Crocker, M. W. (2006). The role of prosody in the interpretation of structural ambiguities: a study of anticipatory eye movements. *Cognition*, 99, B63-B72.

Wilson,D., Wharton,T. (2006). Relevance and prosody. *Journal of Pragmatics,* 36(1), 1559-1579.

Young, A.W., Rowland, D., Calder, A.J., Etcoff, N.L., Seth, A. and Perrett, D.I. (1997). Facial expression megamix: tests of dimensional and category accounts of emotion recognition. *Cognition*, 63, 271-313.

Zeelenberg, R. & Bocanegra, B.R. (2010). Auditory emotional cues enhance visual perception. *Cognition,* 115, 202-206.

7. Acknowledgments

The authors wish to thank Matthieu Couturier for help with programming the experiment, Abhishek Jaywant for help with data acquisition, Stephen Hopkins, Moritz Dannhauer, and Cord Plasse for help with data analysis, and Catherine Knowles for help with tables and figures. This work was supported by a new initiative fund awarded to the authors by the Center for Research on Language, Mind and Brain (CRLMB). Support received from the German Academic Exchange Service (DAAD) to the first author and McGill University (William Dawson Scholar award) to the third author is gratefully acknowledged.

Figure Captions

Figure 1a: This graph shows the mean duration (measured in milliseconds) for each word spoken in different prosodies. xxx stands for the adjective name (angry, happy, frightened, sad, neutral).

Figure 1b: This graph shows the mean fundamental frequency (measured in Hz) for each word spoken in different prosodies. xxx stands for the adjective name (angry, happy, frightened, sad, neutral).

Figure 2: Displayed are example waveforms and spectrograms for stimuli used. Examples of each emotional category tested for two different speakers are displayed. Spectrograms show visible pitch contours and were created with Praat (Boersma & Weenink, 2009).

Figure 3: An illustration of the circular facial expression array setup in the experiment and examples of emotional facial expressions posed by one actor.

Figure 4a&b: Mean eye fixation durations (a) and frequency of looks (b) in the pre-emotional label time window to faces that matched versus mismatched the emotional prosody of the speaker. Error bars refer to standard deviations.

Figure 5a&b: Mean eye fixation durations (a) and frequency of looks to faces (b) that matched the emotion of the semantic instruction (SEM), the speaker's prosody (PROS), and both semantics and prosody (SEM+PROS) when compared to fixations which matched neither the semantics nor the prosody (NONE). Figure 4 (a & b) show the results for the pre-emotional label time window, whereas Figure 5 (a & b) illustrates results for the short post-emotional label time. Error bars refer to standard deviations.

Abstract

This study investigated cross-modal effects of emotional voice tone (*prosody*) on face processing during instructed visual search. Specifically, we evaluated whether emotional prosodic cues in speech have a rapid, mandatory influence on eye movements to an emotionally-related face, and whether these effects persist as semantic information unfolds. Participants viewed an array of six emotional faces while listening to instructions spoken in an emotionally congruent or incongruent prosody (e.g., "Click on the happy face" spoken in a happy or angry voice). The duration and frequency of eye fixations were analyzed when only prosodic cues were emotionally meaningful (pre-emotional label window: "Click on the/…"), and after emotional semantic information was available (post-emotional label window: "…/happy face"). In the pre-emotional label window, results showed that participants made immediate use of emotional prosody, as reflected in significantly longer and more frequent fixations to emotionally congruent versus incongruent faces. However, when explicit semantic information in the instructions became available (post-emotional label window), the influence of prosody on measures of eye gaze was relatively minimal. Our data show that emotional prosody has a rapid impact on gaze behavior during social information processing, but that prosodic meanings can be overridden by semantic cues when linguistic information is task relevant.

(200 words)

Keywords: eye-tracking, gaze, speech processing, affective prosody, semantics

Research Highlights

- We explore if emotional prosody has a rapid influence on eye movements
- Results of fixation duration show rapid impact of emotional prosody on gaze behavior
- Results of fixation frequency show similar rapid, apparently mandatory impact
- Influence of prosody can be overridden by semantic cues if latter are task relevant

Table 1.

Major acoustic features of the emotional utterances presented in the experiment, measured in the pre-semantic and post-semantic time windows.

	Pre-Semantic Time Window					Post-Semantic Time Window				
Prosody	Pitch Mean (Hz)	Pitch Range (Hz)	Duration (ms)	Intensity Mean (dB)	Intensity Range (dB)	Pitch Mean (Hz)	Pitch Range (Hz)	Duration (ms)	Intensity Mean (dB)	Intensity Range (dB)
Angry	206	81	620	56	41	167	170	960	53	33
Frightened	340	80	430	60	31	258	199	850	59	31
Happy	246	198	560	58	34	202	238	910	56	29
Sad	174	92	580	54	31	156	190	960	52	32
Neutral	153	90	580	52	37	162	215	900	52	30

Note: The pre-semantic time window was measured from speech onset to emotional adjective onset, whereas the post-semantic time window was measured from the emotional adjective onset to speech offset.

Table 2.

Mean F0 levels at the beginning of the utterance. The table shows mean F0 values (measured in Hz) for the first 450 ms after sentence onset divided into 50 ms bins.

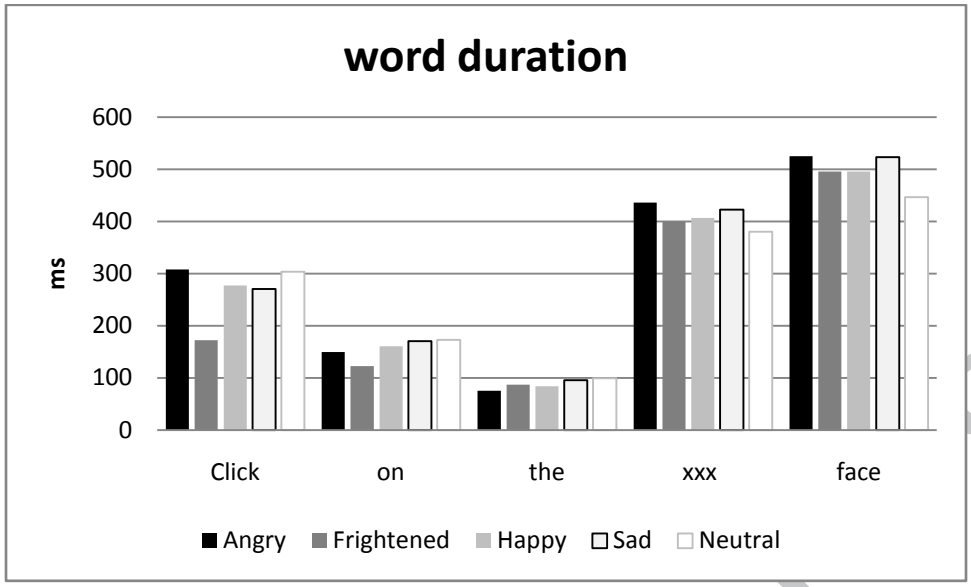
Prosody	0-50ms	50-100ms	100-150ms	150-200ms	200-250ms	250-300ms	300-350ms	350-400ms	400-450ms
Angry	570	253	238	190	183	198	222	222	212
Frightened	334	357	358	358	343	332	327	288	297
Happy	384	288	318	295	270	227	193	201	200
Sad	552	202	181	183	169	167	152	145	156
Neutral	461	189	182	189	127	124	109	103	146

Table 3.

Mean fixation duration and frequency to faces that matched or mismatched the speaker's prosody (pre-semantic time window), or which matched specific combinations of prosody and semantic cues (post-semantic time window), according to the emotion of the prosody.

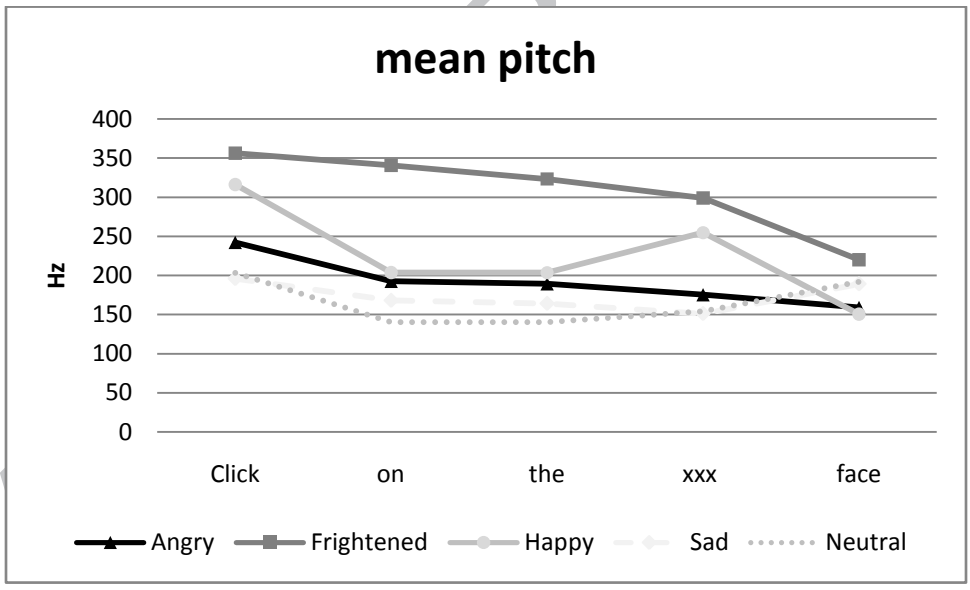
		Pre-Semantic Time Window		Post-Semantic Time Window			
Gaze Measure	Prosody	Prosody Match	Prosody Mismatch	Pros+Sem Match	Semantics Match	Prosody Match	None
Duration (ms)	Angry	276	261	303	296	245	230
	Frightened	286	250	299	298	214	216
	Happy	268	259	332	288	232	224
	Sad	277	266	291	295	238	233
Frequency (# of looks)	Angry	1.02	1.02	1.15	1.10	1.08	1.05
	Frightened	1.01	1.01	1.13	1.13	1.04	1.05
	Happy	1.02	1.02	1.12	1.16	1.05	1.04
	Sad	1.02	1.02	1.16	1.14	1.04	1.04

Figure 1a:

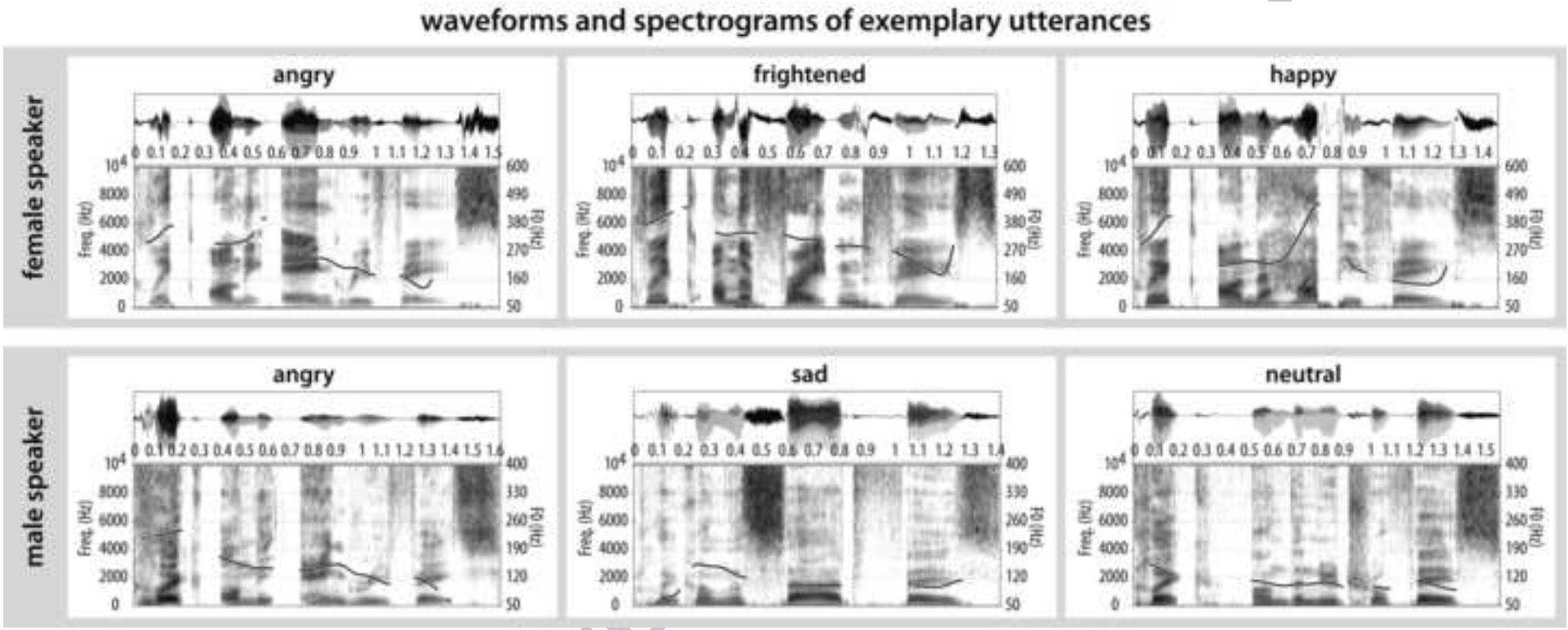


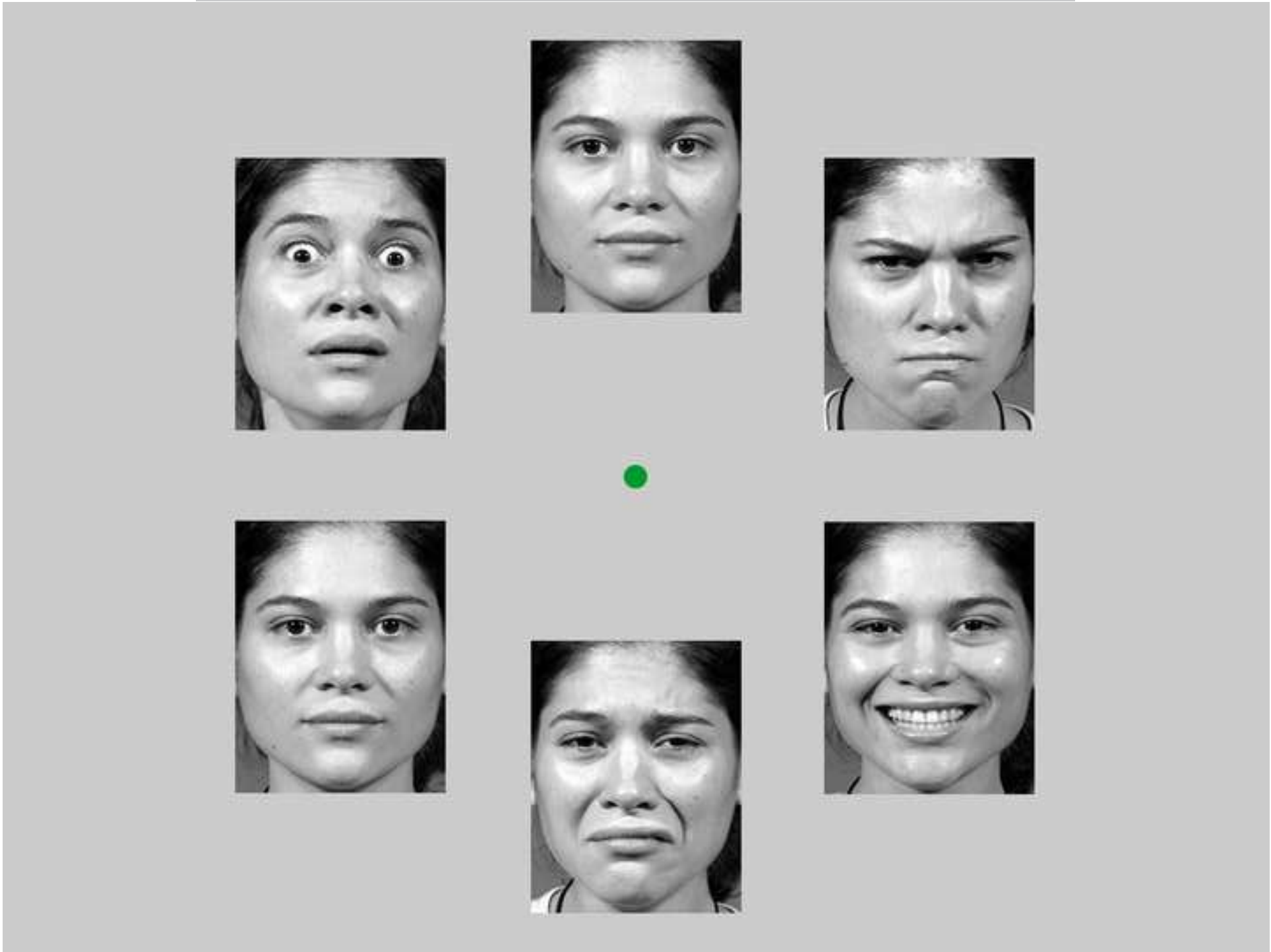
This graph shows the mean duration (measured in milliseconds) for each word spoken in different prosodies. xxx stands for the adjective name (angry, happy, frightened, sad, neutral).

Figure 1b:



This graph shows the mean fundamental frequency (measured in Hz) for each word spoken in different prosodies. xxx stands for the adjective name (angry, happy, frightened, sad, neutral).





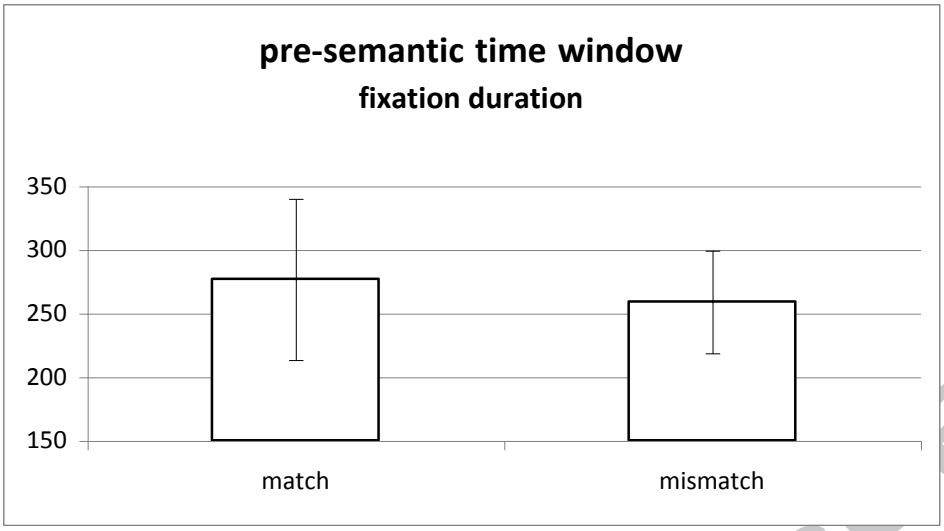


Figure 4a:

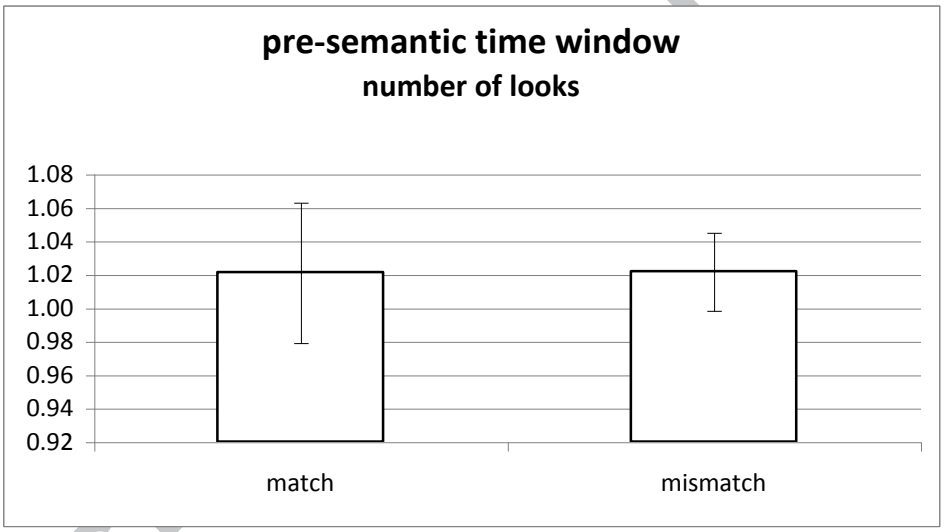


Figure 4b:

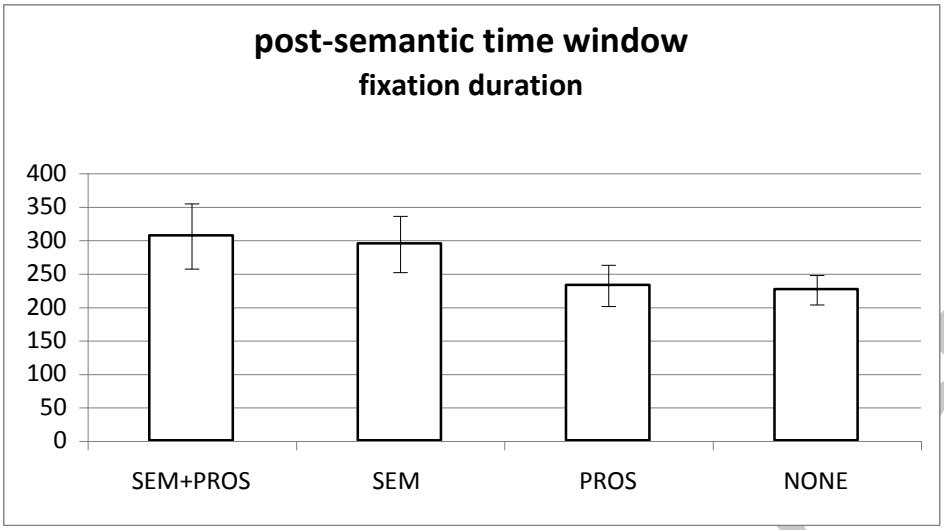


Figure 5a

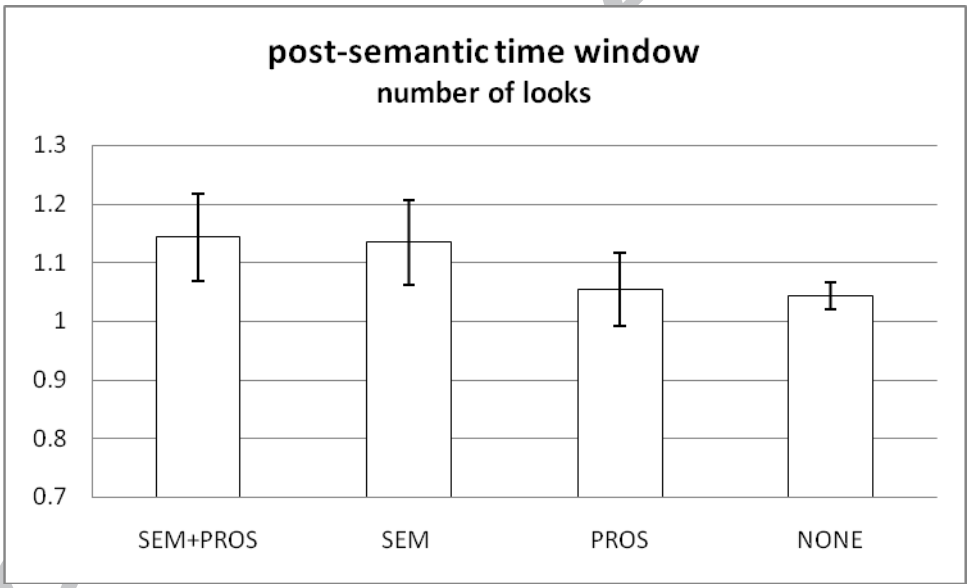


Figure 5b