The Sound of the Crowd: Auditory Information Modulates the Perceived Emotion of a Crowd Based on Bodily Expressions

Article *in* Emotion · August 2011

DOI: 10.1037/a0024785 · Source: PubMed

4 authors:

Joanna Power
National University of Ireland, Maynooth

Gavin Kearney
The University of York

Henry Rice
Trinity College Dublin

Fiona N Newell
Trinity College Dublin

Emotion
2012, Vol. 12, No. 1, 120–131

1528-3542/11/$12.00 DOI: 10.1037/a0024785

The Sound of the Crowd: Auditory Information Modulates the Perceived Emotion of a Crowd Based on Bodily Expressions

Joanna E. McHugh, Gavin Kearney, Henry Rice, and Fiona N. Newell
Trinity College Dublin

Although both auditory and visual information can influence the perceived emotion of an individual, how these modalities contribute to the perceived emotion of a crowd of characters was hitherto unknown. Here, we manipulated the ambiguity of the emotion of either a visual or auditory crowd of characters by varying the proportions of characters expressing one of two emotional states. Using an intersensory bias paradigm, unambiguous emotional information from an unattended modality was presented while participants determined the emotion of a crowd in an attended, but different, modality. We found that emotional information in an unattended modality can disambiguate the perceived emotion of a crowd. Moreover, the size of the crowd had little effect on these crossmodal influences. The role of audiovisual information appears to be similar in perceiving emotion from individuals or crowds. Our findings provide novel insights into the role of multisensory influences on the perception of social information from crowds of individuals.

Keywords: emotion, multisensory integration, crowds, perception, motion perception

Social perception involves interpreting the actions and body postures of others in order to inform us of their attitudes and intentions (Allison, Puce, & McCarthy, 2000). One class of social cue is emotional information. Typically, emotional information is communicated between individuals through the auditory (i.e., voice) and visual (facial or bodily expressions) modalities, representing a naturally bimodal event (Vroomen, Driver, & de Gelder, 2001). As such, specialized mechanisms may underlie the crossmodal binding of this information (de Gelder & Bertelson, 2003) for a more robust perception of emotions expressed by others (Dolan, Morris, & de Gelder, 2001).

Evidence for the effects of audiovisual integration on the perception of emotion has been provided in studies using intersensory bias paradigms (de Gelder & Vroomen, 2000a; Massaro & Egan, 1996) whereby emotional information is presented both to the auditory and visual modalities, and participants are instructed to ignore one modality. Despite these instructions, information from the unattended modality nevertheless affects the recognition of emotion in the attended modality. To date most research in this area has focused on facial and vocal expressions. However, expressions from the body are also important for perceiving the emotional state of others (de Gelder, 2009) and such information, when paired with vocal expressions can provide a reliable multisensory cue for determining emotion. For example, van den Stock and colleagues (van den Stock, Righart, & de Gelder, 2007; van den Stock, Grezes, & de Gelder, 2008) recently used the intersensory bias paradigm to investigate the integration of vocal and emotional body language information, and found evidence in support of crossmodal influences in the recognition of emotion in an individual.

According to Bayesian approaches to multisensory perception, a perceptual outcome is related to the relative reliability of the information in each sensory signal (see, e.g., Ernst & Banks, 2002). For example, during audiovisual stimulation, if the auditory information is relatively noisy and unreliable, visual information will dominate the percept (Alais & Burr, 2004). According to such an approach, the perceptual system weighs the relative reliability of the unisensory inputs before optimally integrating these inputs to achieve a robust multisensory percept (see, e.g., Ernst & Bülthoff, 2004). Previous studies have assessed the role of multisensory inputs on the perceived emotion by manipulating the reliability of the emotional information in a visual image using morphing procedures (de Gelder & Vroomen, 2000a; Young et al., 1997; de Gelder, Teunisse, & Benson, 1997). Using morphed images, De Gelder and Vroomen (2000a) reported that auditory information dominated the perceived emotion more as the ambiguity of the visual information increased. Their finding is therefore consistent with the general predictions of Bayesian approaches to multisensory integration (although such models have yet to be directly applied to cognitive tasks such as emotional recognition). One model which has been applied specifically to the perception of

This article was published Online First August 29, 2011.

Joanna E. McHugh and Fiona N. Newell, School of Psychology and Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland; Gavin Kearney and Henry Rice, Department of Mechanical and Manufacturing Engineering; Trinity College Dublin.

This research was funded by Science Foundation Ireland, Principal Investigator grant (project No. 06/IN.1/I96) awarded to Carol O'Sullivan, Fiona N. Newell, and Henry Rice. We thank Carol O'Sullivan and Rachel McDonnell from the Graphics, Vision, and Visualization Group at Trinity College Dublin, for their aid in preparing the visual stimuli used in the reported studies.

Gavin Kearney is now at Department of Theater, Film and TV, University of York.

Correspondence concerning this article should be addressed to Fiona N. Newell, School of Psychology and Institute of Neuroscience, Lloyd Building, Trinity College, Dublin 2, Ireland. E-mail: fiona.newell@tcd.ie

emotion is the Fuzzy Logical Model of Perception (FLMP, Massaro & Egan, 1996). According to this approach, unisensory inputs are also independently analyzed and integrated according to their reliability, however, the FLMP predicts that emotional information can affect the perceived emotion regardless of its nature, such that, for example, emotional expressions from text could influence the perceived emotion from vocal information (de Gelder & Vroomen, 2000b). This prediction has been supported by recent findings that emotional information from animal vocalizations (van den Stock et al., 2008) or from instrumental music (van den Stock, Peretz, Grézes, & de Gelder, 2009) can be integrated with information from human body expressions to affect the perceived human emotion. However, these authors suggest that there likely remains a privileged link between human vocal and human body language information for the purpose of multisensory integration in the perception of emotion.

While it seems clear that both the visual and auditory senses can provide information relevant to the perception of emotion in a single human character, it remains less clear how and if these sensory signals are combined for the perception of emotion in crowds of people. It is possible that the integration of information across modalities for the purpose of perceiving emotion is independent of the number of characters expressing that emotion. In other words, if a large number of individuals are expressing the same emotion, then the integration of multisensory information for the purpose of the perception of emotion may occur in a similar manner to that in individuals. Recent studies reported by Haberman and Whitney (2007, 2009, 2010) suggest a mechanism which may allow for the emotion of a crowd to be perceived in a similar manner to that of an individual, at least within the visual modality. They reported that when viewing a number of faces the visual system rapidly extracts a summary representation, or average, of the emotional expression of these faces and that this likely occurs in the absence of intentional or attentional processing. By way of further validation Haberman and Whitney (2010) reported evidence that face "outliers" (i.e., faces with expressions which were deviant from the general crowd of faces) were disregarded in the representation of the mean emotion derived from the crowd of faces. This implies that when perceiving the emotion of a crowd, serial processing of individual faces may not be required but rather there appears to be a mechanism which allows for the facial expressions of individuals in a crowd to be represented as holistic. However, other research has shown that for some facial expressions, particularly anger or happy, outliers can be readily detected in a crowd of faces (Pinkham, Griffin, Baron, Sasson, & Gur, 2010; Juth, Lundqvist, Karlsson, & Ohman, 2005). It may be the case, therefore, that outliers can affect the perceived emotion of a crowd dependent on the saliency of that emotional information relative to other characters.

McHugh, McDonnell, O'Sullivan, & Newell (2010) investigated the perceived emotion of a crowd based on visual body expressions of the characters. They found that although overall crowd emotion could be readily categorized, performance depended on the consistency with which an emotion was expressed across all individuals in the crowd. Specifically, categorizing the emotion of a crowd was more efficient when greater proportions of characters expressed one emotion over another. Similar to the morphing of faces used in previous studies investigating facial expressions, visual displays of crowds in which different proportions of characters expressing one of two emotions may represent an emotionally ambiguous percept. As such, emotion in these "mixed" crowds was more difficult to categorize than when all characters consistently expressed one emotion only. The findings of McHugh and colleagues (2010) are compatible with those reported by Haberman and Whitney using faces, in that the majority of items present in the image were encompassed into the overall representation of the emotion of the crowd. As such, the representation of the average emotion of a set of bodies is likely to be more difficult and less reliable as the number of "outliers" increases in the set. Moreover, McHugh et al. reported important interactions based on the nature of the emotional expression, suggesting that visual saliency may also affect the perceived emotion of a crowd.

What was hitherto unknown, however, is whether or not the process involved in representing the average emotion of a set of individuals is affected by emotional information from another modality. The aim of the following experiments was to investigate whether reliable crossmodal information can affect the perception of the emotion of a crowd in another modality. Specifically, reliable auditory information may disambiguate visual information, and vice versa, resulting in a more robust perception of the emotion of a crowd. In Experiment 1a ambiguity in the visual modality was manipulated by varying the proportions of characters in a crowd expressing one of two emotions. These visual images of crowds were presented with the sound of a crowd of characters expressing the same, consistent emotion. In Experiment 1b, ambiguity was manipulated within the auditory modality but all the visual characters in the crowd expressed the same emotion. Based on previous evidence for crossmodal influences in the perception of emotion, we hypothesized that the categorization of the emotion of a crowd in one modality would depend of the nature of the emotional information from the unattended modality, such that if participants were told to ignore the emotional crowd sounds and categorize visual emotional crowds, their responses would be based on the information from both modalities.

In Experiment 2, we investigated a potential effect of crowd size (the number of visual characters or voices in the crowd) on the integration of emotional information across the auditory and visual modalities. Experiment 2 allowed us to assess whether the integration of the auditory and visual information occurs in a serial manner or whether it occurs in parallel as suggested by Haberman and Whitney (2007). Moreover, if responses to the audiovisual information do not differ as a function of crowd size then this could be seen as putative support for Massaro and Egan's (1996) theory that emotion integration occurs regardless of the medium through which the information is delivered.

Experiment 1

The aim of Experiment 1 was to investigate whether emotional information presented in an unattended modality could influence the recognition of emotion in the attended modality. The visual crowds were comprised of dynamic human characters each displaying a particular emotional state using emotional body language only. Auditory crowds were comprised of sounds of groups of individuals vocalizing in a particular emotional state. In Experiment 1a, participants were instructed to ignore the sounds and categorize the emotion of the visual crowd, while in Experiment

1b, participants were told to ignore the images and categorize the emotion of the sound of the crowd. For each experiment, we manipulated the proportion of characters exhibiting one of two emotional states in the attended modality only; the crowds in the unattended modality always represented an unambiguous emotional stimulus. We tested four emotional categories, namely fear, sad, happy and angry. When crowd proportion was manipulated, we used crowds in which 0%, 25%, 50%, 75%, or 100% of the characters depicted one of two emotional categories. The emotion in the unattended modality was always one of the two emotions depicted by the characters in the crowd of the attended modality. We also included a unimodal condition in each of the Experiments 1a and 1b, in which there was either no auditory or no visual stimulus presented respectively. The results from this unimodal condition served as a baseline comparison.

Based on the findings of McHugh et al. (2010), we predicted that the emotional categorization of a visual crowd would be dependent on the proportion of characters in the crowd depicting that emotion. In Experiment 1a we expected that the sound of an emotional crowd would have an effect on the categorization of emotion based on the visual crowds, and for Experiment 1b we expected that the visual emotional crowd would affect the categorization of emotion from the sound of the crowds. Moreover, we expected the categorization of the attended crowds to be biased in the direction of the unattended emotional category. This effect of emotional integration was expected to be strongest when the emotional category of the attended modality was ambiguous (i.e., when 50% of the crowd depicted one of two emotions).

Method

Participants

We recruited 18 participants (15 female, with an age range of 19 to 34 years, and a mean age of 21.2 years) for Experiment 1a and a separate group of 20 participants (eight male, with an age range of 18 to 34 years and mean age of 23.4 years) for Experiment 1b, from the student population at Trinity College Dublin. For this and all further reported experiments, participants gave informed, written consent to partake, and received research credits for their participation. All reported normal or corrected to normal vision.

Stimuli and Apparatus

The visual stimuli for Experiment 1 consisted of crowds containing dynamic human characters expressing emotion via their body language (see McHugh et al., 2010, for full details of the creation of these stimuli). Each character was created by superimposing the motion captured from real actors onto the bodies of virtual characters to create a set of five different characters, each depicting four different versions of each emotional category. There were three female and two male virtual characters and all were depicted with the same clothing and with the faces of the characters masked using a white facial mask. In this way, the emotional state of the character could be determined based on body information only and not on facial expression.

Crowd scenes were then created from these virtual characters using Adobe Premiere 6.5 software. Each crowd scene consisted of a fixed number of 20 characters, with equal numbers of male and female characters in each scene. Each character was pseudorandomly positioned on each scene with the constraints that none were occluded, nor placed at the center of the crowd (where an intertrial fixation cross was positioned). To add to the variation across characters in a crowd scene, characters were oriented in one of five different directions: forward-facing; facing 30°, or facing 60°, toward either the left or right. The characters for each crowd scene were randomly positioned against a uniform gray background (RGB value: 120, 120, 120). The approximate visual angle subtended by each crowd scene was 16° from central fixation. The image of each character within the crowd subtended a visual angle of approximately 2.5° vertically and 1° horizontally. See Figure 1 for an example of the visual crowd stimuli used in Experiment 1.[1]

The four emotion categories investigated were fear, anger, sadness, and happiness. For each of the visual and auditory modalities, emotional crowd continua were created between six possible pairings of these four emotions (following a design used by Young et al., 1997). These pairings were therefore as follows: anger-fear, happy-anger, anger-sad, fear-happy, sad-fear, and sad-happy. For convenience and the purpose of later data analyses, we arbitrarily named the first-named emotion in each continuum as emotion "A" and the second emotional category in each pair as emotion "B." Within each pairing different crowd stimuli were created based on a continuum of proportions of characters depicting one of two emotions assigned to each continuum pair. There were five steps along each continuum with 0%, 25%, 50%, 75%, or 100% of the characters consistently expressing one of the two emotions. Thus, there were four 100% emotional crowd stimuli where all characters consistently expressed one of the four emotion categories. For the other proportions, for example the 75% proportion, 75% of the characters expressed emotion "A" whereas 25% expressed the other emotional category in that pair. Other crowd stimuli comprised of 50% of the characters depicting emotion "A" and 50% depicting emotion "B," or 25% of the characters depicting emotion "A" and 75% emotion "B." These manipulations of the proportion of characters in crowds resulted in four 100% consistent crowds and 18 mixed emotion crowds, giving a total number of different 22 visual crowds.

In Experiment 1a, the auditory stimuli comprised of recordings of 18 actors (11 female, seven male, age range 19–34, mean age 24.5) expressing each of the four different emotional categories vocally. The recording took place in a lecture theater at Trinity College Dublin. Two condenser microphones placed on microphone stands at the left and right hand side of the theater and connected via a stereo preamplifier to a laptop were used to record the sounds. Stimuli were recorded and edited using Adobe Soundbooth software. These actors were instructed to vocalize each emotional state without using words, as the semantics of words may influence the emotional content of the vocalizations. They were cued to begin and stop vocalizing by the researcher. Ten recordings were made of each of angry, fearful, sad, and happy crowd noises, and the four best recordings were subsequently used. Sound stimuli were 2 s in duration in order to match the duration of the visual crowd stimuli.

[1] Examples of the audiovisual dynamic stimuli used in these experiments can be found here: www.tcd.ie/neuroscience/multisensory/demos.html

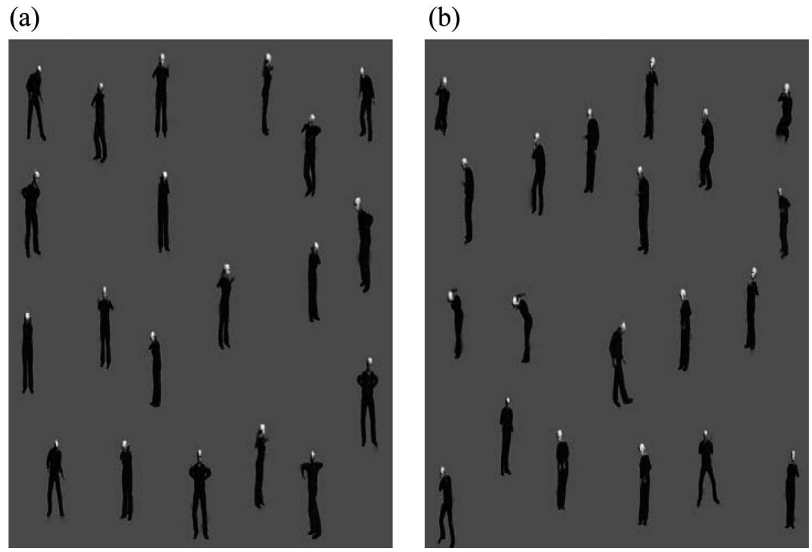

Figure 1. Examples of static images of the crowd stimuli used in Experiments 1 & 2. (a) This example depicts a 100% consistent crowd of 20 characters conveying the emotional category "anger." (b) This example depicts a crowd of 20 characters of whom 50% are conveying the emotional category "fear" and the other 50% are conveying the emotion "sadness."

For the purposes of Experiment 1b, it was necessary to rerecord the auditory crowd stimuli as different proportions of actors vocalizing one of two emotions were required within the same crowd stimulus. This proved impossible with a "live" crowd as the actors found it difficult to express one emotional state vocally while hearing another emotion being expressed by the other actors. This reported difficulty may have been due to the emotional contagion effect (Hatfield, Cacioppo, & Rapson, 1994), or the tendency to mirror the emotions displayed by others, which Hatfield and colleagues noted is particularly prevalent in crowds. For these reasons, it was decided to record smaller groups of actors, all expressing the same emotion, and manipulate these recordings create the sound of a larger crowd of actors using post-production editing techniques.

Auditory crowds were created from the vocal recordings made from a total of five different actors (three female; age range 20–34, mean age 25.2). These actors were recorded in groups of five, four, and three in order to create enough variation to allow us to build auditory crowds consisting of different proportions of voices from one of two emotional states (Experiment 1b) or to build auditory crowds of different sizes of 12, 20, and 32 characters (Experiment 2). Recording took place in a purpose built recording studio in the Department of Electronic and Electrical Engineering in Trinity College Dublin. The recording apparatus consisted of five cardioid microphones, one positioned directly in front of each actor, and a reference omnidirectional microphone centralized about 1.5m in front of the actor group. The actors were instructed to vocalize each of the four emotional states without verbalizing. The group produced 10 different sound examples of each of the four emotional states. Two different segments, each lasting 2 s in duration, were cut from each recording to produce 20 possible auditory samples of each emotion category. After the recordings, the voices of these actors were mixed such that each resultant auditory crowd was comprised of 20 voices (Experiment 1 only), with 50% male and 50% female voices, to match the makeup of the visual crowds.

Post-production comprised of taking the recordings taken from the individual microphones and level matching them based on LAeq (long term averaged A-weighted) levels using ProTools software (Benjamin, 2002). In order to create the auditory crowd stimuli containing mixed voices, multiple instances of each recording were created by taking a maximum of three random segments from each recording and applying pitch-shifting to each sample to create different pitch versions of the original recordings (without altering the speed of playback). Specifically, this variety to the voices was achieved by changing the length of the sound of each voice using time-expansion in ProTools (www.avid.com), and then performing sample rate conversion in Matlab (2008) to change the pitch. These new voice outputs were then panned evenly across a stereo field to ensure that, for the listener, no spatial masking (i.e., the impression that two voices emanate from the same point in space) would be perceived. For each emotional sound of a crowd, the recordings of the individual voices vocalizing one of two different emotions were interleaved across the stereo field. The relative levels of these 'mixed' recordings were matched once again using LAeq measurements, so that one recording was not perceived as louder than another.

Both the visual and the auditory stimuli were presented for 2 s, with synchronized onsets and offsets. The combinations of the emotional expressions and the proportion of crowd members expressing one of two emotions, gave a total of 66 stimuli each for use in Experiments 1a and 1b. The experiment was programmed and data output collected using Presentation software (version 12.0, www.neurobs.com). Stimuli were presented on a standard 15" PC monitor positioned approximately 60 cm away from the participant. The auditory stimuli were delivered via Sennheiser HD 202 headphones. Playback level was subjectively set by the researcher to reflect small crowd sound levels at 1.5m, and sound levels remained consistent between participants in the experiments. Participants were required to wear the headphones at the start of the experiment. Testing took place in a windowless, purpose built testing laboratory in which external sounds were attenuated.

Design

In both Experiments 1a and 1b, the emotion of the crowd in the attended modality (vision or sound respectively) either matched the emotion in the unattended modality (e.g., in Experiment 1a both the visual and auditory crowd expressed emotion category “A”) or did not match the emotion in the unattended modality (e.g., visual emotional category “A” but sound of crowd expressing emotion “B”). As a comparison, we also included a unimodal condition in which no information was provided from the unattended modality.

Experiments 1a and 1b were based on a repeated measures design, with the six emotional category pairings, the proportion of characters depicting one of two emotions (which we referred to as “proportion”: 0%, 25%, 50%, 75%, and 100% of emotion “A”) and the nature of the match between the emotional categories of the attended (i.e., emotion “A”) and unattended modalities (emotion “A,” emotion “B,” or none) as factors. In order to analyze the categorical responses we recoded each of the four possible emotion response categories (i.e., “fear,” “anger,” “happy,” or “sad”) for each continuum as follows: within each continuum one of the emotional categories was arbitrarily labeled emotion “A,” and the other as emotion “B.” Although these emotional labels were arbitrarily assigned, we used the same emotion-to-label pairings across all experiments in this study to ensure consistency across experiments. We also recorded reaction times.

There were two phases to Experiments 1a and 1b, consisting of a training session followed by the main experiment. The same response keys were used for all participants, which were the keyboard buttons “z” (corresponding to anger), “x” (fearful), “>” (happy), and “?” (sad). The index and middle fingers of both hands were used to make a response. The main experiment was based on a four alternative forced choice paradigm, where participants were instructed to categorize the emotion of the visual crowd stimuli in Experiment 1a, or the auditory crowd stimuli in Experiment 1b, as one of the four category emotions (fear, anger, sad, or happy). Although both visual and sound information were presented in both experiments, participants were instructed to ignore the auditory emotion in Experiment 1a, and to ignore the visual emotion in Experiment 1b. Each of the 66 crowd stimuli was presented twice across two experimental blocks, once per block. The presentation order of the trials was random within each block. Participants took a self-timed break between blocks.

Procedure

The participant’s task was to categorize, as fast and accurately as possible, the overall emotion displayed by the crowd of characters in the attended modality (vision in Experiment 1a, and sound in Experiment 1b), while ignoring the emotional category of the unattended modality. During the training session, participants first learned to associate the response keys with each of the emotional categories. To achieve this we presented the name of an emotional category onscreen until a response was made, and participants were required to press the response key associated with that emotional category as fast and accurately as possible. This training block was repeated until participants achieved a criterion response accuracy of at least 90%. On average the block was repeated three times for this criterion to be attained across participants. Only two participants had to repeat this block four times to reach criterion performance.

Participants were then presented with the main experiment. In this phase a trial consisted of the following: a central white fixation cross was presented for 1500 ms on which participants were instructed to focus. For Experiment 1a, a target visual crowd stimulus was presented along with the emotional sound of the crowd. The participant’s task was to categorize the emotion of the visual crowd as angry, fearful, happy, or sad, while ignoring the accompanying auditory information. In Experiment 1b a target auditory stimulus was presented along with an image of an emotional crowd. Here the participant’s task was to categorize the auditory emotion as angry, fearful, happy, or sad, while ignoring the accompanying visual information. In both experiments, each audiovisual stimulus was presented for 2 s, followed by a blank screen. A response at any time triggered the offset of that trial and the onset of the subsequent trial. Across participants, the experiment took an average of 25 min to complete.

Results

Experiment 1a

Reaction times were analyzed for outliers (anything above or below three interquartile ranges of the mean), which were then removed leading to the removal of 2% of the reaction time data. The average reaction time was 1670 ms. First, the data were compared across each of the four emotional categories in the “no sound” condition only. Accuracy and reaction times were compared across emotions as shown in Table 1.

A repeated-measures analysis of variance (ANOVA) was performed on the accuracy data with emotion as a within-subjects factor and a significant effect was found $F(3, 51) = 2.82$, $p < .05$. Tukey post hoc analyses found that categorization accuracy was significantly higher to the sad visual crowds than to either the angry ($p < .05$) or fearful ($p < .05$) visual crowds, with other

Table 1
Mean Proportion of Accurate Responses and Mean Reaction Times Across Participants in Experiment 1a to the 100% Visual Crowd Stimuli in Each of the Four Emotion Categories, With No Sound Presented (Standard Deviations Shown in Parentheses)

Emotional category	Reaction times (ms)	Proportion of accurate responses
Anger	1762 (*SD* = 311)	0.81 (*SD* = 0.33)
Fear	1562 (*SD* = 358)	0.82 (*SD* = 0.16)
Happy	1497 (*SD* = 381)	0.91 (*SD* = 0.12)
Sad	1510 (*SD* = 262)	0.93 (*SD* = 0.127)

differences between the categories failing to reach significance (i.e., happy vs. anger, $p = .056$; happy vs. fear, $p = .09$).

A second repeated measures ANOVA was performed on the reaction times data across the emotion categories and a significant effect was again found $F(3, 51) = 4.3$, $p < .01$. Tukey post hoc analyses revealed that the responses to the angry visual crowds were significantly slower than those to either the happy ($p < .05$) or the sad ($p < .05$) visual crowds, with no other significant differences between the categories. The results of the two ANOVA reported here both confirm that responses were the most efficient in the "no sound" condition to the sad visual crowds, and the least efficient to the angry crowds (see Table 1).

All categorization responses were then recoded as either emotion "A" or emotion "B" based on an arbitrary labeling of each pair of emotions displayed in each crowd stimulus as either "A" or "B," to allow us to average the data across the different emotional pair continua.[2] For example, in the anger-fear continuum, "anger" responses were relabeled as "A" responses and "fear" responses were relabeled as "B" responses.

A 5 × 3 repeated-measures ANOVA was performed on the mean number of "A" emotional category responses across participants with the proportion of characters expressing emotion "A" in the visual crowd scene (0%, 25%, 50%, 75%, and 100%) and emotional category of the sound (sound "A," sound "B" or no sound) as factors. Significant main effects of crowd proportion $F(4, 68) = 149.2$, $p < .01$, and emotion of the sound $F(2, 34) = 10.3$, $p < .01$, were found. Tukey post hoc analyses of the main effect of proportion revealed significant increases in "A" responses with each increase in the proportion of characters in the crowd expressing emotion "A" (at levels of $p < .05$ between each steps on the continuum). Tukey post hoc analysis of the sound effect revealed that there were more category 'A' responses made to the sound "A" condition than either the "B" or the no sound conditions ($p < .01$) but there was no difference between the number of category "A" responses made to the "no sound" or sound "B" conditions.

It is interesting that we found a significant interaction between the factors of crowd proportion and the emotional sound $F(8, 136) = 2.15$, $p < .05$, as shown in Figure 2. A Tukey post hoc analysis of the interaction revealed more category "A" responses made to the sound "A" condition relative to the sound "B" condition at all levels ($p < .01$). Significantly more category "A" responses were also made to the sound "A" condition relative to the "no sound" condition at 0%, 25%, and 50% levels only, with no difference between the sound "A" and the "no sound" conditions at proportion 75% and 100% levels. Finally, there were significantly fewer category "A" responses made to the "sound B" condition than the "no sound" condition at the 50% crowd proportion only and the difference between these conditions at any of the other proportions failed to reach significance ($p > .05$). In other words, at the 50% proportion level there were significantly more "A" responses made to the sound "A" condition than the no sound condition ($p < .05$) and significantly fewer category "A" responses made to the sound "B" condition relative to the no sound condition ($p < .05$). Therefore, when the emotion of the crowd was visually ambiguous, that is, when 50% of the characters visually exhibited one of two emotional states, the emotional sound of the crowd significantly biased the perceived visual emotion toward the category of auditory emotion.

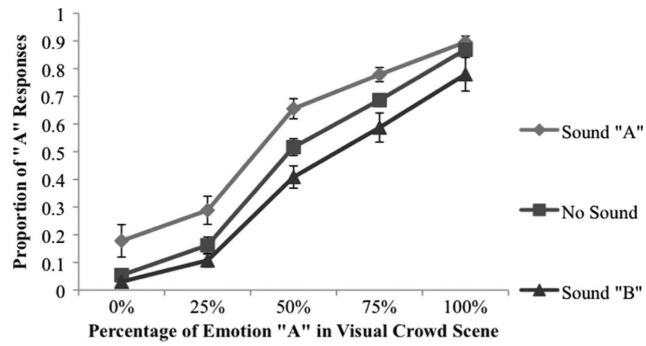

Figure 2. Mean proportion of category "A" responses across participants for all five levels of proportion, in the three sound conditions in Experiment 1a (error bars represent ± 1 *SE* of the mean).

A second repeated measures ANOVA was performed on the reaction time data, again with proportion and sound as factors. There was a significant main effect of proportion $F(4, 68) = 9.3$, $p < .01$. Tukey post hoc analyses on the effect of proportion found that, as expected, reaction times for the crowds at the 0% and the 100% proportion levels were both significantly faster than to crowds with 50% ($p < .05$) or 75% ($p < .05$) proportions. Reaction times to the 0% proportion levels were also faster than those to the 25% proportion ($p < .05$). There was no effect of sound $F(2, 34) < 1$ nor any evidence of an interaction between the factors $F(8,136) = 1.3$, $p > .05$.

Experiment 1b

Reaction times were first analyzed for outliers, which were subsequently removed, leading to the removal of 2% of the reaction time data. The average reaction time was 1985 ms. First, the accuracy and reaction times data were compared across each of the four emotional categories in the "no vision" condition only, as shown in Table 2.

Because emotional categorization performance was relatively low for some categories compared to Experiment 1a, accuracy to each of the emotional categories was compared against chance performance (of 0.25) using a single-sample *t* test. The number of accurate responses for each emotion was found to be significantly better than chance (anger: $t(19) = 3.09$, $p < .01$; fear, $t(19) = 6.9$, $p < .01$; happy, $t(19) = 6.77$, $p < .01$; and sad, $t(19) = 23.2$; $p < .01$).

Separate repeated-measures ANOVA were performed on the accuracy and response time data from the 100% auditory emotional crowds in the "no vision" condition, with emotional category (anger, fear, happy, and sad) as a factor. For accuracy a significant effect of emotional category was found $F(3, 57) = 15.07$; $p < .01$: Tukey post hoc tests revealed that accuracy to the angry auditory crowds was significantly lower than to the fearful ($p < .05$), happy ($p < .01$), or sad ($p < .01$) crowds and accuracy

[2] One consequence of this recoding was that responses to each of the individual emotion categories pairs were not analyzed. These specific analyses were, however, not the main interest of the study, which was whether sounds can affect the perception of the emotional category of the visual crowd.

Table 2
Mean Proportion of Accurate Responses and Mean Reaction Times Across Participants in Experiment 1b to the 100% Auditory Crowd Stimuli in Each of the Four Emotion Categories, With No Visual Information Presented (Standard Deviations Shown in Parentheses)

Emotional category	Reaction times (ms)	Proportion of accurate responses
Anger	1987 (*SD* = 615)	0.43 (*SD* = 0.26)
Fear	1988 (*SD* = 625)	0.65 (*SD* = 0.25)
Happy	1826 (*SD* = 612)	0.74 (*SD* = 0.34)
Sad	1390 (*SD* = 561)	0.93 (*SD* = 0.13)

was significantly lower to the fearful crowds than to the sad auditory crowds ($p < .05$). For the response times, a significant effect of emotional category was also found $F(3, 57) = 16.2$, $p < .01$: response times were significantly faster to the sad auditory crowds than to either the angry (Tukey post hoc, $p < .01$), fearful ($p < .05$), or happy ($p < .05$) crowds, with no other significant differences found.

All category responses were again recoded into one of two emotional categories (emotion "A" or emotion "B") as described in Experiment 1a. A repeated-measures ANOVA was performed with roportion of auditory emotion "A" in the auditory crowd (0%, 25%, 50%, 75%, 100%) and visual emotion condition (visual emotion "A," visual emotion "B," or "no vision") as factors. A main effect of roportion $F(4, 76) = 142.2$, $p < .01$, was found with significant increases in "A" responses with each increase in the proportion of characters in the crowd vocalizing emotion "A" (Tukey post hoc, $p < .05$ for each pair of steps along the continuum). It is important a main effect of visual emotion $F(2, 38) = 31.85$, $p < .01$, was also found. Tukey analyses revealed that there were significantly more category "A" responses to the visual emotion "A" condition than to either the "no vision" condition ($p < .01$) or the visual "B" condition ($p < .01$) and that there were more "A" responses to the "no vision" condition than to the visual "B" condition ($p < .05$).

A significant interaction was also found between visual emotion and proportion $F(8, 152) = 3.145$, $p < .01$, as shown in Figure 3. Tukey post hoc analyses on this interaction revealed that at the 50% proportion level, there was a significantly greater number of "A" responses to the "no vision" condition than to the visual emotion "B" condition ($p < .05$) but responses to the visual emotion "A" condition were not significantly greater than the "no vision" condition. Only at the 75% proportion level were the number of "A" responses made to the "no condition" significantly fewer than to the visual emotion "A" condition and significantly greater than the visual emotion "B" condition ($p < .01$). The differences between the visual emotion responses at the other levels of proportion failed to reach significance.

A second repeated-measures ANOVA was then performed on the reaction times data, with proportion and visual condition as factors. No main effects of proportion $F(4, 76) < 1$, nor of visual condition $F(2, 38) = 1.09$, $p > .05$, were found, nor was there evidence of an interaction between these factors $F(8, 152) = 1.1$, $p > .05$. However, although not significant, reaction times to the 0% and 100% were almost 200 ms faster on average than those to the other crowd proportions (1820 ms and 2004 ms respectively).

Discussion

The results of Experiment 1a suggest that the sound of a crowd influenced the perceived emotion of the visual crowd. In Experiment 1b, the results suggest that the seen emotion of a crowd influences the perceived sound of the emotion of a crowd. These findings suggest a bidirectional influence of sound and vision on the perceived emotion of crowds and are consistent with previous findings on the bidirectional influence of vision and audition in perceiving emotions from an individual character (see, e.g., de Gelder & Vroomen, 2000a and b; van den Stock et al., 2007; 2008). Based on the results of Experiments 1a and 1b, therefore, the integration of emotional information across modalities seems to occur with crowd stimuli in a similar manner to audiovisual integration in the perception of emotion from individuals (e.g., de Gelder & Vroomen, 2000a).

In both Experiments 1a and b the proportion of characters in a crowd expressing one emotion over another had an effect on the perceived emotion of the visual crowd. First, the more characters in a crowd visually or vocally expressing an emotion, the more that emotion dominated the percept. As expected, crowds in which the characters consistently expressed one emotion (i.e., 100% or 0% proportions) were categorized faster than those containing two emotions. The proportion of characters expressing an emotion in the crowd scene was found to be a significant predictor of the emotional category responses in each of the sound conditions in Experiment 1a and the visual conditions in Experiment 1b. These data add further support to the previous finding reported by McHugh et al., (2010) that the emotion of a crowd is not perceived as categorical, and extends this finding to the auditory domain.

The results of Experiment 1a and 1b differ, in that relative to the baseline conditions, (i.e., no sound or no vision respectively) the effect of the information from the unattended modality biased the perceived emotion in the attended modality in different ways. In particular, in Experiment 1b when all of the characters in the crowd consistently expressed the same auditory emotion "A" (i.e., 0% or 100% proportions) the number of correct responses made to that emotion increased when the visual emotion of the crowd was congruent (i.e., also emotion "A"). The reverse was also found in that when the visual crowd expressed a different emotion (i.e., emotion "B") this nevertheless biased the perceived emotional

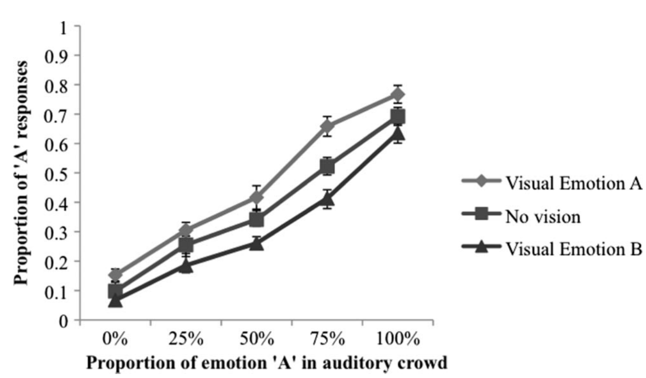

Figure 3. Means of proportion "A" responses to each of the five levels of proportion in the auditory crowds, for the three visual emotion conditions in Experiment 1b (error bars represent ± 1 *SE* of the mean).

sound of the crowd toward the visual emotional category relative to when the auditory and visual emotions were consistent (i.e., 100% emotion "A" and visual emotion "A"). In contrast to these results, the results of Experiment 1a suggest that the auditory emotion had a less of an effect on the perceived visual emotion of the crowd when the majority of characters expressed one emotion over another (i.e., at the 100% emotion "A" proportions). Moreover, when the auditory emotion was consistent with the emotion expressed by the majority of the visual crowd, then sound had less of an effect on the percept relative to when there was no sound present. Only when the emotion in the sound was inconsistent with the emotion expressed by a majority of the visual characters was the perceived emotion affected. Our results suggest an asymmetry in the effects of vision and audition on the perceived emotion of a crowd, with visual information more likely to dominate the percept. These findings are consistent with previous studies, suggesting that emotion recognition is more efficient through vision than audition (Adolphs, 2002; Walk & Homan, 1984).

More pertinently, when the emotion of the crowd was ambiguous (i.e., when 50% of the characters expressed one emotion and the other 50% expressed another emotion), information from the unattended modality had a significant effect on changing the perceived emotion of the crowd in the attended modality. For example, when the sound of an emotional crowd was presented with a visually ambiguous crowd, this sound biased the visual perception of the emotion of the crowd toward the emotion of the sound of the crowd. The corollary was also found in that the sight of an emotional crowd biased the perceived emotion from an ambiguous sound of the crowd. In other words, when more reliable information was available in the unattended modality, this was sufficient to disambiguate the emotion of the crowd in the attended modality. However, stronger bimodal effects on the categorization of emotion from an ambiguous signal were found in Experiment 1a than in 1b. This result possibly reflects visual dominance in the perception of emotion, as argued above, or differences in the relative reliability of the information across vision and audition (as described in Bayesian approaches to sensory integration such as Ernst & Banks, 2002; Alais & Burr, 2004; Burr & Alais, 2006) on the emotion of the crowd.

The results of Experiment 1b suggest a stronger effect of the sight of the crowd on the perceived emotion from the sound of the crowd when 75% of the auditory crowd consistent expressed an emotion rather than the ambiguous crowd of 50%, relative to baseline. A possible explanation for this result is that, relative to the other emotional categories, anger comprised 33% of all emotion "A" category responses in the study, as it was arbitrarily labeled emotion "A" in two of the six emotional continua. As was observed in Experiments 1a and 1b, anger was the least recognizable emotion in both modalities. The assignment of anger as visual emotion "A" may have increased responses at the 75% proportion level relative to the 50% proportion level, due to the relative difficulty recognizing anger in crowds. At the 75% proportion level, therefore, the increased proportion of anger in the auditory modality coupled with the visually presented angry crowd may have facilitated anger recognition, as reflected in a relative increase in category "A" responses.

In sum, despite participants being told to ignore either the auditory information (Experiment 1a) or the visual information (Experiment 1b), the emotional information from the unattended modalities significantly modulated the categorization of the emotion in the attended modalities. This finding was most pronounced when the emotion of the crowds was ambiguous. Our results suggest that emotional information from the auditory and visual modalities was integrated in the perception of the emotion of a crowd and is consistent with previous findings on the perceived emotion from single characters (de Gelder & Vroomen, 2000a; Massaro & Egan, 1996; van den Stock et al., 2007; van den Stock et al., 2008).

In Experiment 1a proportion had an effect on reaction times but the effect of proportion on responses times when the attended modality was sound (Experiment 1b) failed to reach significance. It is likely that the effect of proportion found in Experiment 1a reflects serial processing of the characters within the visual crowd in order to determine the overall emotion of the crowd: faster response times are more likely when all characters consistently express the same emotion, relative to mixed crowds, since fewer characters would require processing to perceive the overall emotion. Since the characters in a visual crowd can be parsed easier than those in an auditory crowd, serial processing is unlikely to affect the duration of response when perceiving the emotion from auditory crowds. To further investigate these effects, in the following experiment we manipulated the size of the crowd to investigate whether the number of characters in a crowd would affect the integration of auditory and visual information in the perception of the emotion of a crowd.

Experiment 2

The results of Experiments 1a and 1b suggest bidirectional, crossmodal influences on the perceived emotion of the crowd, in line with previous results based on the perception of emotion from individuals (de Gelder & Vroomen, 2000a). Hence crossmodal influences on the perceived emotion may be independent of the number of items perceived (either number of visual characters or number of voices). Alternatively, audiovisual integration may be affected by crowd size. For example, perception of emotion may benefit more from congruent crossmodal audio or visual emotional information when crowd size increases, since a larger visual or auditory crowd size may represent an inherently more complex and ambiguous stimulus. To investigate these effects, we created two crowd sizes in each modality: one containing 12 characters, and the other containing 32 characters, for use as stimuli in the following experiment.

Method

Participants

Twenty participants (three male, with an age range from 18 to 41 years and a mean age of 20.95 years) were recruited from the student population at Trinity College Dublin for this study. All reported normal or corrected-to-normal vision and none reported auditory impairment.

Stimuli and Apparatus

The stimuli used in this experiment were similar to those described in Experiment 1 with the exception that the crowds used in the current study differed in the number of characters they contained. Two different visual crowd sizes were created, each with either 12 or 32 characters. The overall visual angle of the crowd scene and size of the characters remained consistent across experiments and crowd sizes such that increasing or decreasing the number of individuals affected crowd density only. The auditory crowd stimuli also comprised of 12 or 32 characters with all characters in the crowd consistently vocalizing the same emotion. These auditory crowds were created according to the same procedure used to make the mixed auditory emotional crowd stimuli described in Experiment 1a, except that one emotion only was present in each auditory emotional crowd. Auditory and visual emotional crowd stimuli were paired such that crowd sizes were always matched, (i.e., auditory 12 character crowds were only ever paired with visual 12 character crowds and the same with the 32 character crowds).

Design and Procedure

The experimental design and procedure here were identical to that used in Experiment 1a: participants had to judge the emotion displayed in the visual crowd while ignoring the accompanying auditory cue. As in the previous experiment, each of the visual crowds was presented in three different auditory conditions; with sound "A," sound "B" or with no sound. This gave a total of 66 stimuli in each crowd size condition, leading to a total of 132 stimuli. Each stimulus was repeated across two experimental blocks, with trial order fully randomized within each block.

Results

The mean accuracy and reaction time responses for each of the four emotion categories in the "no sound" condition were calculated over the 100 emotion conditions and for each crowd size. The mean accuracy and reaction times per emotional category across all participants are presented in Table 3.

Separate repeated measures ANOVA were performed on the accuracy and response time data, with category of 100% emotion category and size of crowd (12 or 32) as factors. For accuracy, was no evidence of an effect of either emotion, $F(3, 57) > 1$, nor crowd size, $F(1, 19) = 3.4$, $p = .08$, although accuracy was slightly better for smaller crowds and lowest for the large crowds. A significant interaction was found between emotion and size, $F(3, 57) = 3.6$, $p < .05$. Tukey post hoc analyses revealed that responses to the fearful crowds were significantly worse for the larger than smaller crowd sizes ($p < .05$). Similarly, performance to the happy crowds was also less accurate for the large than small crowd size ($p < .05$). There were no other differences found. For the response times, again there were no main effects found for either the emotion category, $F(3, 57) = 1.2$, $p > .05$, or for crowd size, $F(1, 19) < 1$, nor was there an interaction between these factors, $F(3, 57) = 1.7$, $p > .05$.

All responses were then recoded as proportion of "A" responses (as described above) across emotional continua for each of the proportion, size, and sound conditions. The data analyzed were proportion of "A" responses and reaction times. Reaction times were analyzed for outliers, which were then removed, leading to the removal of 3% of all of the reaction time data. The average reaction time to all trials was then 2069 ms.

A repeated-measures ANOVA was performed on the emotional category "A" responses, with crowd size (12, 32), proportion (0%, 25%, 50%, 75%, and 100%), and sound condition (sound "A," sound "B," no sound) as factors. A significant main effect of proportion $F(4, 76) = 308.5$, $p < .01$, was found with significant differences between each of the proportion of characters in the crowds (Tukey post hoc, $p < .05$). There was a main effect of sound, $F(2, 38) = 17.9$, $p < .01$, with significantly more "A" responses made to the sound "A" condition than either the sound "B" or the "no sound" conditions (Tukey post hoc test, $p < .05$), but the difference between sound "B" and "no sound" conditions failed to reach significance. There was no effect of crowd size, $F(1, 19) < 1$.

We found a significant interaction between sound and proportion of emotion, $F(8, 152) = 3.5$, $p < .05$, which is plotted in Figure 4. Tukey post hoc analyses revealed that at 0% and 100%, there was no effect of sound. However, at the 25%, 50%, and 75% proportions, there were significantly more "A" responses made to the sound "A" condition than the sound "B" condition ($p < .05$). At the 50% proportion, there were significantly more "A" responses made to the sound "A" than the "no sound" condition ($p < .05$). There was no evidence that significantly fewer "A" responses were made to the sound "B" than the "no sound" conditions at any proportion (although this effect was close to significance at the 50% proportion level, $p = .07$). Also, the interactions between size and proportion, $F(4, 76) = 2.2$, $p > .05$, or size and sound, $F(2, 38) = 1.8$, $p > .05$, failed to reach significance. There was no three-way interaction between crowd size, sound and proportion, $F(8, 152) < 1$.

The same repeated-measures ANOVA performed on the reaction times data, revealed a significant main effect of proportion, $F(4, 76) = 40.4$, $p < .01$, with significantly faster reaction times to both the 0% and 100% conditions than to either the 25%, 50% or 75% conditions (Tukey post hoc analysis, $p < .05$). There was a main effect of sound, $F(2, 38) = 6.7$, $p < .01$, with reaction times to the "no sound" condition significantly slower than those to either the sound "A" or the sound "B" conditions ($p < .05$) and no difference between reaction times to the sound "A" and the sound "B" conditions ($p > .05$). There was no effect of crowd size, $F(1, 19) < 1$. The interaction between proportion and sound was significant, $F(8, 152) = 2.9$, $p < .01$. Tukey post hoc analyses into the interaction between sound and proportion showed that at the 0% proportion of emotion only, reaction times to the sound "B" condition were significantly faster than those to the sound "A" condition ($p < .05$) with no other significant differences between

Table 3
Mean Proportion of Accurate Responses and Mean Reaction Times Across Participants in Experiment 2 to the 100% Visual Crowd Stimuli in Each of the Four Emotion Categories Across Both Crowd Sizes (Standard Deviations in Parentheses)

Emotional category	Reaction times (ms)	Proportion of accurate responses
Anger	2020 (SD = 697)	0.88 (SD = 0.12)
Fear	1958 (SD = 697)	0.84 (SD = 0.12)
Happy	1826 (SD = 688)	0.87 (SD = 0.16)
Sad	1972 (SD = 639)	0.90 (SD = 0.12)

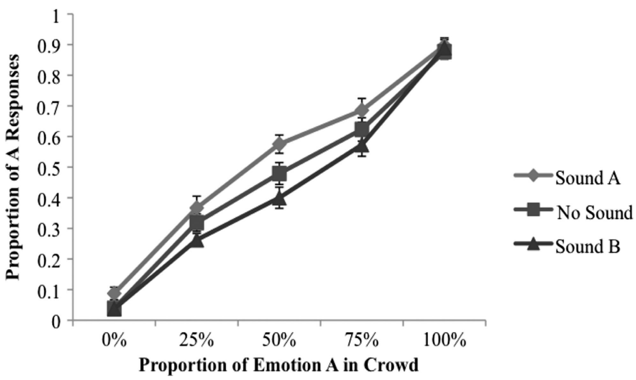

Figure 4. Means of proportion "A" responses to each of the five levels of proportion in the visual crowds, for the three auditory emotion conditions in Experiment 2 (error bars represent ± 1 *SE* of the mean).

sound conditions at any proportion. Neither the interactions between proportion and size, $F(4, 76) = 1.1, p > .05$, size and sound, $F(2, 38) < 1$, nor proportion, size and sound, $F(8, 15) < 1$, were significant.

Discussion

The results of Experiment 2 suggest that there was no effect of the number of characters on the perceived emotion of the crowd through their body expressions nor was the time to categorize the emotion of the visual crowd affected by crowd size. This latter finding was unexpected since the results of a previous study reported by McHugh et al., (2010) reported faster response times for smaller (i.e., 12 characters) than larger (32 character) crowd sizes in judging the emotion of a crowd. However, the previous study was based on unimodal information only, that is, vision, therefore it is possible that the presence of sound information generally affects visual emotional categorization even for large crowd sizes. Finally, crowd size did not interact with any of the other factors such as the emotional category of the sound or the proportion of characters in the crowd expressing one of two emotions.

As in Experiment 1a, the sound "A" condition produced a higher level of "A" responses than the sound "B" or the baseline no sound conditions. However, the presence of emotional sound "B" did not significantly reduce the number of emotion "A" category responses to the visual stimulus relative to baseline, although fewer "A" responses were made in the sound "B" condition. Therefore, although the congruent emotional information in the auditory modality affected the perceived emotion in the visual crowd overall, there was no cost observed when the emotion in the auditory modality was incongruent to that presented visually. However, when the visual emotion of the crowd was ambiguous such that half of the characters expressed one emotion over another, then the difference between the number of "A" responses made to the sound "B" condition relative to baseline was greatest, with fewer "A" responses in this sound condition than baseline (see Figure 3). It is unclear why more "A" responses were made in sound "A" condition relative to baseline but that the complementary cost on the number of "A" responses to the sound "B" conditions was not found. One possibility though is that this is related to the arbitrary choice of emotions as emotion "A" (as discussed in Experiment 1b with relation to the anger emotion). Nevertheless, the pattern of findings mirror those found in the previous experiments suggesting that sound can bias the perceived emotion of a visual crowd.

Previous research found that when an unambiguous auditory emotional signal was presented with an ambiguous facial expression, audition had a greater influence on the perceived facial expression relative to when the facial expression was unambiguous (de Gelder & Vroomen, 2000a). Collignon and colleagues (2008) also found that when the information in one modality was relatively unreliable, the emotional information in the unambiguous modality had a stronger effect on the resultant audiovisual percept. In the present experiment, visual information on the emotions from crowd proportions of 25%, 50%, and 75% would be considered ambiguous or visually unreliable, since these crowds contain characters expressing one of two different emotional categories. Indeed, a larger effect of auditory information was found at these proportion levels, and may be consistent with the predictions of Bayesian approaches to multisensory interactions (Ernst & Banks, 2002; Burr & Alais, 2006). Moreover, when visual information was unambiguous, that is, at the 0% and 100% visual crowd proportions), audition had less of an influence on the perceived visual emotion of the crowd.

General Discussion

The current experiments investigated the effect of crossmodal emotional information in one modality on the emotional category perceived in another modality, and whether this process was affected by the manipulation of crowd size in both modalities. Previous studies on the perception of the emotion of a single individual suggested that multisensory integration is mandatory, since even when participants are told to ignore information in one modality, the emotional content of that information still affected the recognition of emotion in the other modality (e.g., de Gelder & Vroomen, 2000a and b). Recent findings from Haberman and Whitney (2009, 2010) suggest that the number of individuals expressing the emotion does not affect the perceived emotion since the average emotion is rapidly and reliably detected. As such, we expected that the effects of crossmodal influences on the perceived emotion of a crowd of characters would be similar to those reported in studies involving single characters. Moreover, if multisensory integration of emotional information is independent of the nature of the encoded information (as suggested by Massaro & Egan, 1996; van den Stock et al., 2008; 2009) it was expected that the perceived emotion would be independent of the number of items in a stimulus.

However, there is some evidence to suggest that the perception of emotion in crowds differs from that in individuals. For example, visual emotions are categorically perceived from individuals (Etcoff & Magee, 1992; Young et al., 1997) but we previously failed to find evidence of categorical perception of emotion from crowds (McHugh et al., 2010). It is possible that the perceived emotion from a crowd of characters may differ from that perceived from an individual.

The results of Experiment 1a suggest that despite instructions to ignore the auditory emotional information, audition affected the categorization of emotion in the visual crowds. Moreover, the results of Experiment 1b, which investigated the role of unattended visual emotional information on the perceived emotion of the sound of a crowd, suggested that this effect is bidirectional. These findings are consistent with the literature on the perceived emotion from individuals, reporting bidirectional links between emotion perception in the auditory and the visual modalities (De Gelder & Vroomen; 2000a; van den Stock et al., 2007).

In Experiment 2, instead of using crowds of 20 visual characters, a smaller crowd of 12 characters and a larger crowd of 32 characters in each modality were used. We failed to find evidence that the size of a crowd affected the multisensory integration, since audition had the same influence on the perceived visual emotion irrespective of crowd size. The fact that size did not affect perception in this instance indicates that serial processing of the scene from one character to the next is not occurring. This is the same finding reported by Haberman and Whitney (2010) and may support their hypothesis of a mechanism involving the use of a summary statistical representation underlying the perception of an average emotional expression.

There was no effect of sound at either the 0% or the 100% visual proportions of the crowds in Experiment 2. In contrast, however, audition still had an effect on the perceived visual emotion of the crowd in Experiment 1a. One main difference between these experiments was that different numbers of characters were contained in the crowd scenes: in Experiment 1a there were 20 characters whereas in Experiment 2 there were either 12 or 32 characters in each crowd display. In a previous study, McHugh et al., (2010), reported effects of crowd sizes on the perceived emotion of the visual crowd such that responses were more accurate to the small (12) and large (32) crowd sizes than to the medium sized (20) crowds. This previous finding may suggest that visual emotion is relatively easier to resolve in medium size crowds, for reasons yet unknown but are possibly related to crowd density, and therefore vision may be more likely to dominate perception even when both vision and audition are unambiguous. For the smaller and larger crowd sizes, visual emotion may be less efficiently perceived, thus allowing for audition to have more of an influence on the final percept.

Since we investigated the effect of crowd size in the visual modality in Experiment 2, an obvious next step would be to investigate the effect of crowd size in the auditory modality using the same paradigm. However, there are particular issues with investigating auditory crowd scenes. Namely, it is difficult to segregate an auditory crowd scene into individual sources (e.g., Bregman, 1990; Snyder & Alain, 2007) and the ability to derive meaning from individual sounds within an auditory crowd is known to be a perceptually difficult task (Cherry, 1953). While we acknowledge that auditory crowd perception would be interesting for future research, an investigation of crowd size effects in the auditory modality was not feasible in our current study because, for reasons given, the perception of auditory crowds would not be sufficiently similar to the perception of visual crowds to allow us make any useful comparisons.

The results of the current experiments suggest that emotional information from an unattended modality can bias the perceived emotion in the attended modality. These crossmodal influences occur for crowd perception in the same way as demonstrated for the individual perception (de Gelder & Vroomen, 2000a and b; van den Stock et al., 2007; 2008; Massaro & Egan, 1996; Collignon et al., 2008; Pell, 2005). Moreover, our results suggest that there are bidirectional links between emotion processing in the auditory and visual modalities, since participants were unable to ignore information from either modality.

Crowds can be seen to represent experimental stimuli within which ambiguity can be easily manipulated by altering the consistency with which emotions are expressed across all the characters in the crowd scene. The findings from the current study give rise to interesting questions for future research. In particular, the spatiotemporal limits of multisensory influences on the perceived emotion of a crowd are unknown. Moreover, as we alluded to earlier, it is unclear how crowd density or individual groupings can affect the perceived emotion of a crowd. The findings from the present study suggest support for the modal proposed by Haberman and Whitney (2010) that a summary representation of a scene of emotional stimuli is processed. Their model provides an explanation for the finding that performance in the current study appeared unaffected by sample size or by outliers (which would potentially be equivalent to the minority emotions in the 25% proportion crowds). Haberman and Whitney suggested that the mechanism underlying the rapid perception of emotion from a set of faces is an automatic process, which does not require top-down or serial processing to occur. It is not yet known whether such ensemble coding, as termed by Haberman and Whitney, extends to multisensory percepts although it is not unreasonable to assume that a summary representation of a set of emotional voices also occurs during the perceived emotion of an auditory crowd. Moreover, this is consistent with other models such as the FLMP, which purports that emotional cues are integrated regardless of their sensory origin or modality (de Gelder & Bertelson, 2003; Massaro & Egan, 1996). The extent to which ensemble coding can account for multisensory emotion perception requires further elucidation, however, since the finding that crossmodal information influences the perceived emotion in another sensory modality suggests that a summary of the combined multisensory percept does not occur. Nevertheless, it is hoped that the current study represents a step toward clarifying this account of multisensory emotion perception.

References

Adolphs, R. (2002). Neural mechanisms for recognizing emotion. *Current Opinion in Neurobiology, 12,* 169–178.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14,* 257–262.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences, 4,* 267–268.

Benjamin, E. (2002, October). Comparison of objective measures of loudness using audio program material. Presentation to 113th Convention of the Audio Engineering Society, Los Angeles.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: MIT Press.

Burr, D. C., & Alais, D. (2006). Combining visual and auditory information. *Progress in Brain Research, 155,* 243–258.

Cherry, E. C. (1953). "Some experiments on the recognition of speech, with one and with two ears." *Journal of Acoustic Society of America, 25,* 975–979.

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audiovisual integration of emotional experience. *Brain Research, 1242,* 126–135.

de Gelder, B. (2009). Why bodies? Twelve reasons for including bodily

expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364,* 3475–3484.
de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Science, 7,* 460–467.
de Gelder, B., Teunisse, J. P., & Benson, P. J. (1997). Categorical perception of facial expressions: Categories and their internal structures. *Cognition & Emotion, 11,* 1–23.
de Gelder, B., & Vroomen, J. (2000a). The perception of emotion by ear and eye. *Cognition & Emotion, 14,* 289–311.
de Gelder, B., & Vroomen, J. (2000b). Bimodal emotion perception: Integration across separate modalities, cross-modal perceptual grouping or perception of multimodal events? *Cognition and Emotion, 14,* 321–324.
Dolan, R. J., Morris, J. S., & de Gelder, B. (2001). Crossmodal binding of fear in voice and face. *Proceedings of the National Academy of Sciences of the United States of America, 98,* 10006–10010.
Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415,* 429–433.
Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition, 44,* 227–240.
Haberman, J., & Whitney. D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17,* 751–753.
Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 718–734.
Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention Perception and Psychophysics, 72,* 1825–1838.
Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional contagion.* New York: Cambridge University Press.
Juth, P., Lundqvist, D., Karlsson, A., & Ohman, A. (2005). Looking for foes and friends: Perceptual and emotional factors when finding a face in the crowd. *Emotion, 5,* 379–395.
Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and face. *Psychonomic Bulletin & Review, 3,* 215–221.
MATLAB. (2008). The MathWorks Inc., Natick, MA, USA.
McHugh, J. E., McDonnell, R., O'Sullivan, C., & Newell, F. N. (2010). Perceiving emotion in crowds: The role of dynamic body postures on the perception of emotion in crowded scenes. *Experimental Brain Research, 204,* 361–372.
Pell, M. D. (2005). Prosody-face interactions in emotional processing as revealed by the facial affect decision task. *Journal of Nonverbal Behavior, 29,* 193–215.
Pinkham, A. E., Griffin, M., Baron, R., Sasson, N. J., & Gur, R. C. (2010). The face in the crowd effect: Anger superiority when using real faces and multiple identities. *Emotion, 10,* 141–146.
Snyder, J. S., & Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin, 133,* 780–799.
Van den Stock, J., Grèzes, J., & de Gelder, B. (2008). Human and animal sounds influence recognition of body language. *Brain Research, 1242,* 185–190.
Van den Stock, J., Peretz, I., Grèzes, J., & de Gelder, B. (2009). Instrumental music influences recognition of emotional body language. *Brain Topography, 21,* 216–220.
Van den Stock, J., Righart, R., & de Gelder, B. (2007). Body expressions influence recognition of emotions in the face and voice. *Emotion, 7,* 487–494.
Vroomen, J., Driver, R. J., & de Gelder, B. (2001). Is crossmodal integration of emotional expressions independent of attentional resources? *Cognitive, Affective & Behavioural Neuroscience, 1,* 382–387.
Walk, R. D., & Homan, C. P. (1984). Emotion and dance in dynamic light displays. *Bulletin of the Psychonomic Society, 22,* 437–440.
Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., & Perrett, D. I. (1997). Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition, 63,* 271–313.

Received March 21, 2011
Revision received May 27, 2011
Accepted May 31, 2011 ■