

University of Groningen

Eyes on Emotion

de Boer, Minke J.; Başkent, Deniz; Cornelissen, Frans W.

Published in:
Multisensory research

DOI:
[10.1163/22134808-bja10029](https://doi.org/10.1163/22134808-bja10029)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

de Boer, M.J., Başkent, D., & Cornelissen, F.W. (2020). Eyes on Emotion: Dynamic Gaze Allocation During Emotion Perception From Speech-Like Stimuli. *Multisensory research*, 34(1), 17-47.
<https://doi.org/10.1163/22134808-bja10029>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Eyes on Emotion: Dynamic Gaze Allocation During Emotion Perception From Speech-Like Stimuli

Minke J. de Boer^{1,2,3,*}, Deniz Başkent^{1,2} and Frans W. Cornelissen^{1,3}

¹ Research School of Behavioural and Cognitive Neurosciences (BCN), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

² Department of Otorhinolaryngology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

³ Laboratory for Experimental Ophthalmology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Received 20 November 2019; accepted 29 May 2020

Abstract

The majority of emotional expressions used in daily communication are multimodal and dynamic in nature. Consequently, one would expect that human observers utilize specific perceptual strategies to process emotions and to handle the multimodal and dynamic nature of emotions. However, our present knowledge on these strategies is scarce, primarily because most studies on emotion perception have not fully covered this variation, and instead used static and/or unimodal stimuli with few emotion categories. To resolve this knowledge gap, the present study examined how dynamic emotional auditory and visual information is integrated into a unified percept. Since there is a broad spectrum of possible forms of integration, both eye movements and accuracy of emotion identification were evaluated while observers performed an emotion identification task in one of three conditions: audio-only, visual-only video, or audiovisual video. In terms of adaptations of perceptual strategies, eye movement results showed a shift in fixations toward the eyes and away from the nose and mouth when audio is added. Notably, in terms of task performance, audio-only performance was mostly significantly worse than video-only and audiovisual performances, but performance in the latter two conditions was often not different. These results suggest that individuals flexibly and momentarily adapt their perceptual strategies to changes in the available information for emotion recognition, and these changes can be comprehensively quantified with eye tracking.

Keywords

Emotion perception, perceptual strategies, audiovisual integration, gaze allocation, dynamic, eye movements

* To whom correspondence should be addressed. E-mail: minke.de.boer@rug.nl

1. Introduction

Successful social interactions involve not only an understanding of the verbal content of one's conversational partner, but also their emotional expressions. In everyday life, the majority of social interactions takes place as face-to-face communication and emotion perception is thus multimodal and dynamic in nature. Historically, however, emotion perception has been investigated in a single perceptual modality, with static facial emotional expressions being studied most commonly. These unimodal studies have shown one can discriminate between broad emotion categories from visual cues, such as from activations of specific facial muscle configurations (Bassili, 1979; de Gelder *et al.*, 1997; Ekman and Friesen, 1971) but also from specific body movements and postures (de Gelder, 2009; Jessen and Kotz, 2013), as well as from auditory cues, such as prosodic speech information (Banse and Scherer, 1996; Juslin and Laukka, 2003).

The vast amount of literature on multisensory perception in general indicates that observers integrate information in an optimal manner, by weighing the unimodal information based on its reliability prior to linearly combining the now weighted unimodal signals. Because of this, the multimodal benefit, i.e., the strength of the multimodal integration, in perception is largest when the reliability of the unimodal cues is similar and each sense provides unique information. Likewise, when one sense is much more reliable — such as hearing for time interval estimation — this sense will receive a higher weight and the multisensory signal could be roughly equal to the most reliable unisensory signal (see, e.g., Alais and Burr, 2004; Ernst and Banks, 2002; Ernst and Bühlhoff, 2004). However, while it is well known that observers integrate optimally, it is unknown if they also employ specific perceptual strategies when integrating. For example, how different is the visual exploration of an object when the observer is allowed to touch the object compared to when the observer is not allowed to touch the object? Here, we investigated such multisensory perceptual strategies, and the manner in which they adapt to the presence of multiple modalities, by measuring observers' viewing behavior in the context of emotion perception.

The continual adjustments of weighting unimodal information for multisensory perception make audiovisual integration a flexible process. Consequently, it can be expected that the viewing behavior observers employ also reflects this flexibility. It is long known that people naturally tend to foveate the regions of an image that are of interest (e.g., Yarbus, 1967). What is of interest in an image is defined by visual saliency (Itti and Koch, 2000), but also by the nature of the perceptual task (see, e.g., Hayhoe and Ballard, 2005). Võ and colleagues (2012) proposed that gaze allocation is a functional, information-seeking process. They performed an eye-tracking study in which participants

were asked to rate the likeability of videos featuring pedestrians engaged in interviews. When the video was shown with the corresponding audio, participants mostly looked toward the eyes, nose, and mouth. When the audio signal was removed, there was a decrease in fixations to the face in general, and to the mouth in particular. Thus, despite the fact that the visual signal remained unchanged, the viewing behavior changed, indicating that viewing behavior is not only directed by visual information but also by information in other modalities. These findings led the authors to conclude that gaze is allocated on the basis of information-seeking control processes. On the other hand, one could instead argue that gaze was still mostly guided by saliency. Audiovisual synchrony likely increases the saliency in certain image regions, which are then fixated more often. If the audiovisual synchrony disappears when the video is muted, the saliency of the mouth decreases and it is looked at less. On the other hand, Lansing and McConkie (2003), using video recordings of everyday sentences showing only the face of the speaker, found an increase in fixations on the mouth when the video was presented without sound. The participants' task was quite different from that in Vö *et al.* (2012) however, as here participants were required to repeat the spoken sentence. In this study (Lansing and McConkie, 2003), the mouth provides the majority of the information relevant for the task and gaze is thus directed toward it, and even more so when the task is made more difficult by removing the audio. Hence, while both these studies (Lansing and McConkie, 2003; Vö *et al.*, 2012) used similar stimuli, the findings are drastically different, which would indicate that gaze allocation is indeed a flexible information-seeking process.

While speech sounds are mainly produced with mouth movements, many facial features additionally contribute to emotional expressions. Emotion perception from speech may thus be more complex than speech perception in terms of predicting gaze allocation. Naturally, in face-to-face communication, humans do not observe an isolated face, but a dynamic whole body that contributes with gestures and posture that may be relevant for recognizing emotions. It has been shown that observers can, under some conditions, recognize emotions from bodily expressions equally well as they can from facial expressions (see de Gelder, 2009 for a review). Additionally, studies showed that emotional prosody (such as pitch, tempo, and intensity) affects what facial emotion is perceived when the emotion in the voice is incongruent with the emotion in the face (de Gelder and Vroomen, 2000; Massaro and Egan, 1996). It has also been shown that visual attention is guided by emotional prosody, where observers look more often at faces expressing the same emotion than at faces expressing a different emotion (Paulmann *et al.*, 2012; Rigoulot and Pell, 2012). However, these studies on the integration of facial expressions with emotional prosody mostly used static images as visual stimuli. It could thus be that observers did not necessarily attribute the face and voice to the

same person, or the emotions were not being expressed at the same time. In addition, while vocal emotion always unfolds over time, a static image of a facial expression does not, despite the fact that facial expressions are dynamic in real life.

Therefore, in the present study, aiming for enhanced ecological validity, we presented dynamic multimodal emotional stimuli that always contained congruent emotion cues to express one of twelve different emotions, and also included emotions from the same family, such as anger and irritation. The stimuli were obtained from the Geneva Multimodal Emotion Portrayals (GEMEP) core set (Bänziger *et al.*, 2012), which contains audiovisual video recordings of emotional expressions, with actors uttering a short nonsense sentence in an emotional manner. The video recordings show the actor from the waist up and therefore include both facial expressions as well as body, arm, and hand gestures. These stimuli have been shown to be recognizable well above chance level and were rated to be fairly believable and authentic. We used this stimulus set to measure how auditory and visual information is integrated for emotion perception.

For the purpose of this study, we consider information from two modalities as integrated when the addition of a second modality modulates the perception of the first modality (e.g., Etzi *et al.*, 2018; Samermit *et al.*, 2019; Taffou *et al.*, 2013), or vice versa, or when the two modalities are combined into a unified multimodal percept (see Collignon *et al.*, 2008; Kokinous *et al.*, 2015 for similar descriptions). This combination into a unified percept could be indicated by, e.g., a gain in task performance larger or smaller than expected on the basis of independent summation of auditory and visual information or when an illusory percept arises due to the fusion of incongruent visual and auditory information (McGurk effect; McGurk and Macdonald, 1976). Relevant to our study, one form of integration is when observers alter their viewing strategies under different circumstances and tasks (Buchan *et al.*, 2008; Võ *et al.*, 2012).

Here, we used eye tracking to gain insight into observers' viewing strategies and in what way they extract and make use of information from the stimuli. Based on previous studies examining viewing behavior during emotion perception, we cannot make a clear prediction about which areas will be fixated on most of the time, as most of these studies used static stimuli. However, two scenarios are likely: either gaze is mostly guided by information-seeking processes, or gaze is mostly guided by saliency. From the information-seeking perspective, when the task is to decode a speaker's emotional state — the focus of the current study — and congruent audio is added to a video signal, the audio signal may help in decoding the emotional information, as the information in the two modalities overlaps to some extent. Hence, auditory information could render certain visual information largely redundant, such as the motion of a speaker's mouth. Therefore, it may no longer be necessary to look at the

mouth to retrieve that information and gaze can be directed elsewhere to examine different, potentially more unique, information. Alternatively, emotion recognition may rely mostly on salience, in which case an observer would always look at the most expressive region, such as the mouth for happy expressions and the eyes in angry expressions (in line with Smith *et al.*, 2005). In this case we do not expect any changes in viewing behavior in response to the presence or absence of audio. Consequently, a change in viewing behavior in response to a change in modalities available can provide complementary information to task performance as a measure of audiovisual integration. In order to analyze what regions of the stimulus participants were looking at, we employed an Area-of-Interest (AOI) based analysis. Our AOIs were dynamic to capture the dynamic nature of the stimuli. Previous studies have shown that, when observing faces, most fixations are on the eyes, nose, and mouth (Groner *et al.*, 1984; Walker-Smith *et al.*, 1977). In addition, it has been shown that hand movements are frequent in emotion expression (Dael *et al.*, 2012), hence observing these movements might be useful as well for identifying the expressed emotion. Therefore, we focused our analysis on the fixations on the eyes, nose, mouth, and hands, which all could drastically change in location over the time course of the video.

To assess the presence of audiovisual integration, we evaluated whether the accuracy scores for emotion identification differed for audio-only, video-only, and audiovisual stimulus presentation. A difference in accuracy is an indication of integration and the direction this difference is in indicates whether any changes in viewing behavior are indeed functional, i.e., lead to better performance. Several studies have shown that emotion perception improves when participants have access to more than one modality conveying the same emotion (de Gelder and Vroomen, 2000; Massaro and Egan, 1996; Paulmann and Pell, 2011). Conversely, other studies have implied visual information dominates over auditory information and that — consequently — multimodal information may not necessarily improve emotion recognition and the contribution of the audio may be limited (Bänziger *et al.*, 2009; Jessen *et al.*, 2012; Wallbott and Scherer, 1986). These conflicting findings may be the result of differences in the reliability of the auditory and visual information presented in these studies. Collignon and colleagues (2008) found visual dominance when the stimuli were presented without any noise, but found robust audiovisual integration when they added noise to the visual stimulus. The visual dominance was found despite the fact that the unimodal emotion recognition performance (correct recognition rate) was the same for the noiseless visual and auditory stimuli. Thus, it appears that in noise-free environments, visual information is often treated as more reliable. Based on this, we hypothesized that we would find visual dominance in participants' accuracy scores.

2. Materials and Methods

2.1. Participants

In total, 23 young healthy participants volunteered to take part in the experiment (ten male, mean age = 23 ± 2.3 years, range: 20–31). One participant did not pass all screening criteria (described below in Section 2.2.) and was therefore excluded from the experiment before data collection. One other participant was excluded due to severe difficulties in calibrating the eye tracker. Consequently, 21 participants completed the entire experiment (nine male, mean age = 23 ± 2.4 , range: 20–31) and were included in the data analysis. The sample size was initially based on similar previous studies on audio-visual emotion perception (e.g., Collignon *et al.*, 2008; Paulmann and Pell, 2011; Skuk and Schweinberger, 2013; Takagi *et al.*, 2015) and was subsequently modified in order to ensure proper counterbalancing of the experimental blocks. All participants were given sufficient information about the nature of the tasks of the experiment, but were otherwise naïve as to the purpose of the study. Written informed consent was collected prior to data collection. The study was carried out in accordance with the Declaration of Helsinki and was approved by the local medical ethics committee (ABR nr: NL60379.042.17).

2.2. Screening

Prior to the experiment, potential participants' hearing and eyesight were tested to ensure auditory and (corrected) visual functioning was within the normal range.

Normal auditory functioning was confirmed by measuring auditory thresholds for pure tones at audiometric test frequencies between 125 Hz and 8 kHz. A staircase method, similar to typical audiological procedures, was used to determine the thresholds, in a soundproof booth. Testing was conducted at each ear, always starting with the right ear. In order to participate in the experiment, audiometric thresholds at all test frequencies needed to be as good as or better than 20 dB HL for the better ear.

Normal visual functioning was tested with measurements of visual acuity and contrast sensitivity (CS). These tests were performed using the Freiburg Acuity and Visual Contrast Test (FrACT, version 3.9.8; Bach, 1996, 2006). A visual acuity of at least 1.00 and a logCS of at least 1.80 (corresponding roughly to a 1% luminance difference between target and surround) were used as cutoff thresholds to participate in the experiment. Visual tests were performed on the same computer as used in the main experiment.

Additional exclusion criteria were neurological or psychiatric disorders, dyslexia, and the use of medication that can potentially influence normal brain functioning.

2.3. Stimuli

The stimuli used in this study were taken from the Geneva Multimodal Emotion Portrayals (GEMEP) core set (for a detailed description, see: Bänziger *et al.*, 2012), which consists of 145 audiovisual video recordings (mean duration: 2.5 s, range: 1–7 s) of emotional expressions portrayed by ten professional French-speaking Swiss actors (five female). The vocal content of the expressions were two pseudo-speech sentences with no semantic content but resembling the phonetic sounds in western languages (“nekal ibam soud molen!” and “koun se mina lod belam?”). Out of the total set of 17 emotions, 12 were selected for the main experiment. The selection was made to produce a well-balanced design, such that all actors portrayed the selected emotions, and further, these emotions could be distributed evenly on the quadrants of the valence-arousal scale (Russell, 1980; see Table 1), thereby balancing positive and negative emotions as well as high- and low-arousal emotions within the selected stimulus set. This resulted in a total of 120 stimuli used in our experiments. The five remaining emotions that were excluded from data collection were used as practice material to acquaint participants with the stimulus materials and the task.

The audio from all movie files was edited in Audacity (version 2.1.2; <http://audacityteam.org/>), to remove any audible noise or clipping from the audio recordings, and saved as 16-bit WAV-files. To do so, in most cases, the editing consisted of using the built-in ‘Noise Reduction’ effect to reduce background noise as much as possible without affecting the speech signal. In rare cases, the files contained clipping, which was removed by manually

Table 1.

The selected emotion categories used in the experiment. The emotions for the main experiment are distributed over the quadrants of the valence-arousal scale (Russell, 1980). The five additional emotions are used for the practice trials

| Arousal | Valence | |
|------------|--|------------|
| | Positive | Negative |
| High | Amusement | Fear |
| | Joy | Despair |
| | Pride | Anger |
| Low | Pleasure | Irritation |
| | Relief | Anxiety |
| | Interest | Sadness |
| Additional | Disgust Contempt Surprise Admiration Tenderness | |

adjusting the clipped regions of the waveform. Audio recordings were then root-mean-square (RMS)-equalized in intensity level, and re-merged with the corresponding video files (thereby replacing the old audio) using custom-made scripts.

2.4. *Experimental Setup*

Experiments were performed in a silent room, which was dark except for the illumination provided by the screen. Participants were seated in front of a computer screen at a viewing distance of 70 cm with their head in a chin and forehead rest to minimize head movements. Stimuli were displayed and manual responses were recorded using MATLAB (Version R2015b; The Mathworks, Inc., Natick, MA, USA), the Psychophysics Toolbox (Version 3; Brainard, 1997; Kleiner *et al.*, 2007; Pelli, 1997) and the Eyelink Toolbox (Cornelissen *et al.*, 2002) extensions of MATLAB. The stimuli were presented full-screen on a 24.5-inch monitor with a resolution of 1920×1080 pixels (43×24.8 degrees of visual angle). Average screen luminance was 38 cd/m^2 . Stimulus presentation was controlled by an Apple MacBook Pro (early 2015 model). Audio was produced by the internal soundcard of this computer and presented binaurally through Sennheiser HD 600 headphones (Sennheiser Electronic GmbH & Co. KG, Wedemark, Germany). The sound level was calibrated to be at a comfortable and audible level, at a long-term RMS average of 65 dB SPL.

To measure eye movements, an Eyelink 1000 Plus eye tracker, running software version 4.51 (SR Research Ltd., Ottawa, ON, Canada), was used. Gaze data were acquired at a sampling frequency of 500 Hz. The eye tracker was mounted on the desk right below the presentation screen. At the start of the experiment, the eye tracker was calibrated using its built-in nine-point calibration routine. Calibration was verified with the validation procedure in which the same nine points were shown again. The experiment was continued if the calibration accuracy was sufficient (average error of less than 0.5° and a maximum error of less than 1.0°). A drift check was performed both at the start of the experiment and after each break. If the drift was too large (i.e., more than 1.0°), the calibration procedure was repeated.

2.5. *Procedure*

In this study, behavioral and eye-tracking data were obtained to identify accuracy and gaze fixation of emotion identification with dynamic stimuli. In each trial, prior to each stimulus presentation, a central fixation cross appeared for a random duration between 500 and 1500 ms. The response screen followed each stimulus presentation after 100 ms and remained on screen until the participant made his or her response. The order of events in a typical trial is shown in Fig. 1.

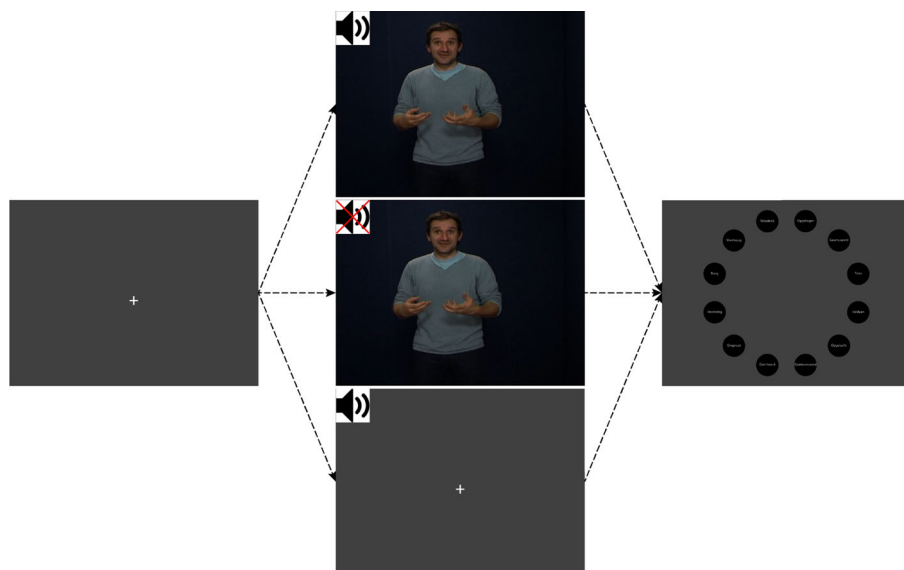


Figure 1. Schematic representation of the events in a single trial. Participants first were shown a fixation cross (left), followed by the stimulus, presented audiovisually (middle top), visually (middle), or aurally (middle bottom). After stimulus presentation, a response screen (right) with labels indicating the possible emotions appeared and remained on screen until the participant made a (forced) response. Emotion labels were in Dutch, from top right going clockwise they are: opgetogen (joy), geamuseerd (amusement), trots (pride), voldaan (pleasure), opgelucht (relief), geïnteresseerd (interest), geïrriteerd (irritation), ongerust (anxiety), verdrietig (sadness), bang (fear), wanhopig (despair), and woedend (anger).

Participants were asked to identify the emotion presented in one of three stimulus presentation modalities: audio-only (A-only), video-only (V-only), or audio and video combined (AV). They were asked to respond as accurately as possible in a forced-choice discrimination paradigm, by clicking on the label on the response screen corresponding with the identified emotion. Emotion labels were shown and explained before the experiment. Participants were further instructed to blink as little as possible during the trial and maintain careful attention to the stimuli.

In total, each participant was presented with all 120 stimuli (twelve emotions \times ten actors) in all three blocks: an A-only block, a V-only block, and an AV block. Block order was counterbalanced between participants. Stimulus order within each block was randomized. Participants were encouraged to take breaks both within and between blocks (breaks were possible after every 40 trials) to maintain concentration and prevent fatigue. Breaks were self-paced and the experiment continued upon the participant pressing the spacebar. Following each break, a drift correction was applied to the eye-tracking calibration.

Fifteen practice trials (five practice trials for each modality) preceded the experiment to familiarize participants with the task and stimulus material. In total, the experiment consisted of 375 trials, including the 15 practice trials, and took at most one hour to complete. Feedback on the given responses was provided during the practice trials only.

2.6. Analyses of Behavioral Data

To assess the presence of audiovisual integration, we tested whether performance for emotion identification differed for A-only, V-only, and AV stimulus presentation. We additionally employed a measure that quantifies the size of the effect from audiovisual integration, i.e., whether audiovisual integration is sub-additive (i.e., lower than expected based on the simultaneous and independent processing of both unisensory modalities), additive (i.e., equal to a summation of the auditory and visual evidence), or supra-additive. A supra-additive effect would be indicative of a gain in performance beyond what is gained by independently summing the information from both modalities (Crosse *et al.*, 2016; Stevenson *et al.*, 2014).

Accuracy scores for each emotion and modality were converted to unbiased hit rates (Wagner, 1993) prior to further analyses. Unbiased hit rates (H_u) were used to account for response biases. Unbiased hit rates were then arcsine-transformed to ensure normality and analyzed in R (version 3.6.0; R Foundation for Statistical Computing, Vienna, Austria — <https://cran.r-project.org>) with repeated-measures ANOVA (*aov_ez* from the *afex* package, version 0.25-1). For the ANOVA, arcsine-transformed H_u was the dependent variable, and modality (with three levels; A-only, V-only, and AV) and emotion (with 12 levels) the fixed-effects variables. Greenhouse–Geisser correction was performed in cases of a violation of the sphericity assumption. Effect sizes are reported as generalized eta-squared (*ges*). Pairwise comparisons were performed to test main effects (comparing different modalities) and interactions (the effect of modality for each emotion) using *lsmeans* from the *emmeans* package (version 1.4.1). For comparing differences between modalities, the Bonferroni correction was applied to make sure our conclusions were not based on a possibly too liberal adjustment. For comparing modality differences between emotions we used the False Discovery Rate (FDR) correction in order to ensure no effects were lost due to strict adjustments of *p*-values due to the many pairwise comparisons made.

For a quantitative assessment of the AV integration effect, we tested if the measured performance for AV exceeded the statistical facilitation produced by A + V. To quantify the predicted H_u for the independent summation of A and V we used the following equation (Crosse *et al.*, 2016; Stevenson *et al.*,

2014):

$$\hat{H}_u(AV) = H_u(A) + H_u(V) - H_u(A) \cdot H_u(V) \quad (1)$$

If the H_u for the AV modality exceeds the predicted H_u , as assessed by a paired t -test, this indicates A and V are integrated in a supra-additive manner (see, e.g., Calvert, 2001; Hughes *et al.*, 1994). Paired t -tests were only performed when at least the differences between AV and V-only and between AV and A-only were significant.

2.7. Analyses of Eye-Tracking Data

Fixations were extracted from the raw eye-tracking data using the built-in data-parsing algorithm of the Eyelink eye tracker. We performed an AOI-based analysis for fixations made during stimulus presentation (only for the AV and V-only modalities as for the A-only modality there is no visual stimulus aside from a fixation cross). Trials with blinks longer than 300 ms during stimulus presentation were discarded. The analysis was restricted to fixations made between 200 ms and 1000 ms after stimulus onset. The first 200 ms were discarded because this is the time needed to plan and execute the first eye movement. No data after 1000 ms were taken into account to limit data analysis to the duration of the shortest movie at 1000 ms.

In the videos, the eyes (left and right), nose, mouth, and hands (left and right) of the speaker were chosen as AOIs. Because the stimuli are dynamic, we created dynamic AOIs. Coordinates of the AOI positions for each movie and each frame were extracted using Adobe® After Effects® CC (Version 15.1.1; Adobe Inc., San Jose, CA, USA). For the face AOIs, these coordinates were obtained by placing an ellipsoid mask on the face area and applying a tracker using the ‘Face Tracking (Detailed Features)’ method, which automatically tracks many features of the face (see Fig. 2 for an example frame with AOIs drawn in). Face track points were visually inspected and manually edited (i.e., moved into the correct place) whenever the tracking software failed to correctly track them.

Coordinates of all obtained face track points for each movie frame were stored in a textfile and used to create rectangular AOIs. For the eyes’ AOI we used the coordinates of the following face track points: ‘Right/Left Eyebrow Outer’ for the x -position of the lateral corner, ‘Right/Left Eyebrow Inner’ for the x -position of the medial corner, ‘Right/Left Eyebrow Middle’ for the top, and the middle between the y -positions of ‘Left Pupil’ and ‘Nose tip’ for the bottom, indicating the eye–nose border. Two individual AOIs were created for the left and right eye, which were later merged for analyses. For the nose AOI: the eye–nose border as the top, the nose–mouth border (middle between the y -positions of ‘Right Nostril’ and ‘Mouth Top’), the x -position of ‘Right Nostril’ for the left corner, and the x -position of ‘Left Nostril’ for the right

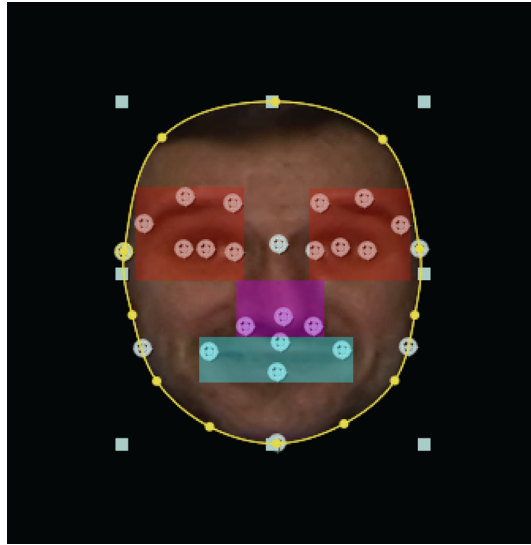


Figure 2. Face tracking in Adobe After Effects CC. The yellow line is the ellipsoid mask after automatic alignment to the contours of the face. Each circled cross is a face track point. The colored rectangles indicate the locations of the different areas of interest (AOIs); the red rectangles denote the right- and left-eye AOIs, the purple rectangle shows the nose AOI, and the blue rectangle specifies the mouth AOI.

corner. For the mouth AOI: the x -position of ‘Mouth Right’ for the left corner, the x -position of ‘Mouth Left’ for the right corner, the nose–mouth border for the top, and the y -position of ‘Mouth Bottom’ for the bottom. Each AOI was expanded by 10 pixels on each side (20 pixels across the horizontal and vertical axes), except at the eye–nose and nose–mouth borders. Overlap between AOIs was avoided. The actual size of each AOI varied across actors and frames e.g. due to some actors being closer to the camera.

For the hand AOIs, the ‘Track Motion’ method was used, in which a single tracker point (per hand) was used to track position. The tracker point was placed approximately in the center of the hand. The track point was manually edited whenever the tracking software failed to correctly track it. This happened often due to the complex movements the hands made in most movies. Figure 3 shows example frames from one movie. After extracting the coordinates, a sphere with a radius of 75 pixels was used to create the AOI.

Then, for each fixation datapoint we checked whether the fixation was on one of the AOIs (with the coordinates from the movie frame cooccurring with the time of the fixation), leading to one binary vector for each AOI with the same length as the length of the fixation data. These vectors were then averaged per trial, giving a mean fixation proportion on each AOI for each trial. Lastly, the means were arcsine-transformed. A mixed linear regression was



Figure 3. Hand tracking using Adobe After Effects CC. In both images, the attach point is at the center (from which the coordinate is extracted), the inner box is the feature region (i.e., what the tracked region looks like), and the outer box is the search region of the tracker (i.e., the region in which the tracker will search for the feature region). Additionally, the tracked points in previous frames can be seen. As can be seen in the left image, tracking works well early in the movie. As the hand starts to change shape later in the movie, however, the tracker errs. This can be seen on the right image where the tracker loses the hand from sight and tracks the arm and background instead.

performed in *R* (using *lmer* from the *lme4* package, version 1.1-21) on correct trials only, as we were most interested in examining whether changes in viewing behavior due to changes in modality availability were adaptive, leading to good performance. In line with the analyses of unbiased hit rates, the model included modality, emotion, and AOI as fixed effects, which were allowed to interact with each other. Random intercepts were included for participant and movie and a random slope for modality was included for both participant and movie if the model still converged (otherwise, only a random slope for modality was included for participants). Overall significance of the main effects and interactions was assessed using the *Anova* function from the *car* package (version 3.0-3). Pairwise comparisons were performed to test whether fixation proportions on different AOIs differed for different modalities and emotions using *lsmeans*. As before, for comparing differences between modalities, the Bonferroni correction was applied while for comparing differences between emotions we used the FDR correction.

Lastly, we ran a second model to test whether fatigue or boredom, which may have occurred due to the lengthy duration of the experiment, had an effect on fixation patterns, by adding experimental block to the model. There was no significant effect of block on fixation patterns ($\chi^2(1) = 1.79$, $p = 0.18$), ruling out the additional effect from potential boredom and fatigue.

3. Results

Participants identified dynamic emotional expressions presented in movies while their eye movements were recorded. The objective of this study was to see if emotions are processed similarly whether conveyed in a unimodal (A-only, V-only) or multimodal (AV) manner, as measured by performance levels and fixation patterns. To achieve this objective, here we present analyses of accuracy and gaze differences for different modalities and emotions. Accuracy and fixation data for individual participants can be found in Supplementary Figs S1, S2, S3, and S4. Confusion matrices for each modality can be found in Supplementary Fig. S5.

3.1. Accuracy Across Modalities and Emotions

Accuracy scores in unbiased hit rate (H_u) and averaged over all participants and testing blocks is shown in Fig. 4. On average, participants performed the task with a mean accuracy of 0.37, well above the chance level of 0.083.

A visual inspection of Fig. 4 suggests performance is lowest for the A-only modality and highest for the AV modality. This was also confirmed by the ANOVA, which had H_u as the dependent variable, and modality and emotion as independent variables. The model showed an overall effect of modality ($F_{2,40} = 42.7$, $p < 0.001$, $ges = 0.18$), a main effect of emotion ($F_{11,220} = 53.1$, $p < 0.001$, $ges = 0.48$), and a significant interaction between

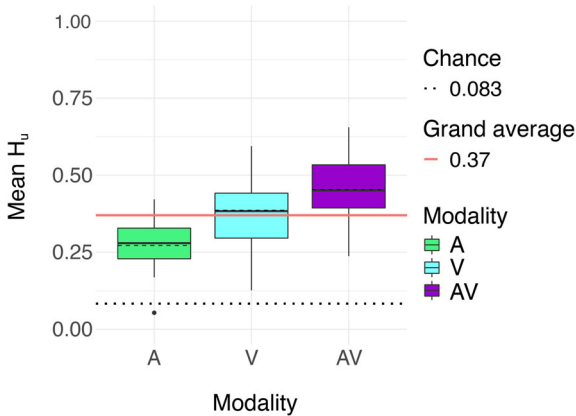


Figure 4. Task performance for each modality, shown as unbiased hit rates and averaged across all participants and blocks. Each box shows the data between the first and third quartiles. The horizontal black solid line in each box denotes the median while the horizontal black dashed line in each box denotes the mean. The whiskers extend to the lowest/highest value still within $1.5 \times$ interquartile range of the first/third quartile. Dots are outliers. The red line indicates the grand average performance (0.37). The black dotted horizontal line indicates chance level performance (0.083).

modality and emotion ($F_{22,440} = 5.2$, $p < 0.001$, $ges = 0.07$). Bonferroni-adjusted pairwise comparisons showed performance was significantly different between all modalities (A-only–AV: $t_{40} = -9.13$, $p < 0.001$; A-only–V-only: $t_{40} = -5.80$, $p < 0.001$; V-only–AV: $t_{40} = 3.34$, $p = 0.006$). Therefore, performance was lowest for A-only (mean accuracy = 45%), intermediate for V-only (mean accuracy = 62%), and highest for AV (mean accuracy = 70%), with all differences between modalities being significant.

Further inspection of the modality-by-emotion interaction showed that, in general, performance was lowest for A-only, intermediate for V-only, and highest for AV, but this was not true for all emotions. In fact, for most emotions (except for Pleasure, Relief and Anxiety), there was no significant difference in performance between V-only and AV. In addition, for some negative valence emotions (Fear and Anger), none of the comparisons between modality pairs produced a significant difference. Lastly, for Pleasure, Relief, and Despair the difference between V-only and A-only was not significant. The complete list of all comparisons is given in Table 2 and further visualized in Fig. 5.

Table 2.

Contrasts for the modality-by-emotion interaction showing the model estimate differences, with the False Discovery Rate (FDR)-adjusted p -values in parentheses. A positive contrast means performance in the first condition was better than in the second of the comparison (and v.v.). Significant differences are indicated in bold

| | Contrast | | |
|---------------------------------------|---------------------|-------------------------|-------------------------|
| | AV–V | AV–A | V–A |
| <i>Positive valence, high arousal</i> | | | |
| Amusement | 0.09 (0.09) | 0.23 (<0.001) | 0.15 (0.005) |
| Joy | 0.09 (0.07) | 0.37 (<0.001) | 0.28 (<0.001) |
| Pride | 0.09 (0.09) | 0.48 (<0.001) | 0.40 (<0.001) |
| <i>Positive valence, low arousal</i> | | | |
| Pleasure | 0.15 (0.005) | 0.23 (<0.001) | 0.08 (0.10) |
| Relief | 0.12 (0.03) | 0.19 (<0.001) | 0.07 (0.16) |
| Interest | 0.08 (0.12) | 0.37 (<0.001) | 0.29 (<0.001) |
| <i>Negative valence, high arousal</i> | | | |
| Fear | 0.09 (0.20) | 0.08 (0.20) | −0.02 (0.74) |
| Despair | 0.05 (0.37) | 0.13 (0.02) | 0.09 (0.12) |
| Anger | 0.04 (0.72) | 0.04 (0.72) | 0.008 (0.87) |
| <i>Negative valence, low arousal</i> | | | |
| Irritation | 0.04 (0.42) | 0.21 (<0.001) | 0.17 (0.001) |
| Anxiety | 0.12 (0.02) | 0.25 (<0.001) | 0.13 (0.01) |
| Sadness | 0.05 (0.28) | 0.17 (0.003) | 0.11 (0.04) |

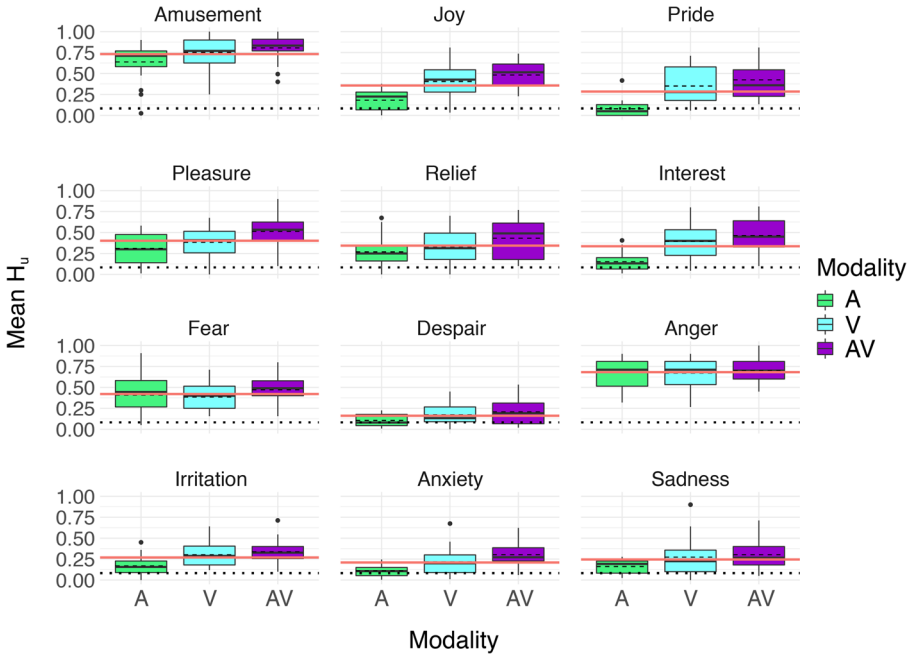


Figure 5. Task performance for each modality, similar to Fig. 4, but shown for each emotion. The red line in each panel indicates the average performance for that particular emotion. The black dotted horizontal line indicates chance level performance (0.083).

While AV performance was significantly higher than both A-only and V-only performance, indicating that AV integration took place, the AV integration effect was sub-additive as performance for AV was significantly lower than predicted on the basis of additivity ($t_{20} = -3.06$, $p = 0.006$; $\hat{H}_u(AV)$: 0.52 ± 0.12 , $H_u(AV)$: 0.45 ± 0.10). Considering individual emotions, only for anxiety, pleasure, and relief performance differed between both AV and V-only and between AV and A-only, and thus, only for these emotions it was further tested whether AV performance was supra-additive. AV performance was not significantly different from the predicted additive performance for Anxiety [$t_{20} = 0.006$, $p = 0.99$; $\hat{H}_u(AV)$: 0.30 ± 0.16 , $H_u(AV)$: 0.30 ± 0.14], for Pleasure [$t_{20} = -1.33$, $p = 0.20$; $\hat{H}_u(AV)$: 0.56 ± 0.05 , $H_u(AV)$: 0.51 ± 0.05] or for Relief [$t_{20} = -1.54$, $p = 0.14$; $\hat{H}_u(AV)$: 0.50 ± 0.05 , $H_u(AV)$: 0.43 ± 0.05], indicating that the AV integration effect was additive in all three emotions.

3.2. Fixation Patterns Across Modalities and Emotions

Fixation proportions, averaged over all stimuli and participants, are shown for all AOIs in Fig. 6. Figure 6a shows how the fixation proportions change over

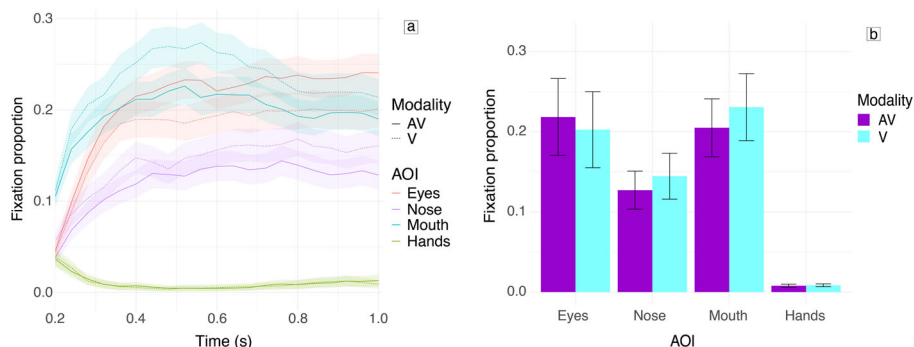


Figure 6. Fixation proportions for correct trials on all areas of interest (AOIs) (face, i.e., eyes, nose, mouth; and hands), across the analyzed time course (a) and averaged over the analyzed time course (b), both averaged over all stimuli and participants. Shaded areas around each line (a) and error bars (b) denote the standard error of the mean (SEM).

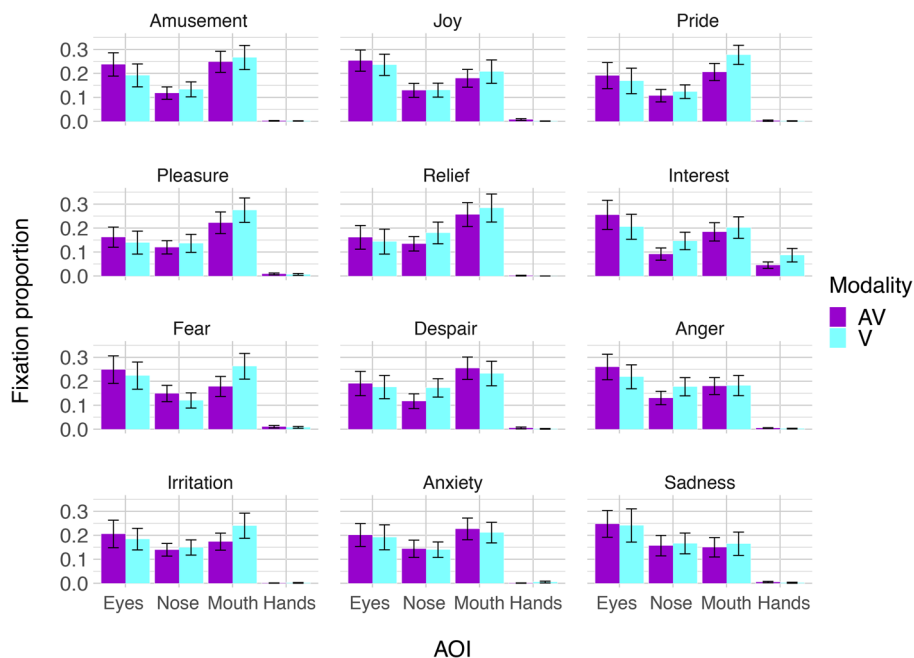


Figure 7. Fixation proportions for correct trials on all areas of interest (AOIs) averaged over the analyzed time course, averaged over participants. The panels show fixation proportions for different emotions. The error bars denote the standard error of the mean (SEM). See Supplementary Fig. S7 for fixation proportions across the analyzed time course for all emotions.

the analyzed time course, while Fig. 6b shows the fixation proportions averaged over the trial. Figure 6 suggests differences in viewing behavior between modalities.

Table 3.

Contrasts for the emotion-by-AOI (area of interest) interaction. The table shows the model estimate difference for the contrasts, with False Discovery Rate (FDR)-adjusted *p*-values in parentheses. A positive contrast means the first AOI was fixated more than the second of the comparison (and v.v.). Significant differences are indicated in bold

| | Contrast | | |
|---------------------------------------|--------------------------|-------------------------|-------------------------|
| | Eyes–mouth | Eyes–nose | Mouth–nose |
| <i>Positive valence, high arousal</i> | | | |
| Amusement | −0.09 (<0.001) | 0.10 (<0.001) | 0.20 (<0.001) |
| Joy | 0.04 (0.13) | 0.12 (<0.001) | 0.07 (0.01) |
| Pride | −0.18 (<0.001) | 0.03 (0.32) | 0.21 (<0.001) |
| <i>Positive valence, low arousal</i> | | | |
| Pleasure | −0.17 (<0.001) | 0.03 (0.22) | 0.20 (<0.001) |
| Relief | −0.20 (<0.001) | −0.03 (0.25) | 0.17 (<0.001) |
| Interest | 0.02 (0.55) | 0.12 (<0.001) | 0.11 (<0.001) |
| <i>Negative valence, high arousal</i> | | | |
| Fear | −0.04 (0.17) | 0.09 (0.003) | 0.12 (<0.001) |
| Despair | −0.12 (0.001) | 0.01 (0.74) | 0.14 (<0.001) |
| Anger | 0.07 (0.01) | 0.11 (<0.001) | 0.04 (0.09) |
| <i>Negative valence, low arousal</i> | | | |
| Irritation | −0.07 (0.048) | 0.04 (0.13) | 0.10 (<0.001) |
| Anxiety | −0.02 (0.42) | 0.09 (0.001) | 0.11 (<0.001) |
| Sadness | 0.005 (0.89) | 0.05 (0.19) | 0.05 (0.20) |

The regression model confirmed this. The model included modality, emotion, and AOI as fixed effects (and their interactions). A random intercept was included for both participant and movie, and a random slope for modality for participants. There was a main effect of AOI ($\chi^2_3 = 3314.1$, $p < 0.001$), a significant interaction between modality and AOI ($\chi^2_3 = 34.2$, $p < 0.001$), and a significant interaction between emotion and AOI ($\chi^2_{33} = 184.2$, $p < 0.001$). Significant main effects and interactions were followed up with *post-hoc* testing, as further detailed below.

Bonferroni-corrected pairwise comparisons showed that, in general, the mouth was fixated more often than the eyes (z -ratio = -7.4 , $p < 0.001$) and nose (z -ratio = -14.9 , $p < 0.001$), the eyes were fixated more often than the nose (z -ratio = 7.5 , $p < 0.001$) and all face AOIs were fixated more than the hands (all $p < 0.001$). Additionally, participants fixated more on the mouth (z -ratio = -3.1 , $p = 0.002$) and nose (z -ratio = -2.3 , $p = 0.02$) and less on the eyes (z -ratio = 3.08 , $p = 0.02$) in the V-only modality compared to the AV modality. There was no difference in fixation proportions on the hands

(z -ratio = -0.1 , $p = 0.92$). Lastly, the results of the emotion by AOI interaction can be found in Table 3 and are visualized in Fig. 7. Because fixations on the hands were so scarce, only comparisons between the face AOIs are shown. In general, the same pattern can be seen for each emotion; most fixations are on the mouth, then the eyes, then the nose, and lastly on the hands (not shown in the table). There is only one exception to this: participants fixated on the eyes more often than on the mouth for Anger (z -ratio = 2.6 , $p = 0.01$).

4. Discussion

The present study examined whether observers flexibly adapt their viewing behavior to the presence of audio during the recognition of videos of emotional expressions. We measured audiovisual integration by examining participants' eye movements and emotion identification performance while they viewed video recordings of dynamic emotion expressions with or without the corresponding audio.

Our main finding is that there is evidence for integration of auditory and visual information when observers recognize emotions, evident from adapted viewing behavior in response to the changes in modality availability. This adaptation in viewing behavior was present even though there was no evidence for supra-additive integration, as derived from task performance. Moreover, adding audio to the video signal changed observers' viewing behavior, even when the addition of audio did not result in any improvement in identification performance. This implies that auditory signals are used in emotion perception for communication when they are present, and when they are not present people cope well by extracting auditory emotional cues visually, for example by observing mouth movements. Together, our results suggest observers flexibly shape their perceptual strategies based on the audiovisual information available.

4.1. Sub-Additivity of Audio and Visual Information During Emotion Recognition With Multimodal Stimuli

Firstly, we asked whether our participants would integrate auditory and visual information when performing the emotion identification task or whether visual information alone would mostly be sufficient. When averaged over emotions, task performance was significantly higher in the AV modality than in either of the unimodal modalities, indicating audiovisual integration took place. These findings are in line with studies that compared only two basic emotion categories and used static visual face stimuli combined either with a spoken word (Massaro and Egan, 1996) or a spoken neutral sentence (de Gelder and Vroomen, 2000), and with a study that compared all six basic emotions and used short audiovisual videos (Paulmann and Pell, 2011).

These previous studies combined show that emotion recognition improves when information from more than one modality is available, provided the multimodal information is congruent (as was the case in our study). However, we found that the audiovisual integration effect was not particularly strong; performance in the AV modality did not exceed performance gain associated with statistical facilitation as is predicted by ‘supra-additivity’ (see, e.g., Calvert, 2001; Hughes *et al.*, 1994) and was even sub-additive. Our data thus show that audiovisual integration took place, but yet led to a smaller gain in performance than would occur if the auditory and visual evidence would be summed.

However, while some researchers suggested supra-additivity to be the hallmark of multisensory integration (see Stein and Meredith, 1993, for a review), originating from the pioneering single-cell electrophysiology of the cat superior colliculus (by, a.o., Meredith and Stein, 1983), others have argued that many multisensory behaviors do not rely on supra-additivity when the presented stimuli are not close to detection threshold (Angelaki *et al.*, 2009; Stanford and Stein, 2007). Since we used stimuli with very rich visual and auditory cues, and also in ideal listening and viewing conditions with no distortions, performance in unimodal conditions was already relatively high. As the inverse effectiveness rule states: the strength of multimodal integration is inversely related to the effectiveness of the unimodal stimuli (Stein and Meredith, 1993). Therefore, it remains a possibility that the AV integration effect was sub-additive for the specific study conducted here; however, if unimodal performance were lower (i.e., closer to chance level), for example due to decreased auditory and visual signals, the integration effect could be stronger and perhaps become supra-additive.

4.2. *Multimodal Viewing Does not Always Facilitate Emotion Recognition*

In addition to the overall effect of an improvement in performance in the AV modality, we analyzed task performance per emotion. We expected more visual dominance for the basic emotion categories included in the used stimulus set (joy, sadness, fear, and anger) and more integration for the fine-grained emotions (e.g., irritation, despair). Our behavioral data indicated that audiovisual integration — i.e., performance in AV being different from performance in the V-only and A-only conditions — did not occur for all emotions. We found that for many emotions, performance did not differ between AV and V-only, while performance for A-only was mostly lower than in both AV and V-only. Therefore, our behavioral findings would suggest decisions were made primarily on the basis of the visual information and contribution from auditory information was limited. Unlike our expectation, visual dominance was present not only in the basic emotion categories we included, but also in many of the fine-grained emotion categories. The only exceptions were three low-arousal emotions: pleasure, relief, and anxiety (see Table 2). Hence, at least

for some fine-grained emotions, combining auditory and visual information increased performance. However, AV performance was never supra-additive for the included emotions.

Our data show a similar pattern to the validation by Bänziger *et al.* (2012) of the stimulus set that was used in the present study. Although these authors did not make all the comparisons we made (in Table 2), their Table 5 (core set rating, 12 repeated emotions only) similarly hints toward visual dominance for many emotion categories investigated. Audiovisual integration, when measured by task performance, thus seems to be the exception, rather than the rule. This is again likely related to inverse effectiveness; performance in the video-only modality was generally higher than performance in the audio-only modality, and hence, the visual dominance observed here could be due to differences in information reliability of this specific stimulus set. This idea is strengthened by findings from Collignon *et al.* (2008), who found visual dominance when audiovisual emotion stimuli were presented without any noise, but evidence for audiovisual integration when they added noise in the visual modality, thus decreasing the reliability of the visual information.

Unreliability of the audio information for emotion recognition may be inherent to this modality. There may be less clear prototypical expressions of specific emotions in the audio (e.g., laughter, crying) than there are in the video (e.g., smile, frown). This could lead to lower reliability for the auditory compared to the visual modality and consequently result in visual dominance. This could explain why visual dominance is commonly found in experiments employing dynamic face/voice stimuli (Bänziger *et al.*, 2009; Takagi *et al.*, 2015; Wallbott and Scherer, 1986) or dynamic body/voice stimuli (Jessen *et al.*, 2012). Alternatively, low reliability of auditory information may be inherent to the stimulus material, for example because the use of non-words makes the auditory cues less salient and thus less reliable, which may explain the discrepancy between our data and some other studies (de Gelder and Vroomen, 2000; Massaro and Egan, 1996) that did find behavioral evidence for an effect of adding audio to a visual stimulus. However, it should be noted that these studies used a static visual stimulus, which may have decreased its salience and/or reliability and consequently increased the utilization of auditory information. Lastly, some methodological decisions may have affected our participants' ability to integrate the audio with the video; the intensity level of all audio recordings was RMS-equalized, which can take away some of the loudness cues related to emotions that occur in everyday life (e.g., a sad expression is generally quieter than an angry expression). Additionally, the audio was presented over headphones and not *via* a speaker, which could lead to some spatial disparity between the auditory and visual cues. Although, in principle, audiovisual temporal synchrony should be a stronger cue than

the spatial co-location, we cannot exclude if participants experienced spatial disparity and therefore focused less on the auditory cues.

It should be noted that audiovisual integration and visual dominance are not necessarily mutually exclusive. While visual information alone might be sufficient for recognizing emotions, the addition of auditory information could still provide more evidence and allow for faster emotion identification, while accuracy rate remains the same. In complex real-life situations with many interfering audiovisual signals, such added evidence may play a more important role than in the ideal conditions of lab testing. Investigating response times or other measures of cognitive processing could therefore be a beneficial addition. However, for our study, the stimuli used were relatively long and participants were only able to respond after the stimulus ended, and therefore, there is a strong possibility that participants already decided on their answer before the stimulus ended. All these factors, if not controlled for, could make response times unreliable. There may be another method to explore this option, namely, in situations where audiovisual integration may become more important, such as under compromised conditions (e.g., noisy audio or blurred video). A decrease in the reliability of the information in one modality could increase the need for integration, possibly leading to supra-additive integration effects on performance. Furthermore, this could clarify whether there is more visual or more auditory dominance, or whether both channels of information equally contribute to an integrated percept.

4.3. There Is No General Tendency to Focus on the Eyes When Recognizing Emotions

In contradiction to popular belief, we did not find a general tendency to fixate on the eyes. There are indications that especially for the recognition of more complex emotions, the eyes are most informative (Nummenmaa, 1964). This view is supported by an ERP study that indicated the eyes as the starting point of emotion recognition. They found that the integration of facial emotional information starts at the eyes, then moves downward across the face, and stops when enough information is integrated to classify an expression (Schyns *et al.*, 2007). Additionally, it has been shown in both healthy observers as well as in observers with Autism Spectrum Disorders that increased gaze duration to the eyes is correlated with higher emotion recognition performance (Bal *et al.*, 2010; Lischke *et al.*, 2012). On the other hand, some eye-tracking studies have indicated that (Western Caucasian) observers distribute fixations evenly across the face (Jack *et al.*, 2009), whereas other studies have shown observers mostly fixate the areas that are diagnostic for specific emotions (e.g., more fixations on the mouth for happy images and more fixations on the eyes for angry images; Eisenbarth and Alpers, 2011). Lastly, there is evidence that fixation patterns are perhaps not only specific for different emotions, but also

shift when a stimulus is dynamic. Blais and colleagues found that observers more or less equally sampled the eyes and mouth when stimuli were static, but fixated mostly on the center of the face when stimuli were dynamic (Blais *et al.*, 2017).

Be that as it may, none of these studies used audiovisual stimuli as was done here and the use of audiovisual stimuli seems to greatly impact where an observer will look. Here, we did not find a general tendency to fixate on the eyes, nor on the center of the face. Additionally, as can be seen from Fig. 7, in line with previous studies our participants fixated mostly on the eyes for Anger stimuli (Calder *et al.*, 2000; Smith *et al.*, 2005), but in contradiction to previous studies they did not mostly fixate on the mouth for Joy stimuli (the close equivalent to the basic emotion happiness used in other studies), but instead sampled the eyes and mouth equally often. For the majority of other emotions, the mouth was fixated most often, followed by the eyes and nose.

4.4. Gaze Behavior During Emotion Perception for Communication Does not Simply Reflect Visual Saliency

Contrary to our behavioral data, our fixation data suggest clear usage of audio information and thus indicate there is at least an interaction between auditory and visual information. When averaged over emotions, observers viewed the mouth less and the eyes more in the AV modality compared to the V-only modality; there was also a decrease in fixations on the nose. These findings suggest observers flexibly adapt viewing behavior to fixate regions that they feel would maximize performance depending on whether audio is present or absent. There are several studies that give indications on why the increased fixations on the nose and mouth in the V-only modality might be beneficial for performance when audio is lacking.

First, the nose has been proposed to be an optimal fixation landmark for global face perception, at least for static facial images (Hsiao and Cottrell, 2008; Peterson and Eckstein, 2012). After all, from this vantage point, it is possible to both rapidly direct the gaze to either the eyes or the mouth, as these regions are more or less equidistant from the nose. Moreover, it may also be possible to simultaneously gather (crude) visual information from both the eyes and the mouth using lower resolution peripheral vision (Posner, 1980). Additionally, biological motion can be processed well in the periphery (Thompson *et al.*, 2007), making fixating on the nose a good strategy if one wishes to retrieve dynamic information from both the eyes and the mouth. It is a fair assumption that in the V-only modality, participants tried to gather as much visual information as possible to compensate for the lack of audio signal, and therefore fixated more on the nose in order to also access visual information from both the eyes and mouth.

Second, increasing the proportion of fixations on the mouth could then serve to gather more fine-grained visual emotional information. Such an increase in fixations on the mouth is not commonly reported in the literature and whether or not it is found seems to depend on the task participants performed. For example, while Lansing and McConkie (2003) also found an increase in mouth fixations in the V-only modality, Võ and colleagues (2012) found a decrease in mouth fixations when sound was muted. However, while their stimuli were similar, their experimental tasks were rather different: both featured videos of people speaking (only face, neck, and shoulders visible) but in the study by Võ *et al.* participants had to rate the likeability of the video, while in the Lansing and McConkie study participants performed a speech identification task. These and our own findings indicate eye gaze reveals how the perceptual strategies flexibly adapt to the available information and the nature of the specific task.

The modality and task dependency of eye gaze indicate that gaze is not simply dictated by visual saliency. If it were, one would expect to always find most fixations on the mouth in dynamic face stimuli. Mouth movements are quite large and thus more salient compared to those of other facial features. Moreover, an increase in fixations on the eyes in the AV compared to the V-only condition is not expected either, as the visual stimulus did not change. Our findings, as those of others (Lansing and McConkie, 2003; Võ *et al.*, 2012), therefore indicate that gaze is guided by an information-seeking process. Moreover, that V-only performance exceeded A-only performance for most emotions, suggests the visual information provided by the mouth can be a vital substitute for the missing auditory information.

4.5. Perceptual Strategies Suggest Auditory Rather Than Visual Dominance

While task performance could be taken to indicate visual dominance for many emotions, there is no compelling evidence for visual dominance to be found in the viewing behavior. Although for many emotions, no significant difference in accuracy was found between AV and V-only, there was a clear effect on viewing behavior when adding audio to the video. Our data suggest that viewing behavior, and by extension the manner in which the task is performed, adapts as a function of both the available information and by the degree to which the information is task-relevant.

That participants' viewing behavior changed depending on the presence of audio, is indicative that gaze is not only guided by the visual information, which remained the same in the AV and V-only modalities, but also by the presence of auditory information. One might therefore even argue for auditory dominance instead of visual dominance for emotion perception. Evidently, when there is audio, one uses it and adapts viewing behavior accordingly, perhaps because some areas do not have to be fixated anymore to obtain the

information present in those areas. As an example, the movements of the mouth can provide cues of the expressed emotion, but the audio (produced by those same mouth movements) likely provides the same cues (as well as some unique information), resulting in redundancy in information across the two modalities. It is therefore no longer necessary to look at the mouth when audio is present and one is free to look for cues of the expressed emotion elsewhere, the eyes perhaps. That this adaptation does not always result in improved task performance could be because there simply is not more information in the visuals, wherever one looks. Our present study cannot yet fully confirm or reject whether emotion perception is guided preferentially and perhaps even compulsory by auditory information. To test this idea, one would have to see changes in behavior, be it viewing behavior or otherwise, in the presence of any audio — e.g., noise — compared to the absence of audio. Regardless, our data show that even when audiovisual integration is not apparent from the task performance, from the adaptations in viewing behavior it is clear the two modalities are integrated and shape the decision-making process. This study thus also underlines the need for measuring more than just task performance if one wishes to draw conclusions on audiovisual integration in emotion perception.

4.6. Limitations and Future Directions

Due to the many comparisons made in Tables 2 and 3 and the corrections therefore applied to the significance values, the comparisons might be underpowered. Future studies can be designed based on the knowledge produced in this study, where a subset of stimuli or conditions could be selected, producing fewer comparisons, or alternatively use a larger sample size, and better statistical power. Additionally, future studies should explore the integration process further by not only manipulating modality availability, but also manipulating information availability within modalities, for example by blurring (parts of) the image or using speech-shaped noise instead of actual emotional speech. Using stimuli specifically designed for it, measuring response times could also be a good addition, to further explore potential AV integration effects, in addition to accuracy performance. Lastly, though it would decrease the ecological validity of the stimuli, future studies could consider the use of (dynamic) incongruent audiovisual stimuli, possibly with differing reliabilities of the audio and video, to explore whether a continuum from visual to auditory dominance exists.

Our fixation data suggest that while the majority of fixations made were directed to our AOIs, a large part of the fixations were elsewhere on the screen. It can be inferred from Figs 6 and 7 that the fixations captured for each condition add up to roughly half of fixations made, although there are quite large individual differences (see Supplementary Figs S3 and S4). This could indicate

our participants either had an interest also for other areas of the screen, which we would have seen as a clustering of these outside AOI fixations on specific regions, such as the abdomen of the actor, or decided to browse around the screen more, which would be evident by fixations dispersed over the screen. However, an inspection of this with heat maps (see Supplementary Fig. S6 for fixation heat maps for all modalities) showed that actually most fixations were indeed directed toward the face, with only a minority of the fixations directed elsewhere, mainly on the body and toward the hands. It therefore seems more likely that participants relatively often looked just outside the AOIs. Additionally, it is peculiar that only few of the fixations made were directed toward the hands of the actors, despite our expectation that observers would use the information that can be gathered from hand gestures. Speculating, it is very well possible that observers need not fixate on the hands in order to retrieve the information they convey; viewing hand movements with peripheral vision might give enough information to recognize the emotion that is being expressed by the gestures. Future studies could test these hypotheses for example by removing the face, forcing observers to use other information.

We analyzed fixation data over an 800-ms time window for all stimuli (200 ms after start until 1000 ms after start, based on the length of the shortest video clip). Because of this, some gaze data was discarded. We chose not to use the full movie as participants may have decided which emotion was being expressed before the end of the movie clip (which is more likely to occur in long movies) and their gaze data after their decision might therefore reflect task-irrelevant viewing behavior. Nevertheless, we find that the pattern of results does not change if we take the full movie into account (see Fig. S8 for a comparison of average fixation proportions for the full movie and the used time window), confirming that the choice to use a 800-ms time window was an appropriate one.

It should be noted that some noise was present in the audio of the original stimulus materials. While careful consideration had been taken to remove this noise from the original stimulus materials, some noise may have been left which could have made the audio less reliable and may have biased performance to visual dominance. However, since our fixation data argue against visual dominance, it seems unlikely that any potentially remaining noise after pre-processing the audio substantially affected task performance.

Finally, it can be argued that the visual information in the stimuli contained two distinct cues for emotion: facial expressions and body expressions. Since this is not the case in the auditory modality, one could say that in the AV modality, participants had access to three emotion cues (face, body, and voice), in the V-only modality to two emotion cues (face and body), but in the A-only modality to only one emotion cue (voice). Following this line of reasoning,

it is thus not surprising that V-only performance was much higher than A-only performance, and that AV and V-only performances did not differ. Future studies should explore this further, for example by comparing performance for face + voice, body + voice, and face + body + voice conditions. In this example, observers always have access to two modalities, but the number of cues — and possibly also the quality of the cues — in the visual modality changes.

4.7. Conclusions

For the perception of emotions, observers generally utilize multiple sources of information when these are available. This was not evident from our behavioral measure of task performance as for many emotions performance on the multimodal task could be quite reliably predicted from performance on the visual task. However, viewing behavior did change based on information source availability even in the absence of a difference in performance. It can therefore be concluded that people change their perceptual strategies depending on the available information in an attempt to maximize performance. Drawing conclusions about integration of auditory and visual information thus is not only defined by the outcome (i.e., task performance), but also by the process (which can be studied with eye tracking). This study, with the use of dynamic multimodal emotion expressions, has taken a small step toward studying the perception of emotions in an ecologically more valid setting than with simpler materials. Further, it highlights the need for using multiple measures of emotion recognition if one wishes to deduce a comprehensive profile of audiovisual integration in emotion perception.

Acknowledgements

We gratefully acknowledge Tanja Bänziger, Marcello Mortillaro, and Klaus R. Scherer for permission for using the GEMEP core set and for publishing sample images from the core set. We also thank Birte Gestefeld, Alessandro Grillini, Rijul Soans, Paolo Toffanin, and Anita Wagner for their help with programming and data analysis. The first author was supported by a BCN-BRAIN grant from The Graduate School of Medical Sciences (GSMS), University of Groningen, the Netherlands. This project was supported by the following foundation: the Landelijke Stichting voor Blinden en Slechtzienden that contributed through UitZicht (Grant number: Uitzicht 2014 – 28 – Pilot). The funding organizations had no role in the design or conduct of this research.

Supplementary Material

Supplementary material is available online at:
<https://doi.org/10.6084/m9.figshare.12416441>

The datasets generated for this study can be found in the DataverseNL repository; <https://hdl.handle.net/10411/XSQS2T>. All data are publicly available.

References

- Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration, *Curr. Biol.* **14**, 257–262. DOI:10.1016/j.cub.2004.01.029.
- Angelaki, D. E., Gu, Y. and DeAngelis, G. C. (2009). Multisensory integration, *Curr. Opin. Neurobiol.* **19**, 452–458. DOI:10.1016/j.conb.2009.06.008.
- Bach, M. (1996). The Freiburg visual acuity test — automatic measurement of visual acuity, *Optom. Vis. Sci.* **73**, 49–53.
- Bach, M. (2006). The Freiburg visual acuity test — variability unchanged by post-hoc reanalysis, *Graefes Arch. Clin. Exp. Ophthalmol.* **245**, 965–971.
- Bal, E., Harden, E., Lamb, D., Van Hecke, A. V., Denver, J. W. and Porges, S. W. (2010). Emotion recognition in children with autism spectrum disorders: relations to eye gaze and autonomic state, *J. Autism Dev. Disord.* **40**, 358–370. DOI:10.1007/s10803-009-0884-3.
- Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression, *J. Pers. Soc. Psychol.* **70**, 614–636. DOI:10.1037/0022-3514.70.3.614.
- Bänziger, T., Grandjean, D. and Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT), *Emotion* **9**, 691–704. DOI:10.1037/a0017088.
- Bänziger, T., Mortillaro, M. and Scherer, K. R. (2012). Introducing the Geneva multimodal expression corpus for experimental research on emotion perception, *Emotion* **12**, 1161–1179. DOI:10.1037/a0025827.
- Bassili, J. N. (1979). Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face, *J. Pers. Soc. Psychol.* **37**, 2049–2058. DOI:10.1037/0022-3514.37.11.2049.
- Blais, C., Fiset, D., Roy, C., Saumure Régimbald, C. and Gosselin, F. (2017). Eye fixation patterns for categorizing static and dynamic facial expressions, *Emotion* **17**, 1107–1119. DOI:10.1037/emo0000283.
- Brainard, D. H. (1997). The psychophysics toolbox, *Spat. Vis.* **10**, 433–436. DOI:10.1163/156856897X00357.
- Buchan, J. N., Paré, M. and Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception, *Brain Res.* **1242**, 162–171. DOI:10.1016/j.brainres.2008.06.083.
- Calder, A. J., Young, A. W., Keane, J. and Dean, M. (2000). Configural information in facial expression perception, *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 527–551. DOI:10.1037/0096-1523.26.2.527.
- Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies, *Cereb. Cortex* **11**, 1110–1123. DOI:10.1093/cercor/11.12.1110.
- Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M. and Lepore, F. (2008). Audio-visual integration of emotion expression, *Brain Res.* **1242**, 126–135.

- Cornelissen, F. W., Peters, E. M. and Palmer, J. (2002). The Eyelink Toolbox: eye tracking with MATLAB and the Psychophysics Toolbox, *Behav. Res. Methods* **34**, 613–617. DOI:10.3758/BF03195489.
- Crosse, M. J., Di Liberto, G. M. and Lalor, E. C. (2016). Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration, *J. Neurosci.* **36**, 9888–9895. DOI:10.1523/JNEUROSCI.1396-16.2016.
- Dael, N., Mortillaro, M. and Scherer, K. R. (2012). Emotion expression in body action and posture, *Emotion* **12**, 1085–1101. DOI:10.1037/a0025737.
- de Gelder, B. (2009). Why bodies? Twelve reasons for including bodily expressions in affective neuroscience, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 3475–3484. DOI:10.1098/rstb.2009.0190.
- de Gelder, B. and Vroomen, J. (2000). The perception of emotions by ear and by eye, *Cogn. Emot.* **14**, 289–311. DOI:10.1080/026999300378824.
- de Gelder, B., Teunisse, J.-P. and Benson, P. J. (1997). Categorical perception of facial expressions: categories and their internal structure, *Cogn. Emot.* **11**, 1–23. DOI:10.1080/026999397380005.
- Eisenbarth, H. and Alpers, G. W. (2011). Happy mouth and sad eyes: scanning emotional facial expressions, *Emotion* **11**, 860–865. DOI:10.1037/a0022758.
- Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion, *J. Pers. Soc. Psychol.* **17**, 124–129. DOI:10.1037/h0030377.
- Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion, *Nature* **415**, 429–433. DOI:10.1038/415429a.
- Ernst, M. O. and Bühlhoff, H. H. (2004). Merging the senses into a robust percept, *Trends Cogn. Sci.* **8**, 162–169. DOI:10.1016/j.tics.2004.02.002.
- Etzi, R., Ferrise, F., Bordegoni, M., Zampini, M. and Gallace, A. (2018). The effect of visual and auditory information on the perception of pleasantness and roughness of virtual surfaces, *Multisens. Res.* **31**, 501–522. DOI:10.1163/22134808-00002603.
- Groner, R., Walder, F. and Groner, M. (1984). Looking at faces: local and global aspects of scanpaths, *Adv. Psychol.* **22**, 523–533.
- Hayhoe, M. and Ballard, D. (2005). Eye movements in natural behavior, *Trends Cogn. Sci.* **9**, 188–194. DOI:10.1016/j.tics.2005.02.009.
- Hsiao, J. H. and Cottrell, G. (2008). Two fixations suffice in face recognition, *Psychol. Sci.* **19**, 998–1006. DOI:10.1111/j.1467-9280.2008.02191.x.
- Hughes, H. C., Reuter-Lorenz, P. A., Nozawa, G. and Fendrich, R. (1994). Visual–auditory interactions in sensorimotor processing: saccades versus manual responses, *J. Exp. Psychol. Hum. Percept. Perform.* **20**, 131–153. DOI:10.1037/0096-1523.20.1.131.
- Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention, *Vis. Res.* **40**, 1489–1506. DOI:10.1016/S0042-6989(99)00163-7.
- Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G. and Caldara, R. (2009). Cultural confusions show that facial expressions are not universal, *Curr. Biol.* **19**, 1543–1548. DOI:10.1016/j.cub.2009.07.051.
- Jessen, S., Obleser, J. and Kotz, S. A. (2012). How bodies and voices interact in early emotion perception, *PLoS ONE* **7**, e36070. DOI:10.1371/journal.pone.0036070.
- Jessen, S. and Kotz, S. A. (2013). On the role of crossmodal prediction in audiovisual emotion perception, *Front. Hum. Neurosci.* **7**, 369. DOI:10.3389/fnhum.2013.00369.

- Juslin, P. N. and Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code?, *Psychol. Bull.* **129**, 770–814. DOI:10.1037/0033-2909.129.5.770.
- Kleiner, M., Brainard, D. and Pelli, D. (2007). What's new in Psychtoolbox-3?, *Perception* **36**, ECVF Abstract Supplement.
- Kokinou, J., Kotz, S. A., Tavano, A. and Schröger, E. (2015). The role of emotion in dynamic audiovisual integration of faces and voices, *Soc. Cogn. Affect. Neurosci.* **10**, 713–720. DOI:10.1093/scan/nsu105.
- Lansing, C. R. and McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences, *Percept. Psychophys.* **65**, 536–552. DOI:10.3758/BF03194581.
- Lischke, A., Berger, C., Prehn, K., Heinrichs, M., Herpertz, S. C. and Domes, G. (2012). Intranasal oxytocin enhances emotion recognition from dynamic facial expressions and leaves eye-gaze unaffected, *Psychoneuroendocrinology* **37**, 475–481. DOI:10.1016/j.psyneuen.2011.07.015.
- Massaro, D. W. and Egan, P. B. (1996). Perceiving affect from the voice and the face, *Psychon. Bull. Rev.* **3**, 215–221. DOI:10.3758/BF03212421.
- McGurk, H. and Macdonald, J. (1976). Hearing lips and seeing voices, *Nature* **264**, 746–748. DOI:10.1038/264746a0.
- Meredith, M. A. and Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus, *Science* **221**, 389–391. DOI:10.1126/science.6867718.
- Nummenmaa, T. (1964). *The Language of the Face (Jyväskylä Studies in Education, Psychology, and Social Research)*. Jyväskylä, Finland.
- Paulmann, S. and Pell, M. D. (2011). Is there an advantage for recognizing multi-modal emotional stimuli?, *Motiv. Emot.* **35**, 192–201. DOI:10.1007/s11031-011-9206-0.
- Paulmann, S., Titone, D. and Pell, M. D. (2012). How emotional prosody guides your way: evidence from eye movements, *Speech Commun.* **54**, 92–107. DOI:10.1016/j.specom.2011.07.004.
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies, *Spat. Vis.* **10**, 437–442. DOI:10.1163/156856897X00366.
- Peterson, M. F. and Eckstein, M. P. (2012). Looking just below the eyes is optimal across face recognition tasks, *Proc. Natl Acad. Sci. U.S.A.* **109**, E3314–E3323. DOI:10.1073/pnas.1214269109.
- Posner, M. I. (1980). Orienting of attention, *Q. J. Exp. Psychol.* **32**, 3–25. DOI:10.1080/00335558008248231.
- Rigoulot, S. and Pell, M. D. (2012). Seeing emotion with your ears: emotional prosody implicitly guides visual attention to faces, *PLoS ONE* **7**, e30740. DOI:10.1371/journal.pone.0030740.
- Russell, J. A. (1980). A circumplex model of affect, *J. Pers. Soc. Psychol.* **39**, 1161–1178. DOI:10.1037/h0077714.
- Samermit, P., Saal, J. and Davidenko, N. (2019). Cross-sensory stimuli modulate reactions to aversive sounds, *Multisens. Res.* **32**, 197–213. DOI:10.1163/22134808-20191344.
- Schyns, P. G., Petro, L. S. and Smith, M. L. (2007). Dynamics of visual information integration in the brain for categorizing facial expressions, *Curr. Biol.* **17**, 1580–1585. DOI:10.1016/j.cub.2007.08.048.

- Skuk, V. G. and Schweinberger, S. R. (2013). Adaptation aftereffects in vocal emotion perception elicited by expressive faces and voices, *PLoS ONE* **8**, e81691. DOI:10.1371/journal.pone.0081691.
- Smith, M. L., Cottrell, G. W., Gosselin, F. and Schyns, P. G. (2005). Transmitting and decoding facial expressions, *Psychol. Sci.* **16**, 184–189. DOI:10.1111/j.0956-7976.2005.00801.x.
- Stanford, T. R. and Stein, B. E. (2007). Superadditivity in multisensory integration: putting the computation in context, *NeuroReport* **18**, 787–792. DOI:10.1097/WNR.0b013e3280c1e315.
- Stein, B. E. and Meredith, M. A. (1993). *The Merging of the Senses*. MIT Press, Cambridge, MA, USA.
- Stevenson, R. A., Ghose, D., Fister, J. K., Sarko, D. K., Altieri, N. A., Nidiffer, A. R., Kurela, L. R., Siemann, J. K., James, T. W. and Wallace, M. T. (2014). Identifying and quantifying multisensory integration: a tutorial review, *Brain Topogr.* **27**, 707–730. DOI:10.1007/s10548-014-0365-7.
- Taffou, M., Guerchouche, R., Drettakis, G. and Viaud-Delmon, I. (2013). Auditory–visual aversive stimuli modulate the conscious experience of fear, *Multisens. Res.* **26**, 347–370. DOI:10.1163/22134808-00002424.
- Takagi, S., Hiramatsu, S., Tabei, K. and Tanaka, A. (2015). Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality, *Front. Integr. Neurosci.* **9**, 1. DOI:10.3389/fnint.2015.00001.
- Thompson, B., Hansen, B. C., Hess, R. F. and Troje, N. F. (2007). Peripheral vision: good for biological motion, bad for signal noise segregation?, *J. Vis.* **7**, 12. DOI:10.1167/7.10.12.
- Võ, M. L.-H., Smith, T. J., Mital, P. K. and Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces, *J. Vis.* **12**, 3. DOI:10.1167/12.13.3.
- Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior, *J. Nonverb. Behav.* **17**, 3–28. DOI:10.1007/BF00987006.
- Walker-Smith, G. J., Gale, A. G. and Findlay, J. M. (1977). Eye movement strategies involved in face perception, *Perception* **6**, 313–326. DOI:10.1068/p060313.
- Wallbott, H. G. and Scherer, K. R. (1986). Cues and channels in emotion recognition, *J. Pers. Soc. Psychol.* **51**, 690–699. DOI:10.1037/0022-3514.51.4.690.
- Yarbus, A. L. (1967). *Eye Movements and Vision*. Plenum Press, New York, NY, USA. DOI:10.1007/978-1-4899-5379-7.