

# Emotion

## **Differential Audiovisual Information Processing in Emotion Recognition: An Eye-Tracking Study**

Yueyuan Zheng and Janet H. Hsiao

Online First Publication, August 18, 2022. <http://dx.doi.org/10.1037/emo0001144>

### CITATION

Zheng, Y., & Hsiao, J. H. (2022, August 18). Differential Audiovisual Information Processing in Emotion Recognition: An Eye-Tracking Study. *Emotion*. Advance online publication. <http://dx.doi.org/10.1037/emo0001144>

# Differential Audiovisual Information Processing in Emotion Recognition: An Eye-Tracking Study

Yueyuan Zheng<sup>1</sup> and Janet H. Hsiao<sup>1, 2</sup>

<sup>1</sup> Department of Psychology, University of Hong Kong

<sup>2</sup> The State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong


Recent research has suggested that dynamic emotion recognition involves strong audiovisual association; that is, facial or vocal information alone automatically induces perceptual processes in the other modality. We hypothesized that different emotions may differ in the automaticity of audiovisual association, resulting in differential audiovisual information processing. Participants judged the emotion of a talking-head video under audiovisual, video-only (with no sound), and audio-only (with a static neutral face) conditions. Among the six basic emotions, disgust had the largest audiovisual advantage over the unimodal conditions in recognition accuracy. In addition, in the recognition of all the emotions except for disgust, participants' eye-movement patterns did not change significantly across the three conditions, suggesting mandatory audiovisual information processing. In contrast, in disgust recognition, participants' eye movements in the audiovisual condition were less eyes-focused than the video-only condition and more eyes-focused than the audio-only condition, suggesting that audio information in the audiovisual condition interfered with eye-movement planning for important features (eyes) for disgust. In addition, those whose eye-movement pattern was affected less by concurrent disgusted voice information benefited more in recognition accuracy. Disgust recognition is learned later in life and thus may involve a reduced amount of audiovisual associative learning. Consequently, audiovisual association in disgust recognition is less automatic and demands more attentional resources than other emotions. Thus, audiovisual information processing in emotion recognition depends on the automaticity of audiovisual association of the emotion resulting from associative learning. This finding has important implications for real-life emotion recognition and multimodal learning.


**Keywords:** emotion recognition, audiovisual information processing, facial expression, eye movements, EMHMM

**Supplemental materials:** <https://doi.org/10.1037/emo0001144.supp>

Emotion recognition is of vital importance in daily human interaction. It demands both temporal and spatial attention, and both emotional facial and vocal information play an important role (Young, 2018; Young et al., 2020). Thus, real-life emotion recognition involves audiovisual processing of dynamic information. However, most previous studies on emotion recognition focused on the processing of unimodal, static images of facial expressions.

While these studies have consistently shown that the recognition of different facial expressions involves different diagnostic features as reflected in eye movements (Schurgin et al., 2014), it remains unclear whether it applies as well to dynamic emotion recognition. In particular, when both emotional visual and auditory information are available, they can both contribute to the recognition and their association can be learned. This association is shown to interact with the attentional mechanisms, including both bottom-up and top-down attention. For example, temporally and spatially aligned multisensory information enhances saliency, suggesting that associated multisensory information facilitates bottom-up attention (Van der Burg et al., 2008). Similarly, a congruent cross-modal stimulus is shown to enhance selective attention to competing alternatives in the perception of ambiguous audiovisual stimuli (Van Ee et al., 2009). Also, audiovisual integration of speech (i.e., the classic McGurk effect where incongruent information in one modality alters the perception of information in another modality) was reduced when the amount of attentional resources available was low (Alsus et al., 2005). These findings suggest the involvement of top-down attention in audiovisual information processing. Talsma et al. (2010) argued that top-down attention plays an important role in multisensory information processing

Yueyuan Zheng  <https://orcid.org/0000-0002-0913-9514>

Janet H. Hsiao  <https://orcid.org/0000-0003-2271-8710>

Data and materials of this study are available on the Open Science Framework at <https://osf.io/7wjcp/>. The study design and analysis were not preregistered. Some of the data and ideas in the article were presented at the 42nd Annual Meeting of the Cognitive Sciences Society. We are grateful to RGC of Hong Kong (Project 17609117 to Janet H. Hsiao). We thank Andy Young for his suggestions and support during this research and Tsao Wing Yan for her help in data collection.

Correspondence concerning this article should be addressed to Janet H. Hsiao, Department of Psychology, University of Hong Kong, Pokfulam Road, Hong Kong, China. Email: [jhsiao@hku.hk](mailto:jhsiao@hku.hk)

when there is competition for attentional resources among multiple stimuli within each modality as selective attention is required to associate appropriate stimuli for the processing. This scenario applies well to emotion recognition, where selective attention is required for selecting diagnostic features from both emotional face and voice stimuli. Thus, the competition for attentional resources between the two modalities can influence recognition performance. This competition may be reflected in eye-movement planning behavior for diagnostic facial features. Indeed, eye movements elicited during an auditory attention task were shown to be predictive of attentional engagement and cued sound location (Braga et al., 2016), suggesting shared neural mechanisms between auditory and visual attention systems. Consistent with this finding, Zheng, Ye, and Hsiao (2022) showed that when watching documentary videos, participants who focused at the center of the screen as opposed to looking more frequently to different screen locations had a better comprehension of the auditory narratives.

Accordingly, in emotion recognition, simultaneous presentation of face and voice information may induce competition for attentional resources, influencing selective attention to diagnostic features in both modalities. Consequently, as compared with viewing only emotional face stimuli, the addition of emotional voice information may make participants look less often to diagnostic facial features for recognition, and their recognition performance may be associated with how well they can attend to diagnostic facial features with concurrent processing of emotional voice information.

Nevertheless, previous research has suggested a strong audiovisual association in emotion recognition; that is, emotional facial or vocal information alone may automatically induce perceptual processes in the other modality. For example, using a McGurk-like paradigm that was adapted to emotional incongruencies, De Gelder and Vroomen (2000) showed that the judgment of emotional face information was biased by incongruent emotional voice information and vice versa. This phenomenon may be because emotional experience can change frequently over time and is multimodal in nature; thus, there is a high demand on using both emotional face and voice information for recognition, resulting in strong audiovisual association (Young, 2018; Young et al., 2020). Indeed, multimodal perception is suggested to be an associative learning process (Connolly, 2014). For example, participants were able to acquire knowledge of arbitrary audiovisual associations through passive exposure (Seitz et al., 2007). Consistent with this speculation, individuals with facial emotion recognition problems are often also affected in voice emotion recognition, particularly in the recognition of fear and anger (Scott et al., 1997). These findings also suggested that there may be variations in the automaticity of audiovisual association in the recognition of different emotions due to differences in the amount of associative learning and demands during daily life. Emotions such as fear and anger may involve strong audiovisual association due to their relevance to survival (Skuse, 2003), whereas emotions learned or developed later in life such as disgust (Rottman, 2014; Widen & Russell, 2008, 2013) may involve weaker audiovisual association due to a reduced amount of associative learning. For the recognition of emotions involving strong audiovisual association, information in one modality may automatically activate associated features in the other modality (as demonstrated using the McGurk-like paradigm; De Gelder & Vroomen, 2000), resulting in weak audiovisual advantage over unimodal conditions. There may also be less

competition for attentional resources between modalities due to the acquired association. Consequently, the interference on selective attention to diagnostic features within each modality may be mitigated, and this phenomenon may be reflected in reduced eye-movement pattern change between audiovisual and unimodal conditions.

Here, we tested these hypotheses through eye tracking. Participants judged emotions of a talking-head video expressing one of the six basic emotions (Ekman et al., 1969) in audiovisual, video-only (without voice information), and audio-only (with a static neutral face image) conditions. We used the eye-movement analysis with hidden Markov models (EMHMM; Chuk et al., 2014) method to analyze eye-movement data since it provides quantitative measures of eye-movement pattern that take both temporal and spatial information into account, allowing us to examine eye-movement pattern change across different modality conditions. We expected that while participants would have better performance in the audiovisual condition in general due to the availability of more information, the recognition of disgust may show the largest audiovisual advantage over unimodal conditions since its relatively weaker audiovisual association may make the contributions from the two modalities more independent from each other. In contrast, for emotions involving strong audiovisual association such as fear and anger, since information presented in a unimodal condition may induce corresponding perceptual processes in the other modality, there may be a smaller advantage in the audiovisual condition over the unimodal conditions in recognition performance. In addition, in this case, eye-movement patterns may not change significantly across the three modality conditions due to the strong audiovisual association. In contrast, when audiovisual association is less automatic such as in the recognition of disgust, voice input may interfere with eye-movement planning, and the amount of eye-movement pattern change due to the interference may be negatively associated with the improvement in recognition performance due to additional voice information: the more the eye-movement pattern changes, the less the performance improves. Since participants' cognitive abilities and autistic traits (E. G. Smith & Bennetto, 2007) may also contribute to the recognition performance, these were measured as control variables in the examination of this association.

## Method

### Participants

Sixty-five participants (44 women and 21 men)<sup>1</sup> between 17 and 22 years old ( $M = 18.91$ ,  $SD = 1.20$ ) were recruited. Participants had similar educational backgrounds. They had normal or corrected-to-normal vision with no self-reported cognitive disabilities or psychological problems. Here, we aimed to examine whether the modality condition effect was different among the six emotions. A power analysis of 3 (Modality Conditions)  $\times$  6 (Emotion Conditions) repeated-measures analysis of variance (ANOVA)

<sup>1</sup> Although our participant recruitment resulted in more female than male participants, no significant gender difference was observed in any of the emotion or modality conditions in either emotion recognition performance or eye-movement pattern. Please see Supplemental Materials A for the relevant analysis results.

based on the means and standard deviations of the emotion recognition performance in a similar study (Livingstone & Russo, 2018) suggested that a sample size of 30 was sufficient for testing the interaction effect ( $\eta^2 = .29$ , power = 1.00,  $\alpha = .05$ ; Superpower; Lakens & Caldwell, 2021). In addition, we examined whether eye-movement pattern change between audiovisual and unimodal conditions could predict corresponding audiovisual advantage in recognition performance (i.e., normalized performance difference between audiovisual and unimodal conditions; please see details below). A power analysis of linear multiple regression indicated that, assuming a medium effect size ( $f^2 = .15$ , power = .80,  $\alpha = .05$ ) and one tested predictor (i.e., eye-movement pattern change), the required sample size was 55. We recruited 65 participants to allow for attrition. This study was approved by the Human Research Ethics Committee of the University of Hong Kong (reference number: EA1908013).

## Materials

The materials consisted of 432 short talking-head video clips taken from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS; Livingstone & Russo, 2018), where the recordings were validated for emotional validity, intensity, and genuineness. Each video clip lasted for 4,800 ms. The 432 video clips were divided evenly into three modality conditions, with 144 clips in each condition. In the audiovisual condition, both speech and video content were displayed; in the video-only condition, video content was displayed without speech content; in the audio-only condition, speech content and a static neutral face image were displayed. In each modality condition, the 144 clips were from 24 performers, with each acting out six basic categories of emotion, including happy, sad, angry, fearful, disgusted, and surprised (Ekman et al., 1969; Figure 1), summing up to 144 clips. Video clips from the same 24 performers were used in the three modality

conditions to better match the stimuli across the conditions. Participants viewed the video clips with a 60.5-cm viewing distance. Accordingly, the width of the face in a video clip spanned about  $8^\circ$  of visual angle, with the nose aligned with the center of the screen (following, e.g., Hsiao, An, et al., 2021; Hsiao & Cottrell, 2008; Hsiao & Liu, 2012). The same speech content “kids are talking by the door” was used in all stimuli; the meaning of the sentence was neutral in valence. We used acted emotional clips due to their stronger intensity than spontaneous ones (Caridakis et al., 2007).

## Apparatus

The eye movements of participants' dominant eye were recorded by an EyeLink 1000 plus eye tracker (the tower mount model; SR Research). The tracking mode was pupil and corneal reflection with 1,000-Hz sampling rate. In data acquisition, the threshold for saccade motion was  $.1^\circ$  visual angle, the threshold for saccade acceleration was  $8,000^\circ$  per square second, and the threshold for saccade velocity was  $30^\circ$  per second. These settings were EyeLink defaults for cognitive research. The resolution of the monitor (19 in.) was  $1,280 \times 1,024$  pixels. A chinrest was placed in front of the monitor to minimize participants' head movements. A Cedrus response box was used to collect behavioral responses.

## Design

The design consisted of two within-subject variables: modality condition (audiovisual vs. video only vs. audio only) and emotion (happy vs. sad vs. angry vs. fearful vs. disgusted vs. surprised). The dependent variables were emotion recognition accuracy and eye-movement pattern as assessed using EMHMM. Repeated-measures ANOVA was used.

**Figure 1**  
*Video Captures of Six Emotions From RAVDESS*



*Note.* Images were adapted from RAVDESS database (Livingstone & Russo, 2018) (CC BY-NC-SA 4.0 license), which had consent obtained from the photographed individual. See the online article for the color version of this figure.

To test our hypothesis regarding audiovisual advantage, in a separate analysis, we examined audiovisual advantage using either the video-only or the audio-only condition as the baseline separately. Specifically, we defined normalized change in performance as  $\text{Normalized Change} = (A - B) / (|A| + |B|)$ , where A stands for performance in the audiovisual condition and B stands for performance in the baseline condition, either the video-only or the audio-only condition. This measure normalized the differences in performance across emotion conditions in the examination of audiovisual advantage. We then examined normalized change in performance between the audiovisual and audio-only conditions and between the audiovisual and video-only conditions separately using repeated-measures ANOVA with emotion as the independent variable.

In addition, to test our hypothesis about the relationship between eye-movement pattern change and audiovisual advantage, we examined what factors, including eye movement, cognitive ability, and autistic trait measures, could predict the advantage of the audiovisual condition over the video-only or audio-only condition in recognition performance through correlation and stepwise multiple regression analyses. Autistic trait measures were included since individuals with autism were found to have difficulty in audiovisual integration as compared with controls (E. G. Smith & Bennetto, 2007).

Procedure

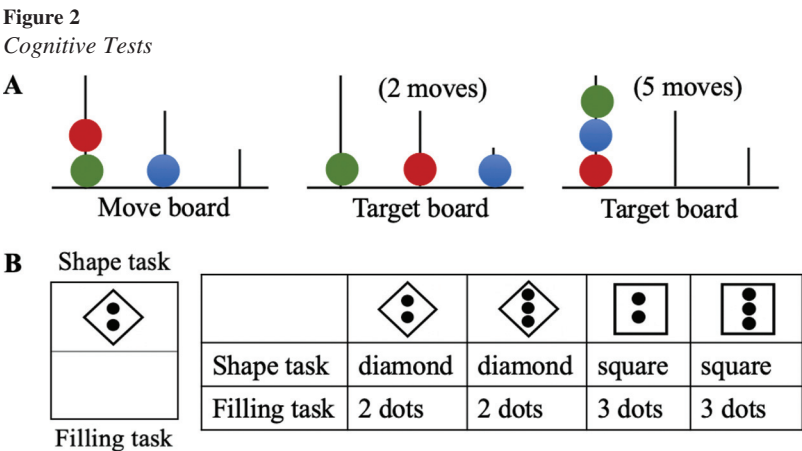
Participants performed an emotion recognition task, followed by cognitive ability tests including verbal and visuospatial two-back tasks to measure their working memory capacity, the Tower of London test to examine their executive function/planning ability, the Multitasking Test to test their task-switching ability, the Trail Making Test to measure their visual attention and switching ability, and the flanker task for testing their selective attention ability. They also filled in the adult version of the Autism Spectrum Quotient (AQ) questionnaire (Baron-Cohen et al., 2001) to assess their autistic traits. These cognitive ability and AQ measures were included as control variables to examine whether online eye-movement behavior could predict the advantage of the audiovisual condition over the video-only or the audio-only condition in

recognition performance if cognitive abilities and autistic traits were taken into account.

In emotion recognition, the 432 video clips were randomly divided into 12 blocks, with an equal number of stimuli from each emotion by modality condition combination in each block. More specifically, each block contained 36 trials, with two trials from each emotion by modality condition combination randomly drawn from the 24 performers. The block order and the trial order in each block were both randomized. Each participant received a different randomization. The standard 9-point calibration and validation of eye-tracking procedure was conducted before each block and whenever drift correction error was more than 1° of visual angle. Each trial started with a solid circle in the middle of the screen for drift correction, followed by a fixation cross presented at the center of one of the four quadrants of the screen at random to direct their attention away from the center. Participants were instructed to look at the fixation cross when it appeared, and the experimenter pressed a key to present the video clip when a stable fixation was observed at the fixation cross. The video clip presentation was at the center of the screen, followed by a 500 ms blank screen. Participants were asked to judge the emotion of the video clip from the six emotion categories as accurately and quickly as possible by pressing corresponding buttons. They could respond any time after the onset of the video clip. The screen turned blank for 500 ms after the response. Response accuracy was measured, and their eye movements when viewing the video clip were recorded and analyzed.

In the two-back tasks (Lau et al., 2010), participants judged whether the presented English letter/symbol location in the current trial was the same as the one presented two trials before in the verbal/visuospatial task, respectively. Each symbol was presented for 1,000 ms, followed by a 2,500-ms blank screen. Accuracy and response time (RT) were measured. Each task had 52 trials.

In the Tower of London test (L. H. Phillips et al., 2001), participants moved three discs of different colors one at a time from an initial position to match a goal position with a minimum number of moves (Figure 2A). Participants completed 12 trials. The total number of moves, execution time, preplanning time before executing the first move, and total time were measured.



Note. Panel A: Examples of the Tower of London test. Panel B: Stimuli used in the Multitasking Test. See the online article for the color version of this figure.



In the Multitasking Test (Stoet et al., 2013), four types of figures with different combinations of shapes and fillings (Figure 2B, right) were presented one at a time in either the top or the bottom half of a box (Figure 2B, left). Participants performed a dual task where they judged the shape of the figure (the shape task) when the figure was shown in the top half and judged the filling (the number of dots) of the figure (the filling task) when it was in the bottom half. The figure was presented for 2,500 ms, followed by a 500-ms blank screen. A shape-only and a filling-only task were tested sequentially before the dual task to measure the baseline performance without task switching. The switching ability was measured as the RT in the dual task minus the average RT during the two no-switching tasks.

In the Trail Making Test (Reitan, 1958), in Part A, participants connected 25 circles from number 1 to 25 sequentially. In Part B, they connected 24 circles with alternating numbers and English letters in sequential order. The RT was recorded separately for the two parts.

In the flanker task (Ridderinkhof et al., 1999), participants judged the direction of an arrow flanked by four other arrows. In congruent trials, the flanking arrows pointed in the same direction as the target arrow, whereas in incongruent trials, they pointed in the opposite direction. In neutral trials, the flankers were nondirectional symbols. Their accuracy and RT were measured.

A 50-item AQ was adopted to measure autistic traits (Baron-Cohen et al., 2001). Each item was scored from 1 to 4, and higher scores corresponded with more autistic-like behavior. Participants' autistic traits were assessed using the scores of three subscales including Social Skills, Communication/Mindreading, and Details/Patterns, as recommended by English et al. (2020) according to their psychometric analysis. English et al. (2020) reported acceptable composite score reliability for the three subscales (Social Skills and Details/Patterns > .70; Communication/Mindreading = .67).

## Eye-Movement Data Analysis

Eye-movement data were first aligned according to the center point between the two eyes across videos. EMHMM (Chuk et al., 2014) was then used to analyze eye-movement data in order to quantify similarities among individual eye-movement patterns, taking both temporal and spatial dimensions of eye movements into account (see also Chuk et al., 2020; Hsiao, Lan, et al., 2021). It has been used to quantify eye-movement patterns in a variety of visual tasks, including face recognition (e.g., An & Hsiao, 2021; C. Y. H. Chan et al., 2018; Chuk, Chan, & Hsiao, 2017; Chuk, Crookes, et al., 2017; Zheng, Chen, et al., 2022), face matching (Hsiao, An, et al., 2021), face emotion recognition (Zhang et al., 2019), passive viewing of faces and images (F. H. F. Chan, Barry, et al., 2020; F. H. F. Chan, Jackson, et al., 2020; F. H. F. Chan et al., 2022; F. H. F. Chan, Suen, et al., 2020; Cho et al., 2022a, 2022b, 2022c), gaze perception (S. K. W. Chan et al., 2022), sustained attention to response (Lee et al., 2021), website viewing (Eckhardt et al., 2013), video viewing (Zheng, Ye, & Hsiao, 2022), scene perception (Hsiao, Lan, et al., 2021), reading (Liao et al., 2022), and visual search (Hsiao, Chan, et al., 2021). EMHMM is based on the assumption that the current fixation in a visual task is conditioned on the previous fixation, and thus eye movements in the task may be considered a Markovian stochastic process, which can be better understood using hidden Markov

models (HMMs). Using this approach, a participant's eye movements in each of the modality condition and emotion combinations were summarized using an HMM (a type of time-series statistical model in machine learning), including person-specific regions of interest (ROIs) and transition probabilities among the ROIs. The hidden states of the HMM corresponded to the ROIs, with Gaussian emissions corresponding to fixation locations. The parameters of the HMM were estimated from the participant's eye-movement data using the variational Bayesian expectation maximization algorithm, with the optimal number of ROIs determined through a variational Bayesian approach. Specifically, we used 1–6 ROIs as the preset range of ROIs for training individual HMMs. Each model with a specific number of ROIs was trained for 300 times, and the model with the highest data log-likelihood was used for the analysis. Since previous studies using EMHMM on face or facial expression recognition typically had a range of 2–4 ROIs as the median number of ROIs (e.g., An & Hsiao, 2021; C. Y. H. Chan et al., 2018; Chuk, Crookes, et al., 2017; Zhang et al., 2019), we used 1–6 ROIs as the preset range to ensure a good coverage of variations among individual data.

Following previous studies (e.g., An & Hsiao, 2021; C. Y. H. Chan et al., 2018; Chuk et al., 2014; Chuk, Chan, & Hsiao, 2017; Zhang et al., 2019), the resulting 1,170 individual HMMs (18 models for each participant) were clustered into two representative groups based on their similarities through the variational hierarchical expectation maximization algorithm (Coviello et al., 2014). Note that a new variational Bayesian hierarchical expectation maximization methodology that uses Bayesian methods to determine the optimal number of clusters (Lan et al., 2021) also suggested two as the optimal number of clusters in the current data. In addition, this new methodology suggested two as the optimal number of ROIs used in the representative models of the group patterns. Thus, we set the number of ROIs to two accordingly, which was also the median number of ROIs among individual models. (The number of ROIs in individual models was also determined automatically through a variational Bayesian approach, and thus different models may have different numbers of ROIs.) The clustering algorithm was run for 300 times with different initializations, and the result with the highest data log-likelihood was used for the analysis. The similarity of an individual's eye-movement pattern to a resulting representative group pattern then was quantified using the log-likelihood of the data being generated by the HMM of the representative pattern. To measure participants' eye-movement pattern in each condition along the dimension of the two representative group patterns, we defined A-B scale as  $A-B \text{ Scale} = (A - B) / (|A| + |B|)$ , where A stands for the log-likelihood of the participant's eye-movement data being classified as the first pattern and B stands for the log-likelihood of the data being classified as the second pattern. A more positive A-B scale value indicates higher similarity to the first pattern as opposed to the second pattern (e.g., An & Hsiao, 2021; Chan et al., 2018; Hsiao, An, et al., 2021; Hsiao, Chan, et al., 2021; Hsiao et al., 2022).

## Transparency and Openness

We report how we determined our sample size, all manipulations, and all measures in the study. Data were analyzed using SPSS (Nie et al., 1975) and Jamovi (Šahin & Aybek, 2019). Data and materials of this study can be found on the Open Science

Framework at <https://osf.io/7wjcpl/>. Some of the data and ideas in the article were presented at the Annual Meeting of the Cognitive Sciences Society in 2020 (Zheng & Hsiao, 2020). The study design and analysis were not preregistered.

## Results

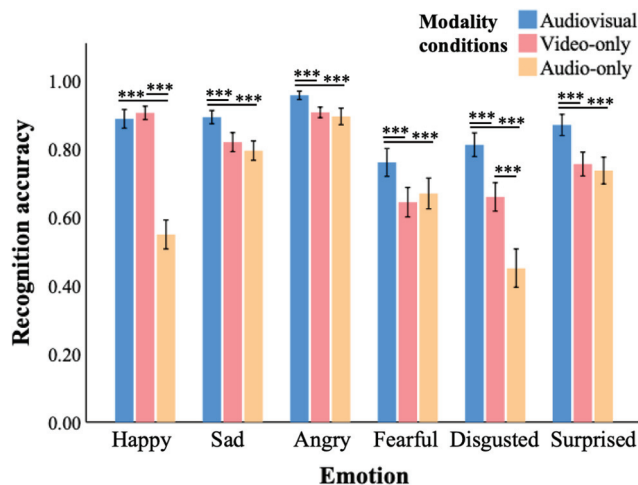
In emotion recognition accuracy,<sup>2</sup> there was a main effect of modality condition,  $F(2, 128) = 429.44, p < .001, \eta^2 = .87$ , 90% confidence interval (CI) [.84, .89].<sup>3</sup> Participants had higher accuracy in the audiovisual than the video-only condition,  $t(64) = 13.21, p < .001, d = 2.20$ , 95% CI [1.74, 2.64], and in the video-only than the audio-only condition,  $t(64) = 16.05, p < .001, d = 1.76$ , [.88, 2.64]. A significant main effect of emotion was also found,  $F(5, 320) = 66.52, p < .001, \eta^2 = .51$ , 90% CI [.44, .55]. People had the best performance in recognizing anger, followed by sadness, happiness and surprise, and disgust. They performed the worst in recognizing fear. There was an interaction between modality condition and emotion,  $F(10, 640) = 52.66, p < .001, \eta^2 = .45$ , [.40, .48]. We then examined the modality condition effect in different emotions separately (see Figure 3). Repeated-measures ANOVA showed a significant modality condition effect in the recognition of all emotions, including happiness,  $F(2, 128) = 277.76, p < .001, \eta^2 = .81$ , [.76, .84]; sadness,  $F(2, 128) = 29.83, p < .001, \eta^2 = .32$ , [.20, .41]; anger,  $F(2, 128) = 26.83, p < .001, \eta^2 = .30$ , [.18, .39]; fear,  $F(2, 128) = 38.00, p < .001, \eta^2 = .37$ , [.26, .46]; disgust,  $F(2, 128) = 119.72, p < .001, \eta^2 = .65$ , [.57, .71]; and surprise,  $F(2, 128) = 31.45, p < .001, \eta^2 = .33$ , [.22, .42]. For happiness, participants' performance did not differ between the audiovisual and video-only conditions,  $t(64) = -1.74, p = .087, d = -.22$ , 95% CI [-.31, -.11], but was higher in the video-only than audio-only condition,  $t(64) = 18.78, p < .001, d = 2.33$ , [1.16, 3.49]. This suggested that they mainly relied on visual information for the recognition of happiness. For disgust, participants were significantly more accurate in the audiovisual than

video-only condition,  $t(64) = 11.10, p < .001, d = 1.38$ , [.69, 2.07], and in the video-only than audio-only condition,  $t(64) = 7.25, p < .001, d = .90$ , [.45, 1.35]. This indicated that visual information was more informative than audio information, and the combination of the two led to the best recognition. For the other emotions, while the best performance was achieved in the audiovisual condition, there was no significant difference between video-only and audio-only conditions, suggesting that emotional face and voice information were similarly informative.

To further understand what led to the differences among modality conditions in disgust recognition, we examined the responses participants made toward disgusted stimuli. A 3 (Modality Condition: audiovisual vs. video only vs. audio only)  $\times$  6 (Response Type: happy vs. sad vs. angry vs. fearful vs. disgusted vs. surprised) repeated-measures ANOVA showed a main effect of response type,  $F(5, 320) = 166.91, p < .001, \eta^2 = .72$ , 90% CI [.68, .75], and an interaction between modality condition and response type,  $F(10, 640) = 12.69, p < .001, \eta^2 = .17$ , [.11, .20]. Post hoc  $t$  tests showed that disgust was misidentified as other response types more frequently in the audio-only condition than in the other two modality conditions, except for the case of sadness as disgust was misidentified as sadness more frequently in the video-only condition than the other two modality conditions (see Figure 4). As shown in Figure 4, in the video-only condition, disgusted facial information was more frequently confused with sad facial information than other emotions ( $ps < .001$ ). Thus, adding concurrent facial information to disgusted voices (audiovisual vs. audio-only condition) significantly reduced the frequency of misidentifying disgusted voices as other emotions except for sadness. Also, adding concurrent vocal information to disgusted faces (audiovisual vs. video-only condition) significantly reduced the frequency of misidentifying disgusted faces as expressing sadness. In contrast, in the audio-only condition, disgusted voices were more frequently misidentified as sad or angry voices than happy or fearful voices ( $ps < .001$ ) and more frequently misidentified as surprised voices than happy voices ( $p < .001$ ). A full confusion matrix for all modality conditions and all emotion categories can be found in Supplemental Materials B.

To examine audiovisual advantage, analysis on normalized change in performance between the audiovisual and video-only conditions showed a significant effect of emotion,  $F(5, 320) = 29.1, p < .001, \eta^2 = .31$ , 90% CI [.24, .37]: Participants had the largest performance increase in recognizing disgust and the least in recognizing happiness (Figure 5A). Similarly, analysis on normalized change in performance between the audiovisual and audio-only conditions showed a significant effect of emotion,  $F(5, 320) = 51.6, p < .001, \eta^2 = .45$ , [.37, .50]: Participants had the largest performance increase in recognizing disgust and happiness (Figure 5B). This result suggested that the recognition of disgust involved the largest audiovisual advantage, consistent with our hypothesis.

**Figure 3**  
Emotion Recognition Accuracy in Different Conditions

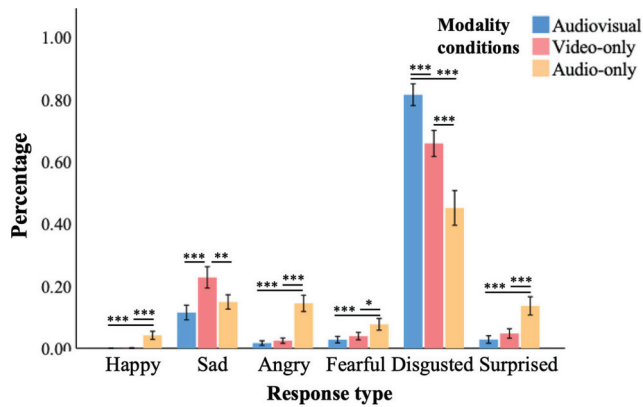


Note. Error bars: 95% CI. See the online article for the color version of this figure.  
\*\*\*  $p < .001$ .

<sup>2</sup> The recognition accuracy of this study had an excellent split-half reliability according to Spearman-Brown coefficient (Parsons et al., 2019),  $r_{sb} = .93$ . It also has acceptable to excellent split-half reliability across individual conditions: audiovisual condition,  $r_{sb} = .86$ ; audio-only condition,  $r_{sb} = .88$ ; video-only condition,  $r_{sb} = .78$ ; fear,  $r_{sb} = .92$ ; sadness,  $r_{sb} = .82$ ; disgust,  $r_{sb} = .88$ ; surprise,  $r_{sb} = .89$ ; happiness,  $r_{sb} = .75$ ; anger,  $r_{sb} = .68$ .

<sup>3</sup> Ninety-percent CI instead of 95% CI is reported for  $F$  tests since  $F$  tests are one sided (Steiger, 2004).

**Figure 4**  
Percentage of Different Responses Made When a Disgusted Emotion Was Presented in Different Modality Conditions



Note. Error bars: 95% CI. See the online article for the color version of this figure.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

In eye-movement data, we discovered two representative patterns as the result of clustering: the nose-focused and eyes-focused patterns (see Figure 6), consistent with a previous EMHMM study on emotion recognition using static face images (Zhang et al., 2019). Participants adopting the nose-focused pattern typically started a trial with a fixation in the nose region/red ROI (99%) and remained looking at the same region afterward (97%), with a small possibility (3%) to transit to the mouth region/green ROI. In contrast, participants adopting the eyes-focused pattern had 94% possibility to first look at the center of the eye region/red ROI and remained looking at the same region afterward. Occasionally (6%), they started from the left eye/green ROI and remained there afterward (94%) or switched to the center of the eye region/red ROI. The two patterns differed significantly (Chuk et al., 2014): Data from those using the nose-focused pattern were more likely to be generated by the nose-focused than eyes-focused HMM,  $t(446) = 17.08$ ,  $p < .001$ ,  $d = .81$ , 95% CI [.40, 1.21], and data from those with the eyes-focused pattern were more likely to be generated by the eyes-focused than nose-focused HMM,  $t(722) = 49.47$ ,  $p < .001$ ,  $d = 1.84$ , [1.30, 2.76]. For current purposes, we referred to the A-B scale for quantifying participants' eye-movement pattern as the nose-eyes scale, with a higher nose-eyes scale indicates higher similarity to the nose-focused pattern.<sup>4</sup>

The results on the nose-eyes scale showed no main effect of modality condition,  $F(2, 128) = 2.52$ ,  $p = .085$ ,  $\eta^2 = .04$ , 90% CI [.00, .10]. There was a main effect of emotion,  $F(5, 320) = 37.98$ ,  $p < .001$ ,  $\eta^2 = .37$ , [.30, .43]: Participants had a more nose-focused pattern when recognizing fear, followed by happiness and surprise; they adopted a more eyes-focused pattern for disgust, followed by sadness and anger. This emotion effect interacted with modality condition,  $F(10, 640) = 31.08$ ,  $p < .001$ ,  $\eta^2 = .33$ , [.27, .36]. We then examined the modality condition effect in different emotions separately (see Figure 7) using repeated-measures ANOVA. Interestingly, no significant difference among the three modality conditions was observed in happiness,  $F(2, 128) = .59$ ,  $p = .558$ ,  $\eta^2 < .01$ , [.00, .04]; sadness,  $F(2, 128) = .31$ ,  $p = .733$ ,  $\eta^2 < .01$ , [.00, .29]; anger,  $F(2, 128) = .46$ ,  $p = .632$ ,  $\eta^2 < .01$ , [.00, .04]; fear,  $F(2, 128) = 2.38$ ,  $p = .096$ ,  $\eta^2 = .04$ , [.00, .09]; and surprise,  $F(2,$

128)  $< .01$ ,  $p = .997$ ,  $\eta^2 < .01$ , [.00, .00]. In contrast, in disgust, a significant modality condition effect was observed,  $F(2, 128) = 50.10$ ,  $p < .001$ ,  $\eta^2 = .44$ , [.33, .52]: Eye-movement pattern in the audio-only condition was more nose focused than the audiovisual condition,  $t(64) = 3.37$ ,  $p = .001$ ,  $d = .42$ , 95% CI [.21, .63], and in the video-only condition was more eyes focused than the audiovisual condition,  $t(64) = -10.49$ ,  $p < .001$ ,  $d = -1.30$ , [-1.84, -.65]. This result was consistent with our hypothesis that for emotions with strong audiovisual association, information in one modality may activate associated information in the other modality automatically, resulting in similar eye-movement patterns across the three modality conditions. In contrast, for emotions with weak audiovisual association such as disgust, adding voice information to the video makes eye movements focus less on the diagnostic eye region. (To compare with the results using a predefined ROI approach, please refer to Supplemental Materials C).

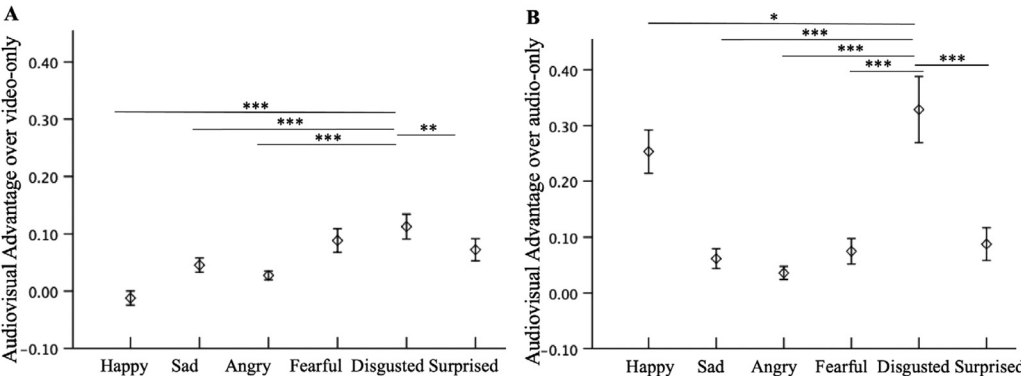
The stepwise multiple regression analysis predicting normalized change in accuracy between the audiovisual and video-only conditions was conducted using normalized change in nose-eyes scale between the two conditions and all cognitive test performance and autistic trait measures.<sup>5</sup> The results showed that normalized change in nose-eyes scale was the only significant predictor,  $\beta = -.32$ ,  $p = .009$ , accounting for a significant portion of variance,  $F(1, 63) = 7.29$ ,  $p = .009$ ,  $R^2 = .10$ , 90% CI [.02, .23]. The tests for multicollinearity indicated a low level of multicollinearity (tolerance = 1.000). It indicated that the less the eye-movement pattern changed, the larger the accuracy increased in the audiovisual condition over the video-only condition. A similar stepwise regression analysis predicting normalized change in accuracy between the audiovisual and audio-only conditions was conducted. The result showed that normalized change in nose-eyes scale was not a significant predictor; instead, significant predictors included execution time of the Tower of London test,  $\beta = .32$ ,  $p = .011$ , and AQ Details/Patterns,  $\beta = .29$ ,  $p = .021$ , accounting for a significant portion of variance,  $F(2, 64) = 5.11$ ,  $p = .009$ ,  $R^2 = .14$ , [.02, .25]. The tests for multicollinearity indicated a low level of multicollinearity for both execution time of the Tower of London test (tolerance = .949) and AQ Details/Patterns (tolerance = .949). It suggested that

<sup>4</sup> The nose-eyes scale measured through EMHMM had an excellent split-half reliability,  $r_{sb} = 1.00$ . It also has excellent split-half reliability across individual conditions: video-only condition,  $r_{sb} = 1.00$ ; audiovisual condition,  $r_{sb} = .99$ ; audio-only condition,  $r_{sb} = .99$ ; fear,  $r_{sb} = .98$ ; sadness,  $r_{sb} = .99$ ; disgust,  $r_{sb} = .99$ ; surprise,  $r_{sb} = .99$ ; happiness,  $r_{sb} = .99$ ; anger,  $r_{sb} = .98$ .

<sup>5</sup> The reliability of cognitive and autistic trait measures in the current data were examined. *N*-back tasks had acceptable to excellent split-half reliability: visual-spatial task:  $r_{sb} = .84$  for accuracy,  $r_{sb} = .95$  for RT; verbal task:  $r_{sb} = .77$  for accuracy,  $r_{sb} = .94$  for RT. The flanker task has acceptable to excellent split-half reliability: congruent trials:  $r_{sb} = .87$  for accuracy,  $r_{sb} = .90$  for RT; incongruent trials:  $r_{sb} = .72$  for accuracy,  $r_{sb} = .90$  for RT. The Tower of London test had acceptable split-half reliability: total time,  $r_{sb} = .59$ ; planning time,  $r_{sb} = .78$ ; executing time,  $r_{sb} = .52$ . Multitasking ability as measured in the Multitasking Test had poor to acceptable split-half reliability:  $r_{sb} = .46$  for accuracy,  $r_{sb} = .69$  for RT. The split-half reliability of the Trail Making Test could not be estimated from the current data; however, previous studies have reported good test-retest reliability (Giovagnoli et al., 1996). AQ measures had acceptable to good split-half reliability: Communication/Mindreading,  $r_{sb} = .77$ ; Social Skills,  $r_{sb} = .88$ ; Details/Patterns,  $r_{sb} = .72$ . We used  $r_{sb} = .50$  as a cutoff to remove measures with low reliability from the analysis (multitasking accuracy was removed accordingly).



**Figure 5**  
*The Normalized Change in Performance (Audiovisual Advantage) Between the Audiovisual and (A) Video-Only and (B) Audio-Only Conditions for Different Emotion Categories*



*Note.* Error bars: 95% CI. In both cases, the recognition of disgust had the largest audiovisual advantage.  
\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

the lower the executive function ability and the higher the autistic traits in Details/Patterns, the larger the accuracy increase in the audiovisual condition over the audio-only condition. As the recognition accuracy data (see Figure 3) suggested that audio information was less informative than visual information in disgust recognition, those who had lower executive function ability and higher autistic traits in Details/Patterns may have more recognition difficulty and consequently benefit more from the availability of the more informative visual information in the audiovisual condition relative to the audio-only condition.

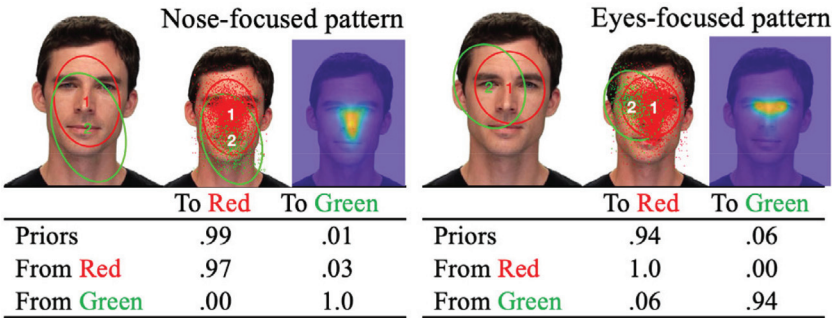
**Discussion**

Recent research has suggested that emotion recognition involves strong audiovisual association due to its multimodal nature and

high demands on accuracy and efficiency (Young, 2018; Young et al., 2020). We hypothesized that in dynamic emotion recognition with both voice and face information, different emotions may differ in the automaticity of audiovisual association, resulting in differential audiovisual information processing. Specifically, for emotions with strong audiovisual association, information in one modality may activate associated information in the other modality automatically, leading to weaker audiovisual advantages. In contrast, for emotions with weak audiovisual association, competition for attentional resources between the two modalities will interfere with selective attention to diagnostic features: the larger the interference, the smaller the audiovisual advantage.

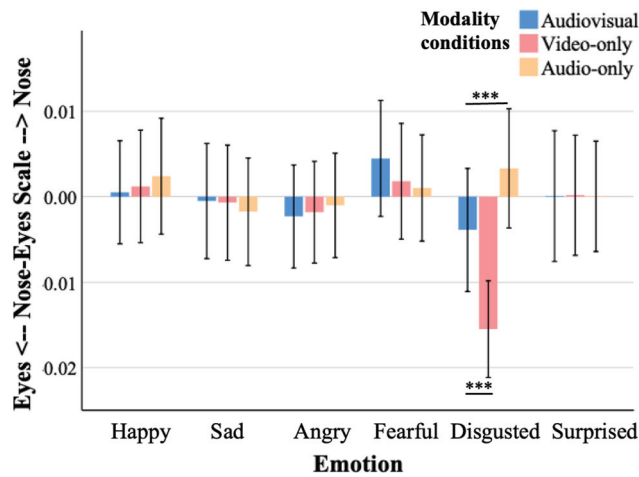
Our results showed that among the six basic emotions, disgust had the largest audiovisual advantage over either the video-only or the audio-only condition. Also, participants' eye-movement

**Figure 6**  
*The Nose-Focused (Left) and Eyes-Focused (Right) Patterns*



*Note.* Ellipses show ROIs as two-dimensional Gaussian emissions. The table shows transition probabilities among the ROIs. Priors show the probabilities that a fixation sequence starts from the ellipse. The image in the middle shows the ROI assignments of the raw fixations (subsampled 10% of the fixations with 0.8 transparency to better visualize the fixation distributions). The assignment of fixations to the ROIs was based on the ROI sequence with the largest posterior probability given the fixation sequence. The image on the right shows the corresponding heatmap. Images were adapted from RAVDESS database (Livingstone & Russo, 2018) (CC BY-NC-SA 4.0 license), which had consent obtained from the photographed individual. See the online article for the color version of this figure.

**Figure 7**  
*Nose-Eyes Scale in Different Conditions*



Note. Error bars: 95% CI. See the online article for the color version of this figure.

\*\*\*  $p \leq .001$ .

pattern did not change significantly across the three modality conditions in the recognition of all emotions except for disgust. This result suggested that in all basic emotions except for disgust, concurrent vocal information improved recognition performance without interfering with eye-movement planning for diagnostic facial features. Interestingly, even in the audio-only condition, where participants viewed a static neutral face with emotional voice, they showed similar eye movements to the audiovisual or video-only conditions, demonstrating mandatory audiovisual association. This result is consistent with the literature on multimodal mental imagery (Nanay, 2018), which suggests that perceptual processing in one sensory modality can be triggered by stimulation in another. When diagnostic facial and vocal features are consistently used together for emotion recognition, they become highly associated, and thus vocal input alone can trigger eye movement for corresponding facial features. Indeed, Schurgin et al. (2014) showed that people could plan eye movements for diagnostic features of a given emotion when viewing a neutral face. Previous patient studies have suggested strong audiovisual association in fear and anger recognition (Scott et al., 1997). Our results further demonstrated strong audiovisual association in happiness, sadness, and surprise recognition.

In contrast, in disgust recognition, participants' eye movements in the audiovisual condition were less eyes focused than the video-only condition and more eyes focused than the audio-only condition. Since diagnostic facial features for disgust recognition, including squinting eyes and wrinkled nose (Green & Guo, 2018; Rottman, 2014; Schyns et al., 2007; M. L. Smith et al., 2005), are better covered in the eyes-focused pattern (which covers both the eye and nose regions; Figure 6), this result suggested that vocal information interfered with eye-movement planning, resulting in a less eyes-focused pattern in the audiovisual than video-only condition. Interestingly, this eye-movement pattern change uniquely predicted the performance change between the two conditions with cognitive abilities controlled: Those whose online eye-movement behavior was affected the least benefited the most from concurrent vocal

information. Indeed, the errors participants made in the video-only condition most often involved misidentifying disgust as sadness, and adding concurrent vocal information significantly reduced the frequency of this misidentification (see Figure 4). As diagnostic facial features for sadness recognition are around the eye and eyebrow regions (Schurgin et al., 2014; Schyns et al., 2007; M. L. Smith et al., 2005) and the moving mouth of our talking-head stimuli may have made visual features around the mouth less salient for emotion identification, a shift to adopt a more nose-focused pattern (which covers the nose and mouth regions) may have been suboptimal. Note also that in the audio-only condition, disgusted voices were most often confused with sad and angry voices, followed by surprised voices (Figure 4; see also Widen & Russell, 2013). Since the diagnostic facial features for these emotions were also around the eye regions (except for surprise, where the mouth region may also be diagnostic; M. L. Smith et al., 2005), adopting a more eyes-focused pattern in the audiovisual condition may also be beneficial for reducing the confusion with these emotions from vocal information.

In contrast, in disgust recognition, the performance increase in the audiovisual relative to audio-only condition was not predicted by eye-movement pattern change between the two conditions. Instead, it was best predicted by executive function ability and autistic traits in Details/Patterns: Those who had low executive function ability and higher autistic traits in Details/Patterns benefited more with the addition of visual information, which was more informative than auditory information in disgust recognition (see Figure 3). We speculated that this phenomenon may be because those with lower executive function ability and higher autistic traits in Details/Patterns may have more difficulty in disgusted voice recognition and consequently benefit more from concurrent visual information. Consistent with this speculation, emotion recognition from speech prosody is shown to be associated with socioemotional adjustment and cognitive and self-regulation abilities in children (Neves et al., 2021), suggesting the role of executive function. Individuals with autism spectrum disorder are reported to have impaired vocal emotion recognition as compared with matched controls, and this impairment is associated with autism spectrum disorder traits and symptoms (Schelinski & von Kriegstein, 2019). In addition, higher autistic traits in attention to detail in healthy adults are found to be associated with better performance in face recognition through the mediation of increased looking at eyes (Davis et al., 2017). This result suggested that participants with higher autistic traits in Details/Patterns in our study may have benefited more from concurrent visual information in the recognition of disgusted voices due to better ability to obtain diagnostic facial information from the eye region.

Among the six basic emotions, disgust is learned and developed the latest in life, with a majority of children as old as 7 years of age still misidentify a disgusted expression as angry (Rottman, 2014; Widen & Russell, 2008, 2013). Thus, it may involve a smaller amount of associative learning for audiovisual information. Consequently, disgust recognition may involve weaker audiovisual association than the other emotions, resulting in the observed audiovisual effects. Note that here we used speech stimuli with emotional voice, which differed from some diagnostic vocalizations of disgust such as "yuk!" and "ugh!" (M. L. Phillips et al., 1998). These diagnostic vocalizations may have stronger audiovisual association with facial features of disgust than emotional voice. In addition, these vocalizations of disgust may lead to more diagnostic facial features for disgust recognition around the

mouth region, which may consequently change participants' eye-movement patterns. Future work will examine these possibilities.

Our results suggested that the automaticity of audiovisual association modulates eye movements and performance in emotion recognition. This finding has important implications for cognitive tasks involving audiovisual information processing. For example, person identification is argued to have weaker audiovisual association than emotion recognition since face and voice identities do not change over time and are often identified separately (Young, 2018; Young et al., 2020). Indeed, people who have face identification problems (prosopagnosia) typically have deficits specific to the visual modality and do not have difficulties in identifying familiar people by voice (Barton & Corrow, 2016). Accordingly, similar to disgust recognition, concurrent voice information may interfere with eye-movement planning for face identification, and those whose eye movements are less interfered may benefit more from concurrent voice information. Similarly, in multimedia learning, inputs from two modalities that have strong association, such as auditory narratives and corresponding visual subtitles, typically facilitate learning, whereas those with weak association may compete for attentional resources, and the performance may depend on one's online information extraction strategy as revealed in eye-movement behavior (Zheng, Ye, & Hsiao, 2022). In a separate, explorative analysis, we found that none of the cognitive ability measures used here could predict participants' eye-movement pattern change between the audiovisual and video-only conditions. Thus, it remains unclear what cognitive abilities are associated with being less interfered by concurrent auditory information in eye-movement planning. It may be related to auditory working memory or other executive functions not measured here, and this requires further investigation.

Note that in the current study, we examined vocal emotion recognition using speech stimuli instead of nonverbal vocalization in order to enhance social relevance of the stimuli. Indeed, Neves et al. (2021) showed that in contrast to speech stimuli, emotion recognition of nonverbal vocalization stimuli was not associated with socioemotional adjustment ability in children. Future work may examine whether the audiovisual association effects reported here can also be observed in emotion recognition in nonverbal vocalization. In addition, while the talking-head videos from the RAVDESS data set (Livingstone & Russo, 2018) have been validated for emotional validity, the meaning of the speech content used in the stimuli "kids are talking by the door" was assumed to have a neutral valence without being verified by human rating. To rule out the possible influence from semantic processing of the speech, future work may consider using pseudolinguistic sentences such as those used in the Emotion Recognition in Multiple Modalities test (Laukka et al., 2021) or the Geneva Emotion Recognition Test (Schlegel et al., 2014; Schlegel & Scherer, 2016).

One limitation of the current study was that in order to examine how emotional voices change participants' eye-movement pattern for viewing (neutral) faces, in the audio-only condition, a static neutral face was presented together with emotional voices. Thus, it involved multimodal inputs, although the visual input did not provide information about the emotion. In contrast, the video-only condition involved unimodal input since no voice stimulus was presented. Thus, the observed performance difference between the audiovisual and audio-only condition did not reflect difference between multimodal versus unimodal processing. Rather, it reflected

difference when visual information provided useful versus neutral information. Thus, our results were not able to be directly compared with previous studies examining performance difference between multimodal versus unimodal conditions. Another limitation of the study was that participants were asked to respond as soon as they recognized the emotion in order to analyze the eye-movement data that were relevant to their emotion recognition response. Thus, we were not able to control for the amount of time participants viewed the stimuli in the analysis of emotion recognition accuracy and eye-movement pattern. Future work may examine whether similar results can be obtained when participants are given a fixed stimulus viewing time.

In conclusion, here we show that audiovisual information processing in emotion recognition depends on the automaticity of audiovisual association of the emotion. For emotions with strong association, information in one modality may activate associated information in the other modality, leading to a weaker audiovisual advantage and similar eye-movement patterns for viewing faces even when diagnostic features were only available in the auditory modality. In contrast, for emotions with weak association such as disgust, although they typically involve larger audiovisual advantages, concurrent vocal information may interfere with online eye-movement planning for diagnostic facial information, and those whose eye-movement behavior is affected less can benefit more from concurrent vocal information. This finding not only informs differential audiovisual information processing in the recognition of different emotions but also has important implications for ways to enhance learning in audiovisual/multimedia environments.

## References

- Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, 15(9), 839–843. <https://doi.org/10.1016/j.cub.2005.03.046>
- An, J., & Hsiao, J. H. (2021). Modulation of mood on eye movement and face recognition performance. *Emotion*, 21(3), 617–630. <https://doi.org/10.1037/emo0000724>
- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, 31(1), 5–17. <https://doi.org/10.1023/A:1005653411471>
- Barton, J. J. S., & Corrow, S. L. (2016). Recognizing and identifying people: A neuropsychological review. *Cortex*, 75, 132–150. <https://doi.org/10.1016/j.cortex.2015.11.023>
- Braga, R. M., Fu, R. Z., Seemungal, B. M., Wise, R. J., & Leech, R. (2016). Eye movements during auditory attention predict individual differences in dorsal attention network activity. *Frontiers in Human Neuroscience*, 10, Article 164. <https://doi.org/10.3389/fnhum.2016.00164>
- Caridakis, G., Castellano, G., Kessous, L., Raouzaoui, A., Malatesta, L., Asteriadis, S., & Karpouzis, K. (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. In C. Boukiss, A. Pnevmatikakis, & L. Polymenakos (Eds.), *Artificial intelligence and innovations 2007: From theory to applications. AIAI 2007. IFIP The International Federation for Information Processing* (Vol. 247, pp. 375–388). Springer. [https://doi.org/10.1007/978-0-387-74161-1\\_41](https://doi.org/10.1007/978-0-387-74161-1_41)
- Chan, C. Y. H., Chan, A. B., Lee, T. M. C., & Hsiao, J. H. (2018). Eye-movement patterns in face recognition are associated with cognitive decline in older adults. *Psychonomic Bulletin & Review*, 25(6), 2200–2207. <https://doi.org/10.3758/s13423-017-1419-0>



- Chan, F. H. F., Barry, T. J., Chan, A. B., & Hsiao, J. H. (2020). Understanding visual attention to face emotions in social anxiety using hidden Markov models. *Cognition and Emotion*, 34(8), 1704–1710. <https://doi.org/10.1080/02699931.2020.1781599>
- Chan, F. H. F., Jackson, T., Hsiao, J. H., Chan, A. B., & Barry, T. J. (2020). The interrelation between interpretation biases, threat expectancies and pain-related attentional processing. *European Journal of Pain*, 24(10), 1956–1967. <https://doi.org/10.1002/ejp.1646>
- Chan, F. H. F., Suen, H., Chan, A. B., Hsiao, J. H., & Barry, T. J. (2022). The effects of attentional and interpretation biases on later pain outcomes among younger and older adults: A prospective study. *European Journal of Pain*, 26(1), 181–196. <https://doi.org/10.1002/ejp.1853>
- Chan, F. H. F., Suen, H., Hsiao, J. H., Chan, A. B., & Barry, T. J. (2020). Interpretation biases and visual attention in the processing of ambiguous information in chronic pain. *European Journal of Pain*, 24(7), 1242–1256. <https://doi.org/10.1002/ejp.1565>
- Chan, S. K. W., Hsiao, J., Wong, A. O. Y., Liao, Y., Suen, Y., Yan, E. W. C., Poon, L.-T., Siu, M. W., Hui, C. L. M., Chang, W. C., Lee, E. H. M., & Chen, E. Y. H. (2022). Explicit and implicit mentalization of patients with first-episode schizophrenia: A study of self-referential gaze perception with eye movement analysis using hidden Markov models. *European Archives of Psychiatry and Clinical Neuroscience*. Advance online publication. <https://doi.org/10.1007/s00406-022-01383-y>
- Cho, V., Hsiao, J. H., Chan, A. B., Ngo, H., King, N., & Anthonappa, R. (2022a). Eye movement analysis of children's attention for midline diastema. *Scientific Reports*, 12, Article 7462. <https://doi.org/10.1038/s41598-022-11174-z>
- Cho, V., Hsiao, J. H., Chan, A. B., Ngo, H., King, N. M., & Anthonappa, R. (2022b). Understanding children's attention to dental caries through eye-tracking. *Carries Research*, 56(2), 129–137.
- Cho, V., Hsiao, J. H., Chan, A. B., Ngo, H., King, N. M., & Anthonappa, R. (2022c). Understanding children's attention to traumatic dental injuries using eye-tracking. *Dental Traumatology*. Advance online publication. <https://doi.org/10.1111/edt.12751>
- Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *Journal of Vision*, 14(11), Article 8. <https://doi.org/10.1167/14.11.8>
- Chuk, T., Chan, A. B., & Hsiao, J. H. (2017). Is having similar eye movement patterns during face learning and recognition beneficial for recognition performance? Evidence from hidden Markov modeling. *Vision Research*, 141, 204–216. <https://doi.org/10.1016/j.visres.2017.03.010>
- Chuk, T., Chan, A. B., Shimojo, S., & Hsiao, J. H. (2020). Eye movement analysis with switching hidden Markov models. *Behavior Research Methods*, 52(3), 1026–1043. <https://doi.org/10.3758/s13428-019-01298-y>
- Chuk, T., Crookes, K., Hayward, W. G., Chan, A. B., & Hsiao, J. H. (2017). Hidden Markov model analysis reveals the advantage of analytic eye movement patterns in face recognition across cultures. *Cognition*, 169, 102–117. <https://doi.org/10.1016/j.cognition.2017.08.003>
- Connolly, K. (2014). Multisensory perception as an associative learning process. *Frontiers in Psychology*, 5, Article 1095. <https://doi.org/10.3389/fpsyg.2014.01095>
- Coviello, E., Chan, A. B., & Lanckriet, G. R. (2014). Clustering hidden Markov models with variational HEM. *Journal of Machine Learning Research*, 15(1), 697–747. <https://arxiv.org/abs/1210.6707>
- Davis, J., McKone, E., Zirnsak, M., Moore, T., O'Keamey, R., Apthorp, D., & Palermo, R. (2017). Social and attention-to-detail subclusters of autistic traits differentially predict looking at eyes and face identity recognition ability. *British Journal of Psychology*, 108(1), 191–219. <https://doi.org/10.1111/bjop.12188>
- de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, 14(3), 289–311. <https://doi.org/10.1080/026999300378824>
- Eckhardt, A., Hsieh, J. J., Chan, A., Maier, C., Chuk, T., Siao, J., & Buettner, R. (2013). Objective measures of IS usage behavior under conditions of experience and pressure using eye fixation data. *Proceedings of the 34th International Conference on Information Systems (ICIS)* (Vol. 3, pp. 2715–2731). Association for Information Systems (AIS).
- Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, 164(3875), 86–88. <https://doi.org/10.1126/science.164.3875.86>
- English, M. C. W., Gignac, G. E., Visser, T. A. W., Whitehouse, A. J. O., & Maybery, M. T. (2020). A comprehensive psychometric analysis of autism-spectrum quotient factor models using two large samples: Model recommendations and the influence of divergent traits on total-scale scores. *Autism Research*, 13(1), 45–60. <https://doi.org/10.1002/aur.2198>
- Giovagnoli, A. R., Del Pesce, M., Mascheroni, S., Simoncelli, M., Laiacina, M., & Capitani, E. (1996). Trail Making Test: Normative values from 287 normal adult controls. *Italian Journal of Neurological Sciences*, 17(4), 305–309. <https://doi.org/10.1007/BF01997792>
- Green, C., & Guo, K. (2018). Factors contributing to individual differences in facial expression categorisation. *Cognition and Emotion*, 32(1), 37–48. <https://doi.org/10.1080/02699931.2016.1273200>
- Hsiao, J. H., An, J., Zheng, Y., & Chan, A. B. (2021). Do portrait artists have enhanced face processing abilities? Evidence from hidden Markov modeling of eye movements. *Cognition*, 211, Article 104616. <https://doi.org/10.1016/j.cognition.2021.104616>
- Hsiao, J. H., Chan, A. B., An, J., Yeh, S.-L., & Jingling, L. (2021). Understanding the collinear masking effect in visual search through eye tracking. *Psychonomic Bulletin & Review*, 28(6), 1933–1943. <https://doi.org/10.3758/s13423-021-01944-7>
- Hsiao, J. H., & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological Science*, 19(10), 998–1006. <https://doi.org/10.1111/j.1467-9280.2008.02191.x>
- Hsiao, J. H., Lan, H., Zheng, Y., & Chan, A. B. (2021). Eye movement analysis with hidden Markov models (EMHMM) with co-clustering. *Behavior Research Methods*, 53(6), 2473–2486. <https://doi.org/10.3758/s13428-021-01541-5>
- Hsiao, J. H., Liao, W., & Tso, R. V. Y. (2022). Impact of mask use on face recognition: An eye-tracking study. *Cognitive Research: Principles and Implications*, 7(1), Article 32. <https://doi.org/10.1186/s41235-022-00382-w>
- Hsiao, J. H., & Liu, T. T. (2012). The optimal viewing position in face recognition. *Journal of Vision*, 12(2), Article 22. <https://doi.org/10.1167/12.2.22>
- Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*. Advance online publication. <https://doi.org/10.1177/2515245920951503>
- Lan, H., Liu, Z., Hsiao, J. H., Yu, D., & Chan, A. B. (2021). Clustering hidden Markov models with variational Bayesian hierarchical EM. *IEEE Transactions on Neural Networks and Learning Systems*. Advance online publication. <https://doi.org/10.1109/TNNLS.2021.3105570>
- Lau, E. Y. Y., Eskes, G. A., Morrison, D. L., Rajda, M., & Spurr, K. F. (2010). Executive function in patients with obstructive sleep apnea treated with continuous positive airway pressure. *Journal of the International Neuropsychological Society*, 16(6), 1077–1088. <https://doi.org/10.1017/S1355617710000901>
- Laukka, P., Bänziger, T., Israelsson, A., Cortes, D. S., Tornberg, C., Scherer, K. R., & Fischer, H. (2021). Investigating individual differences in emotion recognition ability using the ERAM test. *Acta Psychologica*, 220, Article 103422. <https://doi.org/10.1016/j.actpsy.2021.103422>
- Lee, H. H., Chen, Z. L., Yeh, S. L., Hsiao, J. H., & Wu, A. A. (2021). When eyes wander around: Mind-wandering as revealed by eye movement analysis with hidden Markov models. *Sensors*, 21(22), Article 7569. <https://doi.org/10.3390/s21227569>
- Liao, W., Li, S. T. K., & Hsiao, J. H. (2022). Music reading experience modulates eye movement pattern in English reading but not in Chinese reading. *Scientific Reports*, 12, 9144. <https://doi.org/10.1038/s41598-022-12978-9>



- Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE*, 13(5), Article e0196391. <https://doi.org/10.1371/journal.pone.0196391>
- Nanay, B. (2018). Multimodal mental imagery. *Cortex*, 105, 125–134. <https://doi.org/10.1016/j.cortex.2017.07.006>
- Neves, L., Martins, M., Correia, A. I., Castro, S. L., & Lima, C. F. (2021). Associations between vocal emotion recognition and socio-emotional adjustment in children. *Royal Society Open Science*, 8(11), Article 211412. <https://doi.org/10.1098/rsos.211412>
- Nie, N. H., Bent, D. H., & Hull, C. H. (1975). *SPSS: Statistical package for the social sciences* (Vol. 227). McGraw-Hill.
- Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive behavioural measurements. *Advances in Methods and Practices in Psychological Science*, 2(4), 378–395. <https://doi.org/10.1177/2515245919879695>
- Phillips, L. H., Wynn, V. E., McPherson, S., & Gilhooly, K. J. (2001). Mental planning and the Tower of London task. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 54(2), 579–597. <https://doi.org/10.1080/17470210000000000>
- Phillips, M. L., Senior, C., Fahy, T., & David, A. S. (1998). Disgust—The forgotten emotion of psychiatry. *The British Journal of Psychiatry*, 172(5), 373–375. <https://doi.org/10.1192/bjp.172.5.373>
- Reitan, R. M. (1958). The validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills*, 8(3), 271–276. <https://doi.org/10.2466/pms.1958.8.3.271>
- Ridderinkhof, K. R., Band, G. P. H., & Logan, D. (1999). A study of adaptive behavior: Effects of age and irrelevant information on the ability to inhibit one's actions. *Acta Psychologica*, 101, 315–337. [https://doi.org/10.1016/S0001-6918\(99\)00010-4](https://doi.org/10.1016/S0001-6918(99)00010-4)
- Rottman, J. (2014). Evolution, development, and the emergence of disgust. *Evolutionary Psychology*, 12(2), 417–433. <https://doi.org/10.1177/147470491401200209>
- Şahin, M. D., & Aybek, E. C. (2019). Jamovi: An easy to use statistical software for the social scientists. *International Journal of Assessment Tools in Education*, 6(4), 670–692. <https://doi.org/10.21449/ijate.661803>
- Schelinski, S., & von Kriegstein, K. (2019). The relation between vocal pitch and vocal emotion recognition abilities in people with autism spectrum disorder and typical development. *Journal of Autism and Developmental Disorders*, 49(1), 68–82. <https://doi.org/10.1007/s10803-018-3681-z>
- Schlegel, K., Grandjean, D., & Scherer, K. R. (2014). Introducing the Geneva Emotion Recognition Test: An example of Rasch-based test development. *Psychological Assessment*, 26(2), 666–672. <https://doi.org/10.1037/a0035246>
- Schlegel, K., & Scherer, K. R. (2016). Introducing a short version of the Geneva Emotion Recognition Test (GERT-S): Psychometric properties and construct validation. *Behavior Research Methods*, 48(4), 1383–1392. <https://doi.org/10.3758/s13428-015-0646-4>
- Schurigin, M. W., Nelson, J., Iida, S., Ohira, H., Chiao, J. Y., & Franconeri, S. L. (2014). Eye movements during emotion recognition in faces. *Journal of Vision*, 14(13), Article 14. <https://doi.org/10.1167/14.13.14>
- Schyns, P. G., Petro, L. S., & Smith, M. L. (2007). Dynamics of visual information integration in the brain for categorizing facial expressions. *Current Biology*, 17(18), 1580–1585. <https://doi.org/10.1016/j.cub.2007.08.048>
- Scott, S. K., Young, A. W., Calder, A. J., Hellawell, D. J., Aggleton, J. P., & Johnson, M. (1997). Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature*, 385(6613), 254–257. <https://doi.org/10.1038/385254a0>
- Seitz, A. R., Kim, R., van Wassenhove, V., & Shams, L. (2007). Simultaneous and independent acquisition of multisensory and unisensory associations. *Perception*, 36(10), 1445–1453. <https://doi.org/10.1068/p5843>
- Skuse, D. (2003). Fear recognition and the neural basis of social cognition. *Child and Adolescent Mental Health*, 8(2), 50–60. <https://doi.org/10.1111/1475-3588.00047>
- Smith, E. G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology and Psychiatry*, 48(8), 813–821. <https://doi.org/10.1111/j.1469-7610.2007.01766.x>
- Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmuting and decoding facial expressions. *Psychological Science*, 16(3), 184–189. <https://doi.org/10.1111/j.0956-7976.2005.00801.x>
- Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9(2), 164–182. <https://doi.org/10.1037/1082-989X.9.2.164>
- Stoet, G., O'Connor, D., Conner, M., & Laws, K. (2013). Are women better than men at multi-tasking? *BMC Psychology*, 1(1), Article 18. <https://doi.org/10.1186/2050-7283-1-18>
- Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, 14(9), 400–410. <https://doi.org/10.1016/j.tics.2010.06.008>
- Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5), 1053–1065. <https://doi.org/10.1037/0096-1523.34.5.1053>
- van Ee, R., van Boxtel, J. J., Parker, A. L., & Alais, D. (2009). Multisensory congruency as a mechanism for attentional control over perceptual selection. *The Journal of Neuroscience*, 29(37), 11641–11649. <https://doi.org/10.1523/JNEUROSCI.0873-09.2009>
- Widen, S. C., & Russell, J. A. (2008). Children's and adults' understanding of the "disgust face." *Cognition and Emotion*, 22(8), 1513–1541. <https://doi.org/10.1080/02699930801906744>
- Widen, S. C., & Russell, J. A. (2013). Children's recognition of disgust in others. *Psychological Bulletin*, 139(2), 271–299. <https://doi.org/10.1037/a0031640>
- Young, A. W. (2018). Faces, people and the brain: The 45th Sir Frederic Bartlett lecture. *The Quarterly Journal of Experimental Psychology*, 71(3), 569–594. <https://doi.org/10.1177/1747021817740275>
- Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences. *Trends in Cognitive Sciences*, 24(5), 398–410. <https://doi.org/10.1016/j.tics.2020.02.001>
- Zhang, J., Chan, A. B., Lau, E. Y. Y., & Hsiao, J. H. (2019). Individuals with insomnia misrecognize angry faces as fearful faces while missing the eyes: An eye-tracking study. *Sleep*, 42(2), zsy220. <https://doi.org/10.1093/sleep/zsy220>
- Zheng, Y., & Hsiao, J. H. (2020). Audiovisual information processing in emotion recognition: An eye tracking study. In S. M. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd annual conference of the cognitive science society* (pp. 2024–2030). Cognitive Science Society.
- Zheng, Y., Chen, D., Hu, X., & Hsiao, J. H. (2022). The impact of mask use on social categorization. In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th annual meeting of the cognitive science society* (pp. 578–585). Cognitive Science Society.
- Zheng, Y., Ye, X., & Hsiao, J. H. (2022). Does adding video and subtitles to an audio lesson facilitate its comprehension? *Learning and Instruction*, 77, Article 101542. <https://doi.org/10.1016/j.learninstruc.2021.101542>

Received September 21, 2021

Revision received May 24, 2022

Accepted May 27, 2022 ■