Audiovisual Information Processing in Emotion Recognition: An Eye Tracking Study

Yueyuan Zheng (u3514160@connect.hku.hk)
Department of Psychology, University of Hong Kong
Pokfulam Road, Hong Kong

Janet H. Hsiao (jhsiao@hku.hk)
Department of Psychology and the State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong
Pokfulam Road, Hong Kong

Abstract

In audiovisual information processing, auditory information may interfere with eye movement planning in visual processing due to competition for attentional resources. Here we hypothesize that this interference may be mitigated in the recognition of emotions involving strong audiovisual coupling. Participants judged the emotion of a talking head video under audiovisual, video-only, and audio-only conditions. While participants generally performed the best in the audiovisual condition, their eye movement pattern did not change significantly across the three conditions except for the recognition of disgust. In disgust recognition, eye movements in the audiovisual condition were less eyes-focused than the video-only condition, and the larger the difference, the less the audiovisual advantage in performance. Disgust recognition develops later in life and may involve weaker audiovisual coupling. Accordingly, our results suggest that whether emotional voice information facilitates emotion recognition without interfering with eye movement planning depends on the strength of audiovisual coupling in emotion processing.

Keywords: emotion recognition; audiovisual processing; facial expression; eye movement; EMHMM

Introduction

Emotion recognition is of vital importance in daily human interaction. It demands both temporal and spatial attention as facial movements during emotional expressions may contain subtle but critical changes for recognition (Young & Bruce, 2011). In addition, both emotional facial and vocal information play an important role. Thus, real-life emotion recognition involves audiovisual processing of dynamic information. However, most of the 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 (e.g. Smith, Cottrell, Gosselin & Schyns, 2005), as reflected in eye movements (e.g., 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, the two sources of information may influence each other, and this interaction is shown to involve attentional mechanisms (e.g., Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010). Eye movement planning is also highly associated with attentional mechanisms (e.g., Noudoost, Chan, Steinmetz & Moore, 2010). Thus, the presence of auditory information may interfere with eye movement planning in visual processing. Indeed, eye movements elicited during an auditory attention task were shown to be predictive of attentional engagement and cued sound location (Braga, Fu, Seemungal, Wise, & Leech, 2016), suggesting shared neural mechanisms between auditory and visual attention systems. Consistent with this finding, Zheng, Ye and Hsiao (2019) 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 better comprehension of the auditory narratives.

Accordingly, in emotion recognition, as compared with using only emotional face stimuli, the addition of emotional voice information may influence participants' eye movement planning. Consequently, they may look less often to diagnostic features for recognition, and their recognition performance may be associated with how well they can attend to diagnostic visual features under the influence of additional emotional voice information.

Note however that recent research has suggested strong audiovisual coupling in emotion recognition. This phenomenon may be because emotional experience can change frequently over time and is multi-modal in nature, resulting in high demands on audiovisual coupling (Young, 2018). For example, incongruent vocal expressions were shown to modulate perception of facial expression and vice versa (De Gelder & Vroomen, 2000). People with facial emotion recognition problems are often also affected in voice emotion recognition, particularly in the recognition of fear (Sprengelmeyer et al., 1999) and anger (Calder et al., 1996; Scott et al., 1997). This finding also suggested that there may be variation in the strength of audiovisual coupling in the recognition of different emotions due to differences in the demand during daily life. Emotions such as fear and anger may involve strong audiovisual coupling due to their relevance to survival (e.g., Skuse, 2003), whereas emotions learned/developed later in life such as disgust (Phillips, Senior, Fahy, & David, 1998) may involve weaker audiovisual coupling. For the recognition of emotions that typically involve strong audiovisual coupling, the interference of vocal information on eye movement

planning for facial information may be mitigated, since the two sources of information are frequently processed together.

Accordingly, here we tested the hypothesis that while the processing of emotional voice information may interfere with eye movement planning for emotional face information due to competition for attentional resources, this interference may be mitigated in the recognition of emotions that involve strong audiovisual coupling. Participants judged emotions of a talking head video expressing different emotions in audiovisual, video-only (without voice information), and audio-only (with a static neutral face image) conditions with eye tracking. We used the Eye Movement analysis with Hidden Markov Models (EMHMM, Chuk, Chan, & Hsiao, 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 audiovisual conditions. We expected that while participants would have better performance in the audiovisual condition in general due to the availability of more information, for emotions involving strong audiovisual coupling such as fear and anger, vocal information will enhance performance without influencing eye movement planning toward diagnostic facial features. In contrast, when the strength of audiovisual coupling is weak 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.

Method

Participants

65 participants[1] (44 females and 21 males) between 17 to 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 cognitive disabilities or psychological problems.

Materials and Apparatus

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 was about 5 seconds long. The 432 video clips were divided evenly into three audiovisual conditions: 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 were displayed. In each condition, clips of 24 performers acting out six categories of emotion were used, including happy, sad, angry, fearful, disgusted, and surprised (the emotion hexagon; Calder et al., 1996; Figure 1), summing up to 144 slips per condition. 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° of visual angle (following Hsiao & Cottrell, 2008), with the nose aligned with the center of the screen. 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).

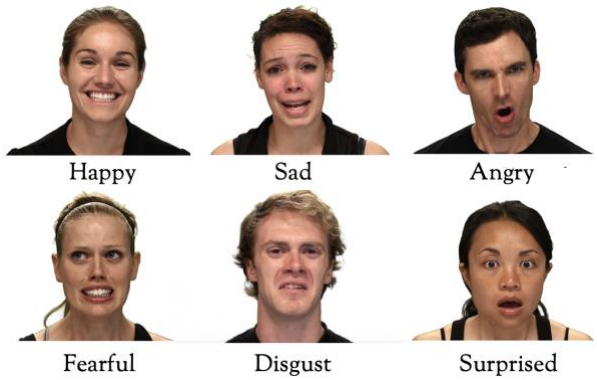


Figure 1: Video captures of six emotions from RAVDESS

EyeLink 1000 plus (tower mount model; SR Research) was used to record eye movements. The sampling rate was 1000 Hz and the resolution of the monitor was 1280 x 1024 pixels. A Cedrus response box was used to collect behavioral responses.

Design

The design consisted of two within-subject variables: audiovisual 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. In a separate analysis, we examined what factors, including eye movement and cognitive ability measures, could predict the advantage of the audiovisual condition over the video-only or audio-only condition through correlation and regression analyses.

Procedures

Participants performed an emotion recognition task, followed by cognitive ability tests including verbal and visuospatial two-back tasks for working memory capacity, Tower of London test for executive function/planning ability, multitasking test for task-switching ability, trail making test for visual attention and switching ability, and

[1] A power analysis of repeated measures ANOVA with 3 measurements (i.e., the 3 audiovisual conditions) assuming a small to medium effect size (f = .17, power = .80, α = .05) showed that the required sample size was 58.

flanker test for selective attention ability. These cognitive ability tests were included to examine what factors, including cognitive abilities and online eye movement behavior, could best predict the advantage of the audiovisual condition over the video-only or the audio-only condition.

In emotion recognition, the 432 videos were presented in a random order in 12 blocks with 36 trials each. 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. Participants were asked to look at the fixation cross when it appeared. The cross lasted for 500 ms and then the video clip was presented. The video clip presentation was followed by a 500 ms blank screen. Participants were asked to judge the emotion of the video clip from the 6 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. Accuracy and reaction time (RT) were 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 RT were measured. Each task had 52 trials.

In the Tower of London test (Phillips, Wynn, McPherson, & Gilhooly, 2001), participants moved 3 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.

In the multitasking test (Stoet, O'connor, Conner, & Laws, 2013), 4 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 2500 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 test (Ridderinkhof, Band, & Logan, 1999), participants judged the direction of an arrow flanked by 4 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 non-directional symbols.

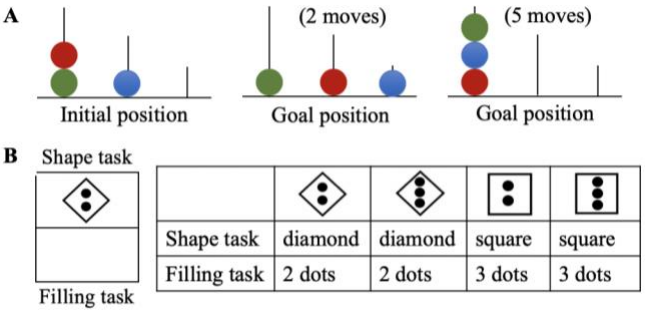


Figure 2: Cognitive tests: A. Examples of the Tower of London test. B. Stimuli used in the multitasking test.

Eye Movement Data Analysis

EMHMM (Chuk, Chan, & Hsiao, 2014) was used to analyze eye movement data. Eye movement data were first normalized according to the center point between the two eyes across videos. A participant's eye movements in each of the audiovisual condition and emotion combinations were summarized using a hidden Markov model (HMM, a type of time-series statistical model in machine learning). The resulting 1170 (18 models x 65 participants) individual models were then clustered to discover two representative patterns. The similarities of individual eye movement patterns to the two representative patterns then were quantified using the log-likelihoods of the data being generated by the representative models (e.g., Chuk, Crookes, Hayward, Chan, & Hsiao, 2017). A similar ANOVA analysis was conducted with the dependent variable being the log-likelihood measures.

Results

In emotion recognition accuracy, there was a main effect of audiovisual condition, $F(2, 128) = 429.4$, $p < 0.001$, $\eta_2 = 0.149$: participants had higher accuracy in the audiovisual than the video-only condition, $t(64) = 17.7$, $p < 0.001$, $d = 2.19$, and in the video-only than the audio-only condition, $t(64) = 14.2$, $p < 0.001$, $d = 1.76$. A significant main effect of emotion was also found, $F(5, 320) = 66.5$, $p < 0.001$, $\eta_2 = 0.226$. People had the best performance in recognizing anger, followed by sadness, happiness and surprise, and disgust. They performed the worst in recognizing fear. Importantly, there was an interaction between audiovisual condition and emotion, $F(10, 640) = 52.7$, $p < 0.001$, $\eta_2 = 0.111$ (Figure 3). For happiness, participants' performance did not differ between the audiovisual and video-only conditions, $t(64) = -1.74$, $p = 0.087$, but was higher in the video-only than audio-only condition, $t(64) = 18.78$, $p < 0.001$, $d = 2.33$. This suggested that they mainly relied on visual information for the recognition of happiness. For disgust, people were significantly more accurate in the audiovisual than video-

only condition, t(64) = 11.10, p < 0.001, d = 1.38, and in the video-only than audio-only condition, t(64) = 7.25, p < 0.001, d = .90. 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.

In eye movement data analysis, we discovered two representative eye movement patterns as the result of clustering: the nose-focused and eyes-focused patterns (Figure 4). This finding was consistent with a previous EMHMM study on emotion recognition using static face images (Zhang, Chan, Lau, & Hsiao, 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 afterwards (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 eye region/red ROI, and remained looking at the same region afterwards. Occasionally (6%) they started from the left eye/green ROI and remained there afterwards (94%). The two representative HMMs differed significantly (Chuk et al., 2014): data from those using the nose-focused pattern were more likely to be generated from the nose-focused than eyes-focused HMM, t(446) = 17.08, p < .001, d = 0.81, and data from those with the eyes-focused pattern were more likely to be generated from the eyes-focused than nose-focused HMM, t(49.47) = 1.892, p = < .001, d = 1.84.

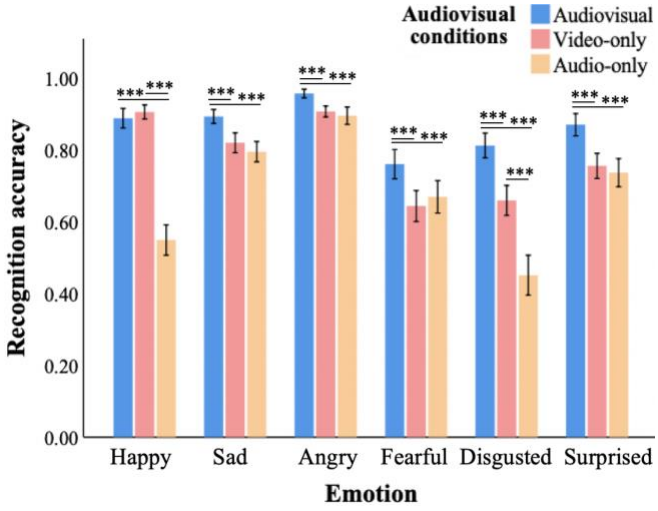
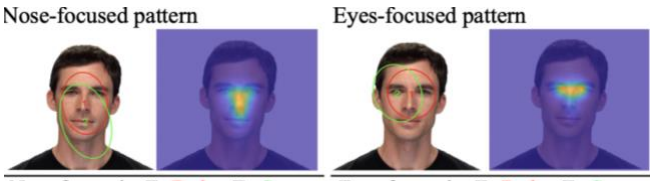


Figure 3: Emotion recognition accuracy in different conditions (error bars: 95% CI; ***p ≤ 0.001).



Nose-focused	To Red	To Green
Priors	.99	.01
From Red	.97	.03
From Green	.00	1.0

Eyes-focused	To Red	To Green
Priors	.94	.06
From Red	1.0	.00
From Green	.06	.94

Figure 4: The nose-focused (left) and eyes-focused (right) patterns. Ellipses show ROIs as 2-D Gaussian emissions. The table shows transition probabilities among the ROIs. Priors show the probabilities that a fixation sequence starts from the ellipse. The image on the right shows the corresponding heatmap.

Following previous studies (e.g., Chan, Chan, Lee, & Hsiao, 2018), we quantified participants' eye movement pattern using the Nose-Eyes scale (N-E scale) as

$$\text{Nose-Eyes scale} = \frac{N - E}{|N| + |E|}$$

Where N is the log-likelihood of the participant's eye movement data being generated by the nose-focused HMM, and E is the log-likelihood of the participant's data being generated by the eyes-focused pattern. This log-likelihood measure reflects the similarity of the participant's eye movement to the representative pattern. A more positive N-E scale indicates higher similarity to the nose-focused pattern, whereas a more negative value indicated higher similarity to the eyes-focused pattern.

In N-E scale, there was a main effect of emotion, F(5, 320) = 37.43, p < 0.001, η_2 = 0.009. 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 effect interacted with audiovisual condition, F(10, 640) = 31.08, p < 0.001, η_2 = 0.012 (Figure 5). Interestingly, for the recognition of happiness, sadness, anger, fear and surprise, no significant difference was observed among the 3 audiovisual conditions, ps > 0.05. In contrast, for disgust, eye movement pattern in the audio-only condition was more nose-focused than the audiovisual condition, t(64) = 3.37, p = 0.001, d = 0.42, and that in the video-only condition was more eyes-focused than the audiovisual condition, t(64) = -10.49, p < 0.001, d = -1.30. This result was consistent with our hypothesis that for emotions with strong audiovisual coupling, additional vocal information facilitates recognition without interfering with visual attention to diagnostic facial features, whereas for emotions with weak coupling such as disgust, adding voice information makes eye movements focus less on the diagnostic eye region.

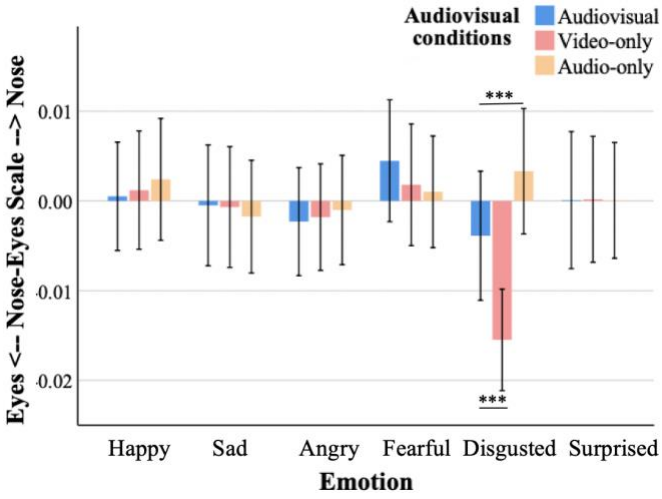


Figure 5: Nose-eyes scale in different conditions (error bars: 95% CI; ***p ≤ 0.001).

We then examined whether the eye movement pattern changes among the audiovisual conditions in disgust recognition were associated with changes in recognition performance. We defined normalized change in performance/eye movement pattern as

$$\text{Normalized change} = \frac{A - B}{A + B}$$

Where A and B refer to performance/eye movement pattern in different conditions. Stepwise multiple regression analysis predicting normalized change in accuracy between the audiovisual and video-only conditions using normalized change in eye movement pattern (N-E scale) between the two conditions and all cognitive task performance measures showed that normalized change in N-E scale was the only significant predictor, β = -.322, p = 0.009, accounting for a significant portion of variance, R^2 = .104, F(1,63) = 7.286, p = .009. It indicated the larger the accuracy increase in the audiovisual condition, the less the eye movement pattern change. A similar stepwise regression analysis predicting normalized change in accuracy between the audiovisual and audio-only conditions showed that execution time of the Tower of London test was the only significant predictor, β = .253, p = 0.042, accounting for a significant portion of variance, R^2 = .064, F(1,63) = 4.322, p = .042. It suggested the larger the accuracy increase in the audiovisual condition, the lower the executive function ability. As the recognition accuracy data (Figure 3) suggested that audio information was less informative than visual information in the recognition of disgust, those who had lower executive function ability 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 coupling due to its multi-modal nature and high demands on accuracy and efficiency (Young, 2018), and the recognition of different emotions may differ in the strength of audiovisual coupling (Sprengelmeyer et al., 1999). Accordingly, here we tested the hypothesis that in audiovisual emotion recognition, vocal information may interfere with eye movement planning for facial information due to competition for attentional resources in emotions with weak audiovisual coupling, and the performance depends on the amount of interference. In contrast, this interference may be reduced in the recognition of emotions with strong audiovisual coupling.

Our results showed that while participants had the best performance in the audiovisual condition in general, their eye movement pattern did not change significantly across the three audiovisual conditions in the recognition of happy, sad, angry, fearful and surprised expressions. This result suggested that concurrent vocal information improved 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 due to strong audiovisual coupling. 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 coupling in the recognition of fear and anger (e.g., Sprengelmeyer et al., 1999). The current results further demonstrated strong audiovisual coupling in the recognition of happiness, sadness, and surprise.

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 most of the diagnostic features for disgust recognition is around the eye region (Phillips et al., 1998), 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 the cognitive ability measures controlled: the less the pattern chance, the more the performance increase. In other words, those whose online eye movement behavior was affected the least benefitted the most from concurrent vocal information. In contrast, the performance increase in the audiovisual relative to audio-only condition was best predicted by executive function ability instead of eye movement pattern

change: those who had low executive function ability benefited more with the addition of visual information, which was more informative than auditory information in emotion recognition.

Among the six basic emotions, disgust is learned and developed the latest in life (Phillips et al., 1998). Thus, disgust recognition may involve weaker audiovisual coupling than the other emotions, resulting in the observed audiovisual effect. In addition, here we used speech stimuli with emotional voice, which differed from the typical diagnostic vocalizations of disgust such as 'yuk!' and 'ugh!' (Phillips et al., 1998). This difference may have created a scenario with weak audiovisual coupling in emotion recognition to reveal its influence on eye movement pattern and performance. Future work will examine this possibility.

The current results suggested that the strength of audiovisual coupling modulates eye movements and performance in emotion recognition. This finding has important implications for audiovisual information processing tasks in general. For example, person identification is argued to have weaker audiovisual coupling than emotion recognition, since face and voice identities do not change over time and are often identified separately (Young, 2018). 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 (e.g., Barton & Corow, 2016). Accordingly, similar to the recognition of disgust, 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 coupling, such as auditory narratives and visual subtitles, typically facilitate learning, whereas those with weak coupling may compete for attentional resources, and the performance may depend on one's online information extraction strategy as revealed in eye movement behavior (Zheng et al., 2019). It remains unclear what cognitive abilities are associated with being less interfered by concurrent auditory information in eye movement planning, as none of the cognitive ability measures used here could predict participants' eye movement pattern change between the audiovisual and video-only conditions. It may be related to auditory working memory or other executive functions not measured here, and this requires further investigations.

In conclusion, here we show that audiovisual information processing in emotion recognition depends on the strength of audiovisual coupling of the emotion. For emotions with strong coupling, vocal information facilitates recognition without interfering with eye movement planning for facial information. In contrast, for emotions with weak coupling such as disgust, concurrent vocal information may interfere with online eye movement planning for 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.

Acknowledgments

We are grateful to RGC of Hong Kong (Project # 17609117 to Hsiao). We thank Professor Andy Young for his suggestions and support during this research.

References

Barton, J. J. S., & Corrow, S. L. (2016). Recognizing and identifying people: A neuropsychological review. *Cortex, 75*, 132-150.

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. *Front. Hum. Neurosci., 10*(2016), 164.

Calder, A. J., Young, A. W., Rowland, D., Perrett, D. I., Hodges, J. R., & Etcoff, N. L. (1996). Facial emotion recognition after bilateral amygdala damage: Differentially severe impairment of fear. *Cogn. Neuropsychol., 13*, 699–745.

Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., & Karpouzis, K. (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. *In IFIP Int. C. AIAI*, 375-388. Springer, Boston, MA.

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. *Psychon. Bull. Rev., 25*(6), 2200-2207.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *J. Vision, 14*(11):8, 1-14.

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.

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

Hsiao, J. H., & Cottrell, G. W. (2008). Two fixations suffice in face recognition. *Psychol. Sci., 9*(10), 998-1006.

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. *J. Int. Neuropsych. Soc., 16,* 1077- 1088.

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), E0196391.

Nanay, B. (2018). Multimodal mental imagery. *Cortex, 105*, 125-134.

Noudoost, B., Chang, M. H., Steinmetz, N. A., & Moore, T. (2010). Top-down control of visual attention. *Curr. Opin. Neurobiol., 20*(2), 183–190.

Phillips, L.H., Wynn, V.E., McPherson, S., & Gilhooly, K.J. (2001). Mental planning and the Tower of London task. *Q. J. Exp. Psychol. - A, 54*, 579–597.

Phillips, M. L., Senior, C., Fahy, T., & David, A. S. (1998). Disgust - the forgotten emotion of psychiatry. *Brit. J. Psychiat.*, *172*(5), 373-375.

Reitan, R. M. (1958). The validity of the Trail Making Test as an indicator of organic brain damage. *Percept. Mot. Skills, 8*, 271-276.

Ridderinkhof, K.R., Band, G.P., & Logan, D. (1999). A study of adaptive behavior: effects of age and irrelevant information on the ability to inhibit one's actions. *Acta Psychol.*, *101*, 315–337.

Schurgin, M. W., Nelson, J., Iida, S., Ohira, H., Chiao, J. Y., & Franconeri, S. L. (2014). Eye movements during emotion recognition in faces. *J. Vision*, *14*(13), 14-14.

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*, 254–257.

Skuse, D. (2003). Fear Recognition and the Neural Basis of Social Cognition. *Child Adol. Ment. H-UK*, *8*(2), 50-60.

Smith, M., Cottrell, G., Gosselin, F., & Schyns, P. (2005). Transmitting and Decoding Facial Expressions. *Psychol. Sci.*, *16*(3), 184-189.

Sprengelmeyer, R., Young, A. W., Schroeder, U., Grossenbacher, P. G., Federlein, J., Bu¨ttner, T., & Przuntek, H. (1999). Knowing no fear. *P. Roy. Soc. B.-Biol. Sci*, *266*, 2451–2456.

Stoet, G., O'connor, D., Conner, M., & Laws, K. (2013). Are women better than men at multi-tasking? *BMC Psychol.*, *1*(1), 18.

Talsma, D., Senkowski, D., Soto-Faraco, S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.*, *14*, 400–410.

Young, A. W. (2018). Faces, people and the brain: The 45th Sir Frederic Bartlett Lecture. *Q. J. Exp. Psychol.*, *71*(3), 569-594.

Young, A. W., & Bruce, V. (2011). Understanding person perception. *Brit. J. Psychol.*, *102*, 959–974.

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.

Zheng, Y., Ye, X., & Hsiao, J. H. (2019). Does video content facilitate or impair comprehension of documentaries? The effect of cognitive abilities and eye movement strategy. In A.K. Goel, C.M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, 1283-1289.