



Emotion-specific modality effects in auditory and visual perception of emotion

Yongseong Lee¹ · Youngju Lee¹ · Dongha Lee¹

Received: 21 March 2026 / Accepted: 25 June 2026

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2026

Abstract

Emotion is typically perceived in dynamic multisensory contexts. Although previous studies have demonstrated multimodal gains in audiovisual recognition (i.e., general enhancement in audiovisual conditions), they have not clarified how visual and auditory information contribute to affective judgments or whether their contribution differs by emotion categories. The current study investigated how emotion perception varies across modalities and examined whether these variations depend on emotion categories. We conducted an emotion perception experiment using audio-only (AO), video-only (VO), and audio-visual (AV) speech stimuli. Participants evaluated the valence (positive/negative) and arousal (weak/strong) of stimuli depicting seven emotions (angry, calm, disgust, fearful, happy, sad, and surprise). Overall, AO stimuli tended to be perceived as more negative in valence and weaker in arousal compared to VO or AV stimuli. In representational similarity analysis, the response patterns for AV were more similar to VO than to AO. Happy and disgust stimuli were rated as weaker in arousal and less positive or less negative in valence in AO condition than in other conditions. For surprise, VO stimuli were rated more negatively than AO and AV. However, angry, fearful, and sad emotions did not differ between modalities. These findings indicate that affective judgments are not solely driven by general enhancement from combined audiovisual cues; rather, the relative importance of sensory modalities varies by emotion, suggesting that specific unimodal cues can be as informative as multimodal information. These modality-specific contributions to emotion perception have implications for clinical assessment, social communication, and affective neuroscience.

Public significance statement

Emotion is typically understood through multisensory information in the real-world; however, it is frequently perceived even in the absence of sensory input, such as when wearing a face mask during the COVID-19 pandemic or during telephone conversations. This study examines the role of sensory input in emotion perception. In situations requiring emotion perception that relies on specific inputs, understanding how the influence of sensory input varies across emotional categories is essential to facilitate more effective communication.

Introduction

Accurate comprehension of others' emotions is essential in social interactions. Emotion perception utilizes various sensory information, such as facial expression and voice. As facial expressions are crucial for accurate emotion perception, most studies of emotion perception have utilized static facial

expressions as emotional stimuli. However, in real life, individuals perceive emotions using multisensory information. For instance, only auditory information is used in phone conversations. Thus, multisensory stimuli should be considered in emotion perception research.

Previous studies of emotional modality have confirmed multimodal gain in emotion recognition (i.e., general enhancement in audiovisual conditions). Multimodal gain occurs when classifying emotional stimuli, as multimodal stimuli lead to better performance than do unimodal stimuli (De Gelder & Vroomen, 2000; Kreifelts et al., 2007; Zheng & Hsiao, 2023). Additionally, multimodal stimuli show greater neural activation in functional magnetic resonance imaging (fMRI) compared to unimodal stimuli, activating the bilateral posterior superior temporal gyrus (pSTG) and right thalamus (Kreifelts et al., 2007).

Yongseong Lee and Youngju Lee contributed equally to this work.

✉ Dongha Lee
donghalee@kbri.re.kr

¹ Cognitive Science Research Group, Korea Brain Research Institute, 61 Cheomdan-ro, Daegu, Dong-gu 41062, Republic of Korea

The influence of emotional modality can vary across emotional categories. Specifically, different emotions are associated with distinctive visual information perceived from diagnostic regions of the face. Differences in eye movements support this hypothesis. Previous studies using facial expressions have demonstrated that different emotions involve different eye movement patterns (Calvo et al., 2014, 2018; Nusseck et al., 2008; Schurgin et al., 2014; Smith et al., 2005; Zheng & Hsiao, 2023). Participants frequently focused on the eye regions of angry, fearful, or sad faces, and motion recognition performance improved when fixations were allocated to the eye region. Conversely, happy, disgusted, and surprised faces were primarily recognized from the mouth and nose regions. These distinct fixation patterns indicate that the diagnostic value and salience of facial features vary substantially across emotion categories. According to the principles of multisensory integration, sensory inputs are optimally weighted by their relative reliability and informativeness for a given task (Campanella & Belin, 2007; Ernst & Banks, 2002). Therefore, given that the diagnostic utility of visual cues varies among different emotions, the relative perceptual weight—or reliance—placed on the visual versus auditory modality may also differ depending on the emotion involved.

Additionally, emotion categories are associated with sensory information to varying degrees. When recognizing emotions from a unimodal stimulus, the other unimodal modalities are recalled. For example, audiovisual associations from either visual or auditory modalities are common (Zheng & Hsiao, 2023). In prior studies, fearful and angry emotions involved in survival had stronger audiovisual associations in emotion recognition, whereas acquired emotions such as disgust had weaker associations (Rottman & Hastie, 2014; Skuse, 2003; Widen & Russell, 2013). A recent study found that audiovisual associations differ across emotions, with weak emotions exhibiting less multimodal gain, indicating the influence of modality on emotion recognition (Zheng & Hsiao, 2023). However, categorical approaches to emotion recognition provide limited insight into how emotions differ across modalities, as such approaches rely on discrete labels and classification accuracy measures. Such approaches cannot specify whether modality-specific differences arise primarily in valence, arousal, or both. Therefore, a dimensional approach is necessary to capture modality-by-emotion interactions quantitatively.

An emotional-dimensional approach, rather than a categorical one based solely on classification accuracy, may help identify how information in each sensory modality contributes to emotion perception and how their integration changes emotion perception. According to the dimensional theory, emotions can be categorized along two orthogonal axes: valence and arousal (Hamann, 2012; Russell, 1980, 2003). Valence ranges from positive to negative, while arousal ranges from weak to strong. This approach can capture emotional details and integrate with neuroscience data (Barrett, 2006; Lee et al., 2023; Posner et al., 2005).

Different sensory modalities may differentially affect each dimension of emotion perception. Moreover, Ekman's basic emotions can be mapped onto this dimensional space (Ekman, 1992; Gu et al., 2019). This indicates that facial expressions classified into basic emotional categories can be sufficiently analyzed using a two-dimensional approach. In this context, we use the term emotion-specific modality effects as a descriptive label for the interaction between sensory modality and emotion category in perceived valence and arousal, consistent with the dimensional approach to emotion.

We utilized multivariate pattern analysis (MVPA) to analyze the dimensional property of emotion. In MVPA, representational similarity analysis (RSA) and multidimensional scaling (MDS) are preferred for comparing and mapping relationships among multi-voxel patterns in fMRI research (Kriegeskorte et al., 2008). Notably, MDS has been used in emotion research to confirm the distribution of emotions along valence and arousal axes (Delacherie et al., 2011; Lee et al., 2023; Russell & Bullock, 1985). RSA enables direct comparisons to be made between various data types, including neural, behavioral responses, computational models, and cross-species, using representational dissimilarity matrices (RDMs). RDMs are created for behavioral data by calculating Euclidean distances between responses. Then, based on the dimensional theory of emotion, RSA, which utilizes the RDMs created using Euclidean distances among emotional states in a two-dimensional space, enables comprehensive and integrative quantitative analysis of emotion perception.

The current study investigated how emotion perception varies quantitatively across sensory modalities. Specifically, sensory modalities differentially shape emotion perception along valence and arousal dimensions, with visual information often contributing strongly to valence-related evaluations and auditory information demonstrating particular relevance for arousal-related perception (Li et al., 2025). Additionally, previous studies suggest that perceiving audiovisual emotions tends to resemble visual-only perception rather than auditory-only perception (Collignon et al., 2008). Based on prior evidence, we hypothesized that emotion perception would differ between visual and auditory modalities, with each sensory modality exerting distinct influences on valence and arousal. In particular, we predicted greater differences for visual information when the perceptual pattern in the visual modality resembled that in the audiovisual stimulus.

Method

Participants

For the current study, we reanalyzed data from the seventy Korean participants (35 women) aged between 20 and 35 years were recruited (mean=25.3 years, standard deviation

(SD)=3.1 years) in our previous study (Lee et al., 2026). All were right-handed and had corrected visual acuity. This study was approved by the Institutional Review Board of the Korea Brain Institute, and all participants provided written informed consent. Following a previous study (Zheng & Hsiao, 2023), power analysis for a repeated-measures ANOVA was conducted in G*power using Cohen's f of 0.9045 (equivalent to an η^2 of 0.45), power 0.95, and alpha 0.05. The analysis suggested 14 participants to capture the interaction effect (Faul et al., 2007). Nevertheless, the sample size was increased to 70 to ensure sufficient statistical power.

Experimental materials

A total of 840 speech stimuli (560 video and 280 audio) were collected from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS; Livingstone & Russo, 2018). Based on the goodness scores that assessed overall emotion recognition (Livingstone & Russo, 2018), we selected the five highest-rated male (actor number: 3, 5, 7, 9, and 19) and female actors (actor number: 2, 6, 8, 12, and 18). The average ratings for male and female actors were 5.5 (SD=0.2) and 6.1 (SD=0.3), respectively. The stimuli included three modalities (AO, VO, and AV) and strong versions of seven emotions

(angry, calm, disgust, fearful, happy, sad, and surprise). Neutral emotions were excluded because they lacked strong versions, and responding to their valence and arousal may be challenging. Some original stimuli initially exceeded 4 s. Therefore, we truncated the ends of these longer stimuli. After editing, all stimuli were checked by a researcher to ensure that the audio and visual signals remained natural and that the emotional expressions were not disrupted. The average duration of the final edited stimulus files was 3.68 s (see <https://osf.io/759a4/>). During the experiment, the presentation time for all stimuli was fixed at exactly 4 s; thus, stimuli shorter than 4 s were extended with silence at the end of the presentation.

The monitor diagonal length was 27 inches, with a resolution of 1920 × 1080 pixels and a refresh rate 60 Hz. A chinrest was utilized to maintain a 57 cm distance between the monitor and the participant's eyes. The video stimulus was presented on the screen at approximately 38.9° × 23.2° in visual angles. The word "Sound" appeared in white at the center of the screen for audio stimuli.

Experimental procedure

Figure 1 illustrates the experimental procedure for each modality. The stimulus was presented for 4 s, followed by

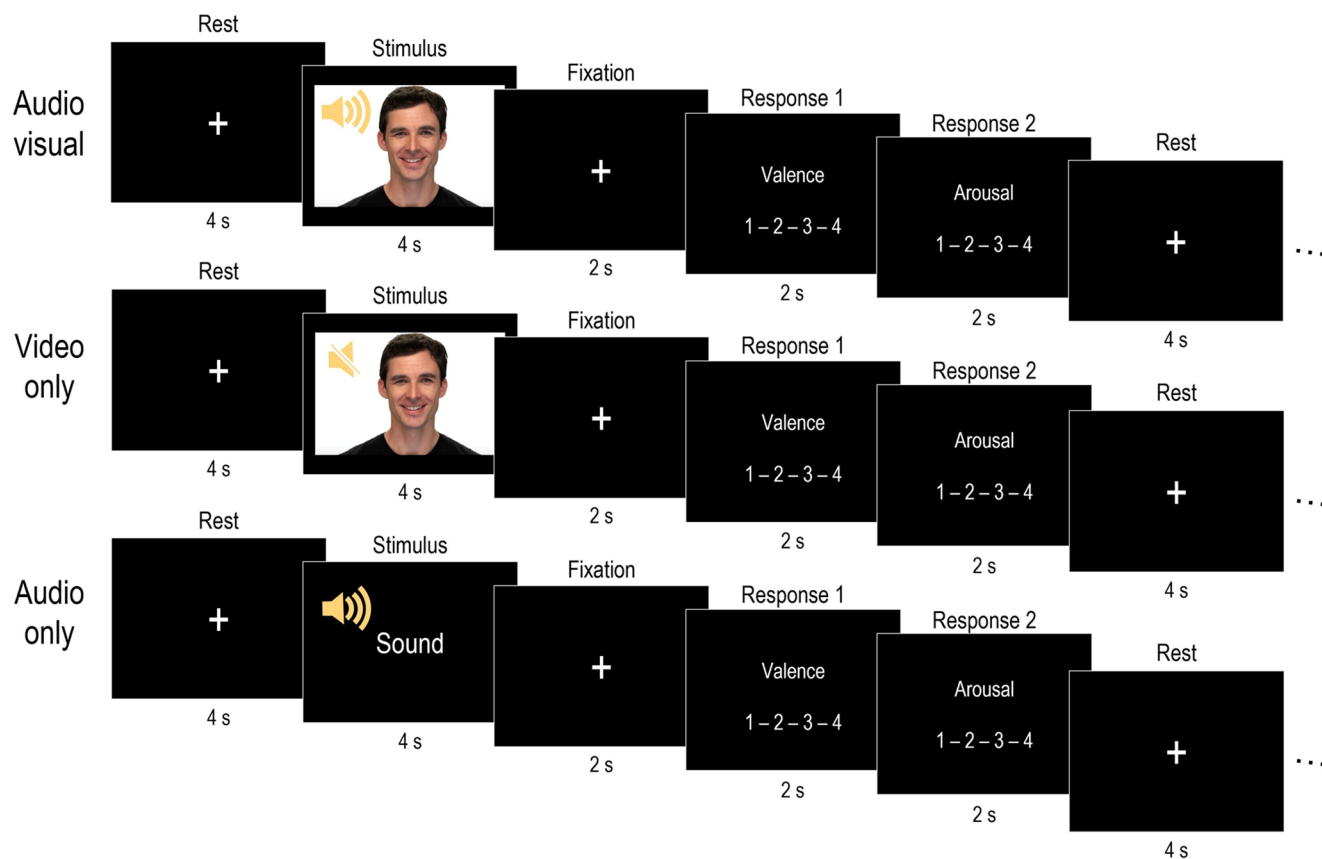


Fig. 1 Experimental trials for each modality

2 s of fixation. Then, the participants evaluated the valence and arousal for 2 s in each phase. The evaluations employed 4-point Likert scales: for valence, response options were “very negative,” “negative,” “positive,” and “very positive,” and for arousal, the options were “very weak,” “weak,” “strong,” and “very strong.” Trials without responses were removed before analysis.

Two actors (one male and one female) were assigned as stimuli for each participant, counterbalancing each block. Modality was presented differently across blocks, resulting in six blocks for each participant. In the modality block, we expected the order effect from AV to AO or AV to VO would be larger than that from AO to AV or VO to AV. Accordingly, participants were assigned to one of two counterbalanced modality-order sequences: AV-first ($N=35$) or AV-last ($N=32$), in which the AV condition appeared either first or last in the task. To empirically check the presence of these anticipated order effects, we conducted exploratory supplementary analyses of these two groups. These analyses revealed a significant order effect in arousal ratings, but not in valence ratings, confirming the necessity of our counterbalanced design (see Supplementary Table S1 for detailed results).

The stimuli for one block included 28 trials, featuring seven emotions and two dialogues (e.g., “Kids are talking by the door,” “Dogs are sitting by the door”) repeated once. This resulted in 168 trials in the experiment. The participants could take sufficient breaks between each block. They were instructed to rate the valence and arousal levels of each stimulus using a 4-point Likert scale during the experiment and completed seven practice trials featuring a single actor portraying all seven emotions; this actor was excluded from the main experiment.

Behavioral analysis of valence and arousal responses

We excluded participants based on their response averages if (1) either the valence or arousal rating was missing in more than 20% of all trials; (2) the same valence or arousal rating was made in more than 90% of all trials; or (3) either valence or arousal average reaction time exceeded 3 SD from the mean across all participants. Three participants were excluded owing to missing data (missing rate: 38%, 41%, and 60%), resulting in 67 participants (32 women, mean age=25.2 years, $SD=3.1$ years) for the analysis. Then, we removed all trials with missing data (1.67% of the total) for either the valence or arousal rating from the 67 participants. Valence and arousal ratings were coded on a scale of 1 to 4, with those close to 1 indicating a more negative and weaker or less intense feelings, respectively.

We performed a repeated-measures analysis of variance (ANOVA) to investigate the differences between modalities based on emotions. Because repeated-measures ANOVA only indicates differences in mean valence and arousal across isolated dimensions, it is challenging to directly determine whether holistic response patterns (e.g., whether AO and VO resemble AV) vary across modalities. Therefore, we conducted representational similarity analysis (RSA) to investigate the relational structure of the entire emotion space (Kriegeskorte et al., 2008).

We modeled participant-level RDMs using the following procedure to define the exact observations entering the matrices: First, we computed each participant’s mean behavioral ratings for 28 distinct stimulus categories (7 emotions $\times$ 2 actors’ sex $\times$ 2 rounds). For the combined valence-and-arousal RDM, the mean valence and arousal ratings for each condition were plotted as coordinates in a two-dimensional space. The observations populating the cells of the resulting 28×28 RDM per participant were the Euclidean distances calculated between these coordinate points for every possible pair of conditions. Euclidean distance was specifically chosen because it captures absolute geometric distances in psychological affective space, accounting for critical shifts in affective intensity and polarity, which purely correlational metrics might overlook (e.g., Russell, 1980). We conducted RSA for valence, arousal, and both valence and arousal.

For statistical inference, we computed Spearman’s rank correlations between the lower-triangular portions of the individual-level RDMs across modalities. These correlation coefficients were Fisher Z-transformed to approach a normal distribution and then subjected to repeated-measures ANOVA, followed by Bonferroni-corrected paired t-tests. The group-average RDM was calculated as the mean of all participants’ RDMs, for visualization purposes only. We applied multidimensional scaling (MDS) to these group-average RDMs and utilized Procrustes analysis, aligning the coordinates of the AO and VO modalities to the AV reference space to enable direct geometric comparisons.

Results

Statistical evaluation

Greenhouse-Geisser correction was employed when the sphericity assumption was violated. All post-hoc tests were Bonferroni-corrected. Regarding the valence ratings, the results revealed a significant main effect of modality, $F(1.8, 119.5)=7.05, p=.002, \eta^2_G=0.008$. The main effect of emotion was also significant, $F(4.6, 304.9)=796.06, p<.001, \eta^2_G=0.839$. The interaction between modality and emotion was significant, $F(8.0, 526.9)=53.90, p<.001, \eta^2_G=0.206$.

Post-hoc analysis revealed that AO conditions were evaluated more negatively than AV, $t(66) = -3.26$, $p = .005$, Cohen's d_z ($d_z = t/\sqrt{N}$) $= -0.398$, and VO conditions, $t(66) = -2.54$, $p = .039$, Cohen's $d_z = -0.310$. However, VO and AV conditions did not differ significantly, $t(66) = 1.24$, $p = .654$, Cohen's $d_z = 0.152$. In the post-hoc analysis for emotion, most emotions significantly differed from each other, $|t(66)| \geq 4.17$, $p \leq .002$, $|Cohen's d_z| \geq 0.509$; however, differences were not significant between disgust and fearful, $t(66) = 0.18$, $p = 1.000$, Cohen's $d_z = 0.022$, disgust and sad, $t(66) = -2.29$, $p = .528$, Cohen's $d_z = -0.280$, and fearful and sad stimuli did not, $t(66) = -2.40$, $p = .400$, Cohen's $d_z = -0.293$.

An additional analysis was conducted to examine the interaction between modality and emotion, specifically how modality affected individual emotion perception, as indicated by valence ratings. Regarding valence (Fig. 2), we analyzed the differences in modality for each emotion to determine if specific emotions were associated with particular modalities. Angry stimuli were rated more negatively in the AO than in the AV condition (see Table 1). Calm stimuli were rated less positively in AO than in AV and VO conditions. Disgust stimuli were rated less negatively in AO than in AV and VO conditions. Happy stimuli were rated less positively in AO than in AV and VO

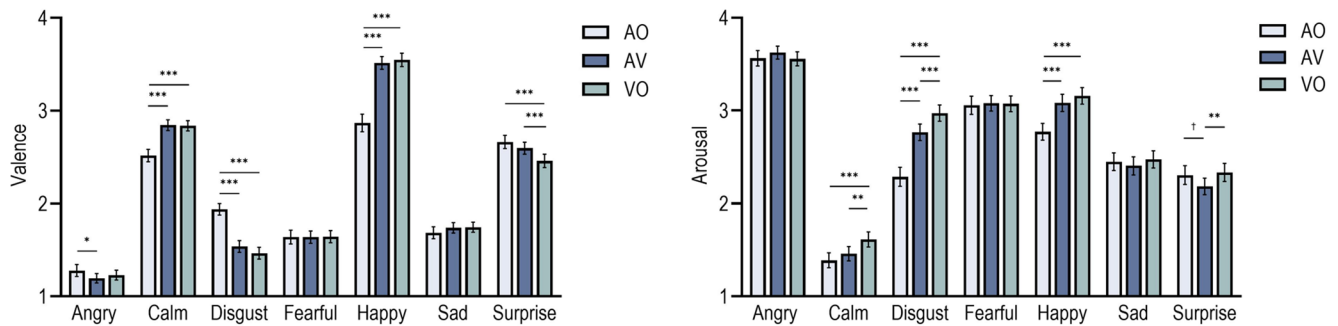


Fig. 2 Average valence and arousal ratings for audio-only, audio-visual, and video-only stimuli, grouped by seven emotions. The left and right panels show valence and arousal ratings, respectively.

AO=audio-only; VO=video-only; AV=audio-visual. The error bars represent within-subject standard errors. * $p < .05$, ** $p < .01$, *** $p < .001$, † $p < .10$

Table 1 Post-hoc comparisons of modality within each emotion for valence and arousal ratings. AO=Audio-only, AV=Audio-visual, VO=Video-only

Emotion	Comparison	Valence			Arousal		
		$t(66)$	p	Cohen's d_z	$t(66)$	p	Cohen's d_z
Angry	AO–AV	2.72	0.025*	0.332	–1.68	0.293	–0.205
	AO–VO	1.41	0.493	0.172	0.13	1.000	0.016
	AV–VO	–1.25	0.653	–0.153	1.84	0.211	0.225
Calm	AO–AV	–6.88	<0.001***	–0.841	–1.60	0.342	–0.196
	AO–VO	–7.91	<0.001***	–0.966	–4.68	<0.001***	–0.572
	AV–VO	0.18	1.000	0.022	–3.78	0.001**	–0.462
Disgust	AO–AV	8.16	<0.001***	0.997	–8.24	<0.001***	–1.007
	AO–VO	10.54	<0.001***	1.288	–9.51	<0.001***	–1.162
	AV–VO	1.87	0.198	0.229	–4.04	<0.001***	–0.494
Fearful	AO–AV	0.004	1.000	0.001	–0.52	1.000	–0.064
	AO–VO	–0.08	1.000	–0.010	–0.40	1.000	–0.049
	AV–VO	–0.10	1.000	–0.012	0.12	1.000	0.015
Happy	AO–AV	–10.71	<0.001***	–1.308	–6.64	<0.001***	–0.811
	AO–VO	–12.75	<0.001***	–1.558	–7.05	<0.001***	–0.861
	AV–VO	–0.86	1.000	–0.105	–1.64	0.319	–0.200
Sad	AO–AV	–1.44	0.466	–0.176	0.87	1.000	0.106
	AO–VO	–1.80	0.228	–0.220	–0.72	1.000	–0.088
	AV–VO	–0.37	1.000	–0.045	–1.40	0.502	–0.171
Surprise	AO–AV	1.45	0.458	0.177	2.23	0.087†	0.272
	AO–VO	4.23	<0.001***	0.517	–0.45	1.000	–0.055
	AV–VO	5.14	<0.001***	0.628	–2.91	0.015*	–0.356

* $p < .05$, ** $p < .01$, *** $p < .001$, † $p < .10$

conditions. Conversely, surprise stimuli were rated more negatively in VO than in AO and AV conditions. These emotions did not differ significantly in other condition comparisons. Additionally, fearful and sad stimuli did not differ significantly across modalities.

Regarding the arousal rating, there was a significant main effect of modality, $F(2.0, 128.8)=30.49$, $p < .001$, $\eta^2_G = 0.033$, and emotion, $F(4.4, 291.2)=316.09$, $p < .001$, $\eta^2_G = 0.668$. Similar to valence, modality and emotion had a significant interaction effect, $F(8.6, 567.8)=19.04$, $p < .001$, $\eta^2_G = 0.057$. In the post-hoc analysis of modality, AO was less arousing than AV and VO conditions, AO–AV: $t(66)=-4.87$, $p < .001$, Cohen's $d_z=-0.595$; AO–VO: $t(66)=-7.38$, $p < .001$, Cohen's $d_z=-0.902$. Furthermore, arousal for VO stimuli was evaluated as stronger than that for AV stimuli, $t(66)=3.23$, $p = .006$, Cohen's $d_z = 0.395$. The post-hoc analysis of emotions revealed no significant differences between fearful and happy stimuli, $t(66)=1.22$, $p=1.000$, Cohen's $d_z = 0.149$, and sad and surprise stimuli, $t(66)=2.19$, $p = .674$, Cohen's $d_z = 0.268$. However, most emotions significantly differed from each other, $|t(66)| \geq 4.89$, $p < .001$, $|Cohen's d_z| \geq 0.597$.

An additional analysis was conducted to examine the interaction effect of modality and emotion on arousal ratings (Fig. 2). Calm stimuli were rated as stronger in VO than in AO and AV conditions (see Table 1). Disgust stimuli differed significantly across all pairs of modalities: weaker in AO than in AV and VO conditions, and stronger in VO than in AV conditions. Happy stimuli were rated as less arousing in AO than in AV and VO conditions. Surprise stimuli were rated as more intense in the VO than in the AV condition, and marginally more intense in the AO than in the AV condition. For these emotions, the other pair of conditions did not exhibit a significant difference. Additionally, angry, fearful, and sad stimuli did not differ significantly across modalities.

Representational similarity analysis and multidimensional scaling

Figure 3 displays group-level RDMs, correlation distance graphs, and multidimensional scaling plots to analyze correlations in behavioral response patterns by modality. For modality-related valence RDMs, the VO RDM was highly similar to the AV RDM. However, the AO RDM was less similar to both AV and VO RDMs (Fig. 3A). To compare the response pattern correlations across modalities, we calculated correlation distances ($1-r$) across modalities at the participant level, transformed correlation distances to Fisher's Z , and then conducted a repeated-measures ANOVA and post-hoc analysis with the Bonferroni correction. A significant difference was observed by modality,

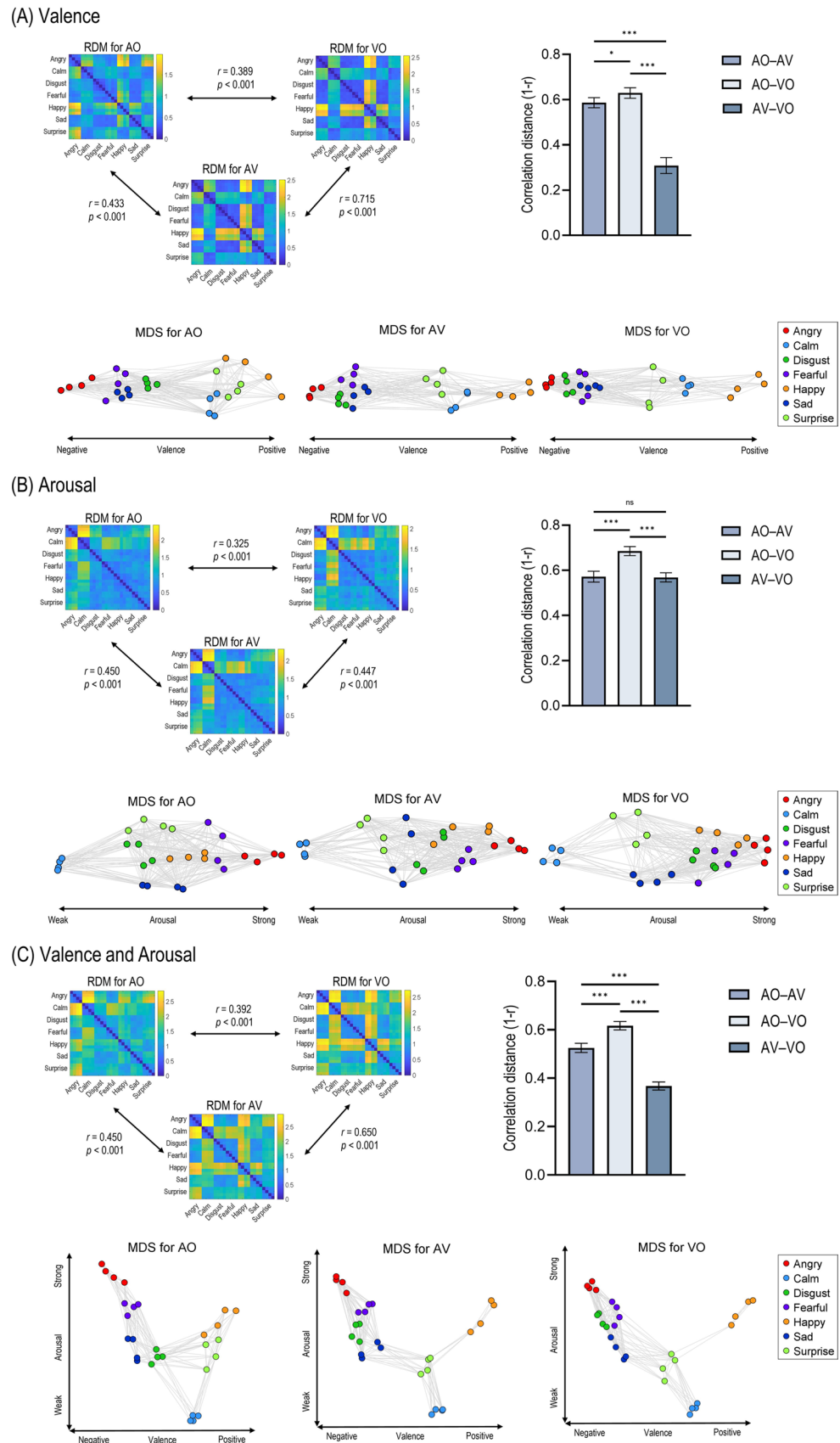
$F(2, 132)=178.04$, $p < .001$. Post-hoc analysis indicated a significantly higher correlation distance of AO–AV than for AV–VO, $t(66)=-15.30$, $p < .001$. The correlation distance of AO–VO was significantly higher than that of AO–AV and AV–VO, (AO–VO) – (AO–AV): $t(66)=-2.97$, $p = .004$; (AO–VO) – (AV–VO): $t(66)=-13.22$, $p < .001$. In the MDS plots, emotions were distributed along the valence dimension in a two-dimensional space. Interestingly, disgust and surprise stimuli were distributed in a more positive direction in the AO plot than in other modality plots.

Regarding the arousal RDMs related to modality, the correlation between AO RDM and AV RDM was similar to that between VO RDM and AV RDM. However, AO RDM was less similar to VO RDM. A repeated-measures ANOVA and post-hoc analysis with Bonferroni correction were conducted as for valence. A significant difference was found by modality, $F(2, 132)=21.16$, $p < .001$. The correlation distance of AO–VO was significantly higher than that of AO–AV and AV–VO, (AO–VO) – (AO–AV): $t(66)=-6.74$, $p < .001$; (AO–VO) – (AV–VO): $t(66)=-5.87$, $p < .001$. However, the correlation distances of AO–AV and AV–VO did not differ significantly, $t(66)=0.13$, $p=1.000$. In the MDS plots, emotions were distributed along the arousal dimension in a two-dimensional space. Disgust and happy stimuli were distributed at lower arousal levels in the AO plot than in other modality plots.

Regarding valence and arousal RDMs, the correlation for VO RDM was highly similar to that for AV RDM. However, AO RDM was less similar to both AV RDM and VO RDM (Fig. 3C)). Repeated-measures ANOVA and post-hoc tests indicated a significant difference in modality, $F(2, 132)=104.26$, $p < .001$. The correlation distance of AO–VO was significantly higher than that of AO–AV and AV–VO, (AO–VO) – (AO–AV): $t(66)=-7.07$, $p < .001$; (AO–VO) – (AV–VO): $t(66)=-13.46$, $p < .001$. The correlation distance of AO–AV was significantly higher than that of AV–VO, $t(66)=-7.87$, $p < .001$. In the MDS plots, emotions were distributed along the valence and arousal dimensions in a two-dimensional space (Fig. 3C). These distributions indicate that the RDMs were successfully modeled with a clear distinction among emotional representations. Notably, disgust and surprise stimuli were positioned closer to the positive emotions in the AO plot than in other modality plots. Additionally, happy stimuli were similarly situated nearer to the other emotions in the AO plot than in other modality plots.

We visualized the combined MDS for the three modalities and connected the corresponding stimuli across modalities to illustrate their representational trajectories (Fig. 4). Interestingly, for disgust, happy, and surprise stimuli, VO was more similar to AV. In contrast, AO was more isolated from other modalities.

Fig. 3 Representational geometry of emotional dimensions of valence and arousal. RDM=representational dissimilarity matrix; MDS=multidimensional scaling; AO=audio-only; VO=video-only; AV=audio-visual. **(A)** RDMs and correlation distance graphs for valence show that the AO matrix was more dissimilar than the AV and VO matrices. MDS plots for valence indicate that emotions are clearly distinguished along the valence axis. **(B)** RDMs and correlation distance graphs for arousal reveal no significant difference in how AO and VO matrices differed from the AV matrix. MDS plots for arousal show that emotions are well-distinguished along the arousal axis. **(C)** Valence and arousal ratings were combined in a two-dimensional space and were modeled using RDMs for valence and arousal. The correlation distance graph shows that the AO matrix was more dissimilar than the AV and VO matrices, such as valence RSA. MDS plots for valence and arousal show that emotions are well-distinguished along the valence-arousal axis. The error bars represent standard error means. * $p < .05$, ** $p < .01$, *** $p < .001$, ns: not significant



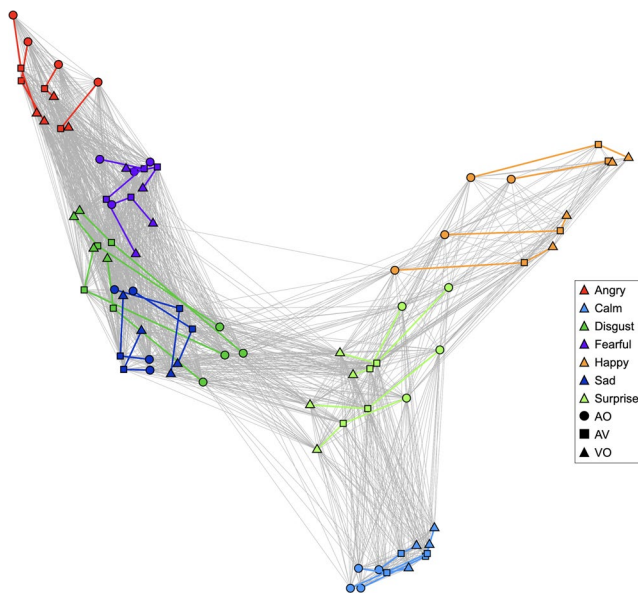


Fig. 4 Representational trajectories across modalities. Multidimensional scaling (MDS) of audio- only (AO), video-only (VO), and audio-visual (AV) was integrated into a single embedding space. Corresponding stimuli across modalities were connected to visualize their representational trajectories. Each modality is shown with a different marker: AO, circle; AV, square; and VO, triangle. This visualization displays condition-specific trajectories, emphasizing how representations shift across modalities. Notably, for disgust, happy, and surprise stimuli, AO was more isolated from other modalities

Discussion

The current study investigated how sensory modality influences emotion perception and whether certain emotions rely on specific modalities. We utilized MDS and RSA using dimensions of three emotional modalities—AO, VO, and AV. The findings demonstrated emotion-specific modality effects. As we hypothesized, the visual and auditory modalities distinctly influenced emotion perception with respect to valence and arousal levels. Moreover, while there was an overall tendency for individuals to perceive emotions as more negative and weaker for AO stimuli than for those with visual information, this effect was primarily driven by specific emotions such as happiness and disgust. Conversely, evaluations of angry, fearful, and sad emotions remained robustly consistent across modalities, whereas assessments of surprise exhibited a unique pattern reflecting a greater influence of auditory cues.

The present study showed that individuals perceive emotional valence differently in the auditory modality than in other modalities. The relationship between valence and arousal was V-shaped in the VO and AV conditions (Fig. S1), which aligns with previous findings of a consistent shape across emotional stimuli (Gjoreski et al., 2022; Kron et al., 2015; Kuppens et al., 2013). However, the shape in

the AO condition was different than that in other modalities. Specifically, individuals perceived happy stimuli less positively in the AO condition. Additionally, the response patterns in the valence and arousal RSA were less similar in AO than in other modalities (Fig. 3(C)). Separate examination of valence and arousal showed that the disparity in AO response patterns was more pronounced for valence (Fig. 3(A)). These results indicate a greater influence of visual information, especially when assessing the valence of emotions. It suggests that individuals perceive auditory emotional information without visual cues, such as that in phone calls, differently than in other modalities.

In contrast, arousal perception did not differ significantly between auditory and visual modalities. Specifically, the response pattern similarity between VO and AV conditions and that of AO and AV conditions did not differ in the arousal RSA (Fig. 3(B)), suggesting that arousal perception did not favor either modality. Moreover, arousal in the VO condition was perceived as stronger than in the AV condition. Previous research has found that multimodal information is more effective than unimodal information in emotion recognition (Zheng & Hsiao, 2023). Multimodal information is accompanied by high BOLD responses in brain imaging (Kreifelts et al., 2007) and elicits stronger arousal responses due to its increased sensory input. However, the present findings suggest that multimodality may not lead to greater enhancement in every case. Arousal has been explored less frequently than valence (Zsido, 2024); therefore, more research is needed to investigate this outcome.

We observed that modality effects in emotion perception varied across emotional categories. Specifically, visual information played an important role in perceiving disgust, happiness, and calm. This aligns with a previous study, which reported that individuals rely more on visual information when recognizing disgust and happiness (Zheng & Hsiao, 2023). Moreover, modality-related differences in the valence patterns of calm were similar to those of happiness, indicating that calm may have been evaluated as a relatively positive emotion in the present study. Future studies would benefit from including neutral emotions as a baseline condition when comparing calm with other emotions. By contrast, other emotions were perceived consistently across modalities. Specifically, although the valence of anger was rated less negatively in the AO than in the AV condition, overall ratings were consistent across modalities. Moreover, fearful and sad had consistent valence and arousal ratings across modalities, as indicated by the ANOVA results. These findings are in line with prior studies on the characteristics of gaze movements, and the audiovisual associations with these emotions may have been reflected, even with limited sensory modalities. However, because the current study did not directly measure eye movements or manipulate

audiovisual associations, interpretations of these mechanisms remain largely speculative. These mechanisms should be tested in future studies using eye-tracking methods or by manipulating variables related to audiovisual associations.

Interestingly, surprise exhibited a distinctive pattern across sensory modalities. In our results, surprise stimuli were rated more negatively in the VO condition than in both the AO and AV conditions, whereas valence ratings did not differ significantly between AO and AV. These results indicate that the absence of auditory information leads to a more negative affective judgment of surprise, highlighting the critical role of auditory cues in modulating its valence. The uniqueness of surprise perception across modalities aligns with previous studies, which have reported modality-specific differences in arousal (Li et al., 2025) or demonstrated auditory dominance in recognition accuracy (Chong et al., 2024). A common feature in previous studies and the current results is that, unlike other emotions, the evaluation of surprise is significantly influenced by auditory information. Consequently, the results show that auditory information strongly contributes to evaluating surprise stimuli, suggesting that employing fully integrated AV stimuli may facilitate more comprehensive affective judgments in research involving surprise.

The current study has several limitations. First, one limitation regarding the observed modality effects is potential asymmetry in the direction of emotion and self-relevance across sensory stimuli. Emotion perception and recognition are influenced by the direction of visual information, such as gaze and head orientation (Adams & Kleck, 2005; Hess et al., 2007; Milders et al., 2011). Furthermore, the self-relevance of emotional stimuli influences the intensity of emotion, involving the amygdala, fusiform gyrus, and medial prefrontal cortices (N'Diaye et al., 2009). The current study used visual forward-facing stimuli to simulate social interaction in participants. Therefore, the visual information may have provided emotional direction or self-relevance, enhancing emotion perception. In contrast, auditory stimuli in the current study might have lacked direction or emotional self-relevance. The auditory stimulus consisted of a neutral sentence containing random information presented by an actor. Due to a lack of self-relevant information in the auditory stimuli, the participants might have deemphasized auditory information. Therefore, the observed modality effects may partly reflect the greater directionality and social relevance of the visual stimuli rather than modality differences alone. Future studies should use auditory information that includes properties such as dialogue with the participant's name (Deng et al., 2022) or a calling situation that induces one-to-one interaction.

Second, neutral stimuli and weaker emotional stimuli were not included, which could limit the interpretation of valence-space organization, particularly for calm and surprise. Although calm exhibited a valence pattern similar to

that of happy, and surprise demonstrated a pattern relatively consistent with auditory dominance across modality conditions, these interpretations should be considered tentative because they have not been compared with responses to neutral stimuli. Future studies should include neutral emotional stimuli as a baseline condition. Additionally, modality effects may differ for lower-intensity or more ambiguous emotional expressions. Thus, future studies should compare normal- and strong-intensity stimuli to examine whether modality effects vary with emotional intensity or ambiguity.

Finally, another limitation of the study is the potential influence of cross-cultural and cross-linguistic factors. All participants were young Korean adults, whereas the RAVDESS stimuli were produced by North American English-speaking actors. Participants were instructed to focus on the emotional expression rather than the semantic meaning of speech, and the auditory stimuli consisted of semantically neutral sentences. However, we cannot exclude the possibility that familiarity with English prosody, accent, or culturally specific emotional expressions influenced emotion perception, particularly in the AO and AV conditions. Future research should include measures of language proficiency and cultural exposure and compare culturally matched and mismatched stimuli to determine whether modality-specific influences generalize across linguistic and cultural groups.

Further studies using other channels in emotion perception are needed (e.g., eye movement patterns, neural signal patterns) to elaborate on current behavioral results. Specifically, happy and disgust are characterized by predominant gaze movements to the nose and mouth regions than are other emotions (e.g., angry, fearful) which primarily involve a gaze toward the eyes (Calvo et al., 2014, 2018; Nusseck et al., 2008; Schurgin et al., 2014; Smith et al., 2005; Zheng & Hsiao, 2023). This indicates that the perception of disgust and happy emotions from visual stimuli is characterized by intricate visual information, including that conveyed by the nose and mouth, which potentially reflects a stronger contribution of visual information to emotion perception. Furthermore, a neural study found that multimodal stimuli were associated with higher accuracy in emotion recognition (Kreifelts et al., 2007). As such, it is necessary to determine whether the modality effects on emotion perception identified in this study are also supported by neural evidence. Moreover, studies of modality effects should consider inducing competition between auditory information and visual information in the experimental design, such as by using incongruent stimuli.

The current study used a dimensional approach and confirmed that the modality effects of emotion perception vary across emotional categories, suggesting that the relative influence of visual versus auditory cues depends on the nature of the emotion. Our findings illustrate the distinct patterns of

audiovisual emotion evaluation and have implications for various domains. For example, in clinical psychology, modality-dependent emotion perception is known to be impaired or systematically altered in specific populations, such as older adults (Lin et al., 2024, 2025) and individuals with schizophrenia (Vogel et al., 2016). By elucidating typical patterns of modality reliance, our results can help tailor diagnostic tools for these groups. The results are of utility to neuroscience, as they can inform the exploration of dynamic neural integration processes, providing a behavioral basis for how the brain optimally weights sensory inputs based on task demands (Klasen et al., 2011). In human-computer interaction, our results can be used to enhance emotion-aware technologies by guiding the development of multimodal systems that mimic human-like shifts in sensor reliance (Poria et al., 2017). Finally, our study can help explain interpersonal communication dynamics in social psychology by extending existing frameworks of social cue integration to specific affective judgments (Zaki, 2013).

Abbreviations

AO	Audio-Only
AV	Audio-Visual
VO	Video-Only
RSA	Representational Similarity Analysis
RDMs	Representational Dissimilarity Matrices
MDS	Multidimensional Scaling

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00426-026-02340-3>.

Acknowledgements Not applicable.

Author contributions **Yongseong Lee** : Investigation, Methodology, Software, Data Curation, Formal analysis, Visualization, Writing - Original Draft, Writing - Review & Editing. **Youngju Lee** : Investigation, Data Curation, Formal analysis, Writing - Original Draft, Writing - Review & Editing. **Dongha Lee** : Conceptualization, Funding acquisition, Supervision, Writing - Review & Editing.

Funding This research was supported by KBRI basic research program through Korea Brain Research Institute funded by Ministry of Science and ICT (26-BR-03-01, 26-BR-03-02, 26-BR-02-11) and the Bio&Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00401794).

Data availability All stimuli and preprocessed data for the current study are available on the Open Science Framework at [<https://osf.io/759a4>]. Data were analyzed using MATLAB 2024b and R 4.4.2. Our study hypothesis, design, and analysis were not preregistered.

Declarations

Ethics approval and consent to participate This study was approved by the Institutional Review Board of the Korea Brain Research Institute, and all participants were informed about the experimental process and received a consent form (KBRI-202111-HR-003-08).

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

- Adams, R. B. Jr., & Kleck, R. E. (2005). Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion*, 5(1), 3–11. <https://doi.org/10.1037/1528-3542.5.1.3>
- Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1), 20–46. https://doi.org/10.1207/s15327957pspr1001_2
- Calvo, M. G., Fernández-Martín, A., & Nummenmaa, L. (2014). Facial expression recognition in peripheral versus central vision: Role of the eyes and the mouth. *Psychological Research Psychologische Forschung*, 78(2), 180–195.
- Calvo, M. G., Fernández-Martín, A., Gutierrez-Garcia, A., & Lundqvist, D. (2018). Selective eye fixations on diagnostic face regions of dynamic emotional expressions: KDEF-dyn database. *Scientific Reports*, 8(1), 17039. <https://doi.org/10.1038/s41598-018-35259-w>
- Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, 11(12), 535–543. <https://doi.org/10.1016/j.tics.2007.10.001>
- Chong, C. S., Davis, C., & Kim, J. (2024). A cantonese audio-visual emotional speech (CAVES) dataset. *Behav Research Methods*, 56(5), 5264–5278. <https://doi.org/10.3758/s13428-023-02270-7>
- Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research*, 1242, 126–135. <https://doi.org/10.1016/j.brainres.2008.04.023>
- De Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion*, 14(3), 289–311. <https://doi.org/10.1080/026999300378824>
- Dellacherie, D., Bigand, E., Molin, P., Baulac, M., & Samson, S. (2011). Multidimensional scaling of emotional responses to music in patients with temporal lobe resection. *Cortex; A Journal Devoted To The Study Of The Nervous System And Behavior*, 47(9), 1107–1115. <https://doi.org/10.1016/j.cortex.2011.05.007>
- Deng, N., Sun, Y., Chen, X., & Li, W. (2022). How does self name influence the neural processing of emotional prosody? An ERP study. *PsyCh Journal*, 11(1), 30–42. <https://doi.org/10.1002/pchj.499>
- Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3–4). <https://doi.org/10.1080/02699939208411068>. 169–200.
- Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870), 429–433. <https://doi.org/10.1038/415429a>
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Gjoreski, M., Kiprijanovska, I., Stankoski, S., Mavridou, I., Broulidakis, M. J., Gjoreski, H., & Nduka, C. (2022). Facial EMG sensing for monitoring affect using a wearable device. *Scientific Reports*, 12(1), 16876. <https://doi.org/10.1038/s41598-022-21456-1>
- Gu, S., Wang, F., Cao, C., Wu, E., Tang, Y. Y., & Huang, J. H. (2019). An integrative way for studying neural basis of basic emotions with fMRI. *Frontiers in Neuroscience*, 13, 628. <https://doi.org/10.3389/fnins.2019.00628>
- Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in Cognitive Sciences*, 16(9), 458–466. <https://doi.org/10.1016/j.tics.2012.07.006>

- Hess, U., Adams, R. B., & Kleck, R. E. (2007). Looking at you or looking elsewhere: The influence of head orientation on the signal value of emotional facial expressions. *Motivation and Emotion*, 31(2), 137–144. <https://doi.org/10.1007/s11031-007-9057-x>
- Klasen, M., Kenworthy, C. A., Mathiak, K. A., Kircher, T. T., & Mathiak, K. (2011). Supramodal representation of emotions. *Journal of Neuroscience*, 31(38), 13635–13643. <https://doi.org/10.1523/JNEUROSCI.2833-11.2011>
- Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. *Neuroimage*, 37(4), 1445–1456. <https://doi.org/10.1016/j.neuroimage.2007.06.020>
- Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4. <https://doi.org/10.3389/neuro.06.004.2008>
- Kron, A., Pilkiw, M., Banaei, J., Goldstein, A., & Anderson, A. K. (2015). Are valence and arousal separable in emotional experience? *Emotion*, 15(1), 35–44. <https://doi.org/10.1037/a0038474>
- Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, 139(4), 917–940. <https://doi.org/10.1037/a0030811>
- Lee, Y., Seo, Y., Lee, Y., & Lee, D. (2023). Dimensional emotions are represented by distinct topographical brain networks. *International Journal of Clinical and Health Psychology*, 23(4), 100408. <https://doi.org/10.1016/j.ijchp.2023.100408>
- Lee, Y., Lee, Y., Kim, Y., & Lee, D. (2026). Sensory modality affects emotional dimension perception and frontoparietal brain responses. *International Journal of Clinical and Health Psychology*, 26, 100699. <https://doi.org/10.1016/j.ijchp.2026.100699>
- Li, Y., Li, A., Tao, J., Li, F., Erickson, D., & Akagi, M. (2025). Contributions of audio and visual modalities to perception of Mandarin Chinese emotions in valence-arousal space. *Acoustical Science and Technology*, 46(1), 55–63. <https://doi.org/10.1250/ast.e24.41>
- Lin, Y., Ye, X., Zhang, H., Xu, F., Zhang, J., Ding, H., & Zhang, Y. (2024). Category-sensitive age-related shifts between prosodic and semantic dominance in emotion perception linked to cognitive capacities. *Journal of Speech Language and Hearing Research*, 67(12), 4829–4849. <https://doi.org/10.1044/2024.JSLHR-23-0081>
- Lin, Y., Xu, F., Ye, X., Zhang, H., Ding, H., & Zhang, Y. (2025). Age and sex differences in emotion perception are influenced by emotional category and communication channel. *Psychology and Aging*, 40(1), 17. <https://doi.org/10.1037/pag0000828>
- 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. <https://doi.org/10.1371/journal.pone.0196391>
- Milders, M., Hietanen, J. K., Leppanen, J. M., & Braun, M. (2011). Detection of emotional faces is modulated by the direction of eye gaze. *Emotion*, 11(6), 1456–1461. <https://doi.org/10.1037/a0022901>
- N'Diaye, K., Sander, D., & Vuilleumier, P. (2009). Self-relevance processing in the human amygdala: gaze direction, facial expression, and emotion intensity. *Emotion*, 9(6), 798–806. <https://doi.org/10.1037/a0017845>
- Nussek, M., Cunningham, D. W., Wallraven, C., & Bulthoff, H. H. (2008). The contribution of different facial regions to the recognition of conversational expressions. *Journal of Vision*, 8(8), 11–23. <https://doi.org/10.1167/8.8.1>
- Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37, 98–125. <https://doi.org/10.1016/j.inffus.2017.02.003>
- Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3), 715–734. <https://doi.org/10.1017/S0954579405050340>
- Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, 140(1), 109. <https://doi.org/10.1037/a0031903>
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161. <https://doi.org/10.1037/h0077714>
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172. <https://doi.org/10.1037/0033-295x.110.1.145>
- Russell, J. A., & Bullock, M. (1985). Multidimensional scaling of emotional facial expressions: similarity from preschoolers to adults. *Journal of Personality and Social Psychology*, 48(5), 1290. <https://doi.org/10.1037/0022-3514.48.5.1290>
- Schurgin, M. W., Nelson, J., Iida, S., Ohira, H., Chiao, J. Y., & Francorneri, S. L. (2014). Eye movements during emotion recognition in faces. *Journal of Vision*, 14(13), 14. <https://doi.org/10.1167/14.13.14>
- 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, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological Science*, 16(3), 184–189. <https://doi.org/10.1111/j.0956-7976.2005.00801.x>
- Vogel, B. D., Brück, C., Jacob, H., Eberle, M., & Wildgruber, D. (2016). Effects of cue modality and emotional category on recognition of nonverbal emotional signals in schizophrenia. *BMC psychiatry*, 16(1), 218. <https://doi.org/10.1186/s12888-016-0913-7>
- 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>
- Zaki, J. (2013). Cue integration: A common framework for social cognition and physical perception. *Perspectives on Psychological Science*, 8(3), 296–312. <https://doi.org/10.1177/1745691613475454>
- Zheng, Y., & Hsiao, J. H. (2023). Differential audiovisual information processing in emotion recognition: An eye-tracking study. *Emotion*, 23(4), 1028–1039. <https://doi.org/10.1037/emo0001144>
- Zsido, A. N. (2024). The effect of emotional arousal on visual attentional performance: a systematic review. *Psychological Research Psychologische Forschung*, 88(1), 1–24. <https://doi.org/10.1007/s00426-023-01852-6>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.