Discover Artificial Intelligence

RESEARCH Open Access

Facial expression-based assessment of emotional engagement under multimodal stimuli

Andrea Generosi[1], Milena Martarelli[2], Paolo Castellini[2] and Maura Mengoni[2*]

*Correspondence:
Maura Mengoni
m.mengoni@univpm.it
[1]Department of Information Science and Technology, Università Pegaso, Napoli, Italy
[2]Department of Industrial Engineering and Mathematical Sciences, Università Politecnica delle Marche, Ancona, Italy

Abstract
This study investigates the feasibility of an emotion-derived engagement proxy computed from facial expression recognition (FER) outputs in response to auditory and audiovisual stimuli, in an immersive and acoustically isolated setting relevant to experience evaluation. Thirty-seven participants were exposed to three stimulus blocks: Sound, Sound with Video (same audio content and duration), and a validated strong-elicitation benchmark (FilmStim excerpts). The Sound vs. Sound with Video contrast was used as an expected sensitivity check, while FilmStim served to contextualise response magnitude. Within the fixed-sequence protocol, the proposed FER-derived engagement proxy discriminated the three blocks in the expected direction. We discuss limitations related to fixed order, demographic composition, and FER uncertainty, and outline how the approach can complement self-report and sensor-based measures in early-stage evaluation workflows.

Keywords Facial expression recognition, Emotion recognition, Deep learning, Computer vision

1 Introduction

Over the past decade, Visual and auditory stimuli can be defined as objects or events eliciting a response by producing a visual or aural sensory experience that lead to a change in perceptions of the environment [1]. This environment is the center-stage not only in business communities but also in cultural activities to enhance user experience, boost product attractiveness and heightened excitement.

The auditory experience of a product is much more than just a "sensory response to an acoustical stimulus." Users often associate auditory characteristics, such as reliability and quality, with the product itself. The sound a product makes has a substantial impact on users' perceptions, emotional states, and their overall relationship with the product, ultimately affecting their purchase decisions, preferences, and expectations regarding its performance. In this sense, the auditory dimension is not merely about hearing a sound (e.g., identifying it as loud or sharp); it is fundamental to establishing an emotional and psychological bond between the user and the product. The auditory aspect has become

an integral part of product design, requiring focused attention due to its substantial influence on the success of a product [2]. Modern consumers increasingly demand products with refined auditory features, and the sonic attributes of a product have become vital elements influencing brand perception and purchasing choices [3, 4]. Understanding the contribution of auditory cues to the complete user experience is therefore essential, particularly in products where sound plays an important role.

Visual and audio combined visual stimuli are widely available in our daily lives. From cinemas to museums, clothing, entertainment, in almost all commercial products visual appeal is indispensable. It is to strengthen or increase emotional responses with the subject [5]. reviewed the effects of visual or auditory stimuli on visual attention tasks. The manipulated emotional arousal then measured through behavioral, ocular, or neural responses. The review found that task relevant arousing stimuli captured and held attention, regardless of how they were presented [1]. examined how visual and auditory cues shape emotional responses to inform the creation of multisensory museum experiences. Using both self-report questionnaires and psychophysiological recordings, it found that, according to self-reports, sound and images are equally effective at eliciting specific emotions. However, the physiological data showed that, for most emotions, sound produced higher arousal levels than visual stimuli. The development of new products involves a considerable amount of testing aimed at understanding how sounds are perceived. This research investigates the emotional impact of both auditory and visual stimuli. Drawing on the findings of [2] and [6], who identified differences in affective reactions to auditory versus visual attributes, we aimed to determine whether auditory and visual stimuli could influence customers' emotional responses. Specifically, we studied the impact of three types of stimuli, i.e. "Sound" (auditory-only), "Sound with video" (combined auditory-visual stimuli), and "Elicitation Video" (multimedia video stimuli, i.e. visual and audio, designed to elicit certain emotions), on emotional perception. The core manipulation used to address our research objective is the comparison between Sound and Sound with Video, where the auditory track is identical and duration-matched and the only systematic difference is the addition of congruent visual information. The Elicitation Video condition (FilmStim excerpts) is included as a validated benchmark of strong elicitation to contextualise the magnitude of responses, rather than as a modality-matched control. By employing facial expression recognition (FER), we operationalise an emotion-derived engagement proxy computed from FER emotion probabilities (Eq. 1) and test its feasibility and sensitivity in an immersive and acoustically isolated setting. Specifically, the Sound vs. Sound with Video contrast is used as an expected sensitivity check (same audio content and duration), while FilmStim clips serve as a validated strong-elicitation benchmark; the manuscript does not claim a sequence-independent causal modality effect. In this manuscript, we use the term engagement to denote a specific construct: high-arousal, threat-related (orienting/defensive) emotional activation, operationalised by Eq. (1). Accordingly, the proposed engagement score should not be interpreted as a measure of immersion, presence, or positive motivational engagement.

For years, the scientific community has explored methods to objectively and quantitatively assess human emotional engagement, particularly through the analysis of external signals such as facial expressions, vocal characteristics, and physiological markers. Among the pioneers in this field is Paul Ekman, who in 1971 proposed the existence of a distinct set of universal emotions (i.e. anger, disgust, fear, happiness, sadness, and

surprise) that are expressed through facial expressions shared across different cultures [7]. These foundational theories have significantly shaped subsequent studies on emotion recognition involving various forms of stimuli, including auditory and visual content. Ekman's contributions, as well as those of others, have laid the foundation for the development of automated systems that can recognize and categorize emotions, with applications ranging from product design to human-computer interaction.

In this work, we focus on a specific facet of engagement, namely high-arousal, threat-related (orienting/alerting) emotional engagement, rather than engagement in a broad motivational sense. Accordingly, we operationalise engagement through an emotion-derived index that emphasises surprise, anger and fear as indicators of attentional capture and defensive/orienting responses. This focus motivates the use of aversive versus relaxing stimuli as controlled exemplars of design-relevant emotional dimensions (e.g., annoyance/tension versus calmness/comfort), and it directly motivates our experimental contrasts.

1.1 Correlation between sound and emotions

Numerous studies have demonstrated a substantial correlation between auditory stimuli and emotional responses. For example, Lee et al. examined the relationship between subjective ratings, sound metrics, and brain activity using Electroencephalogram (EEG) data [8]. Medvedev et al. studied the influence of soundscapes on physiological factors such as skin conductivity and heart rate, concluding that acoustic environments can support physiological recovery after experiencing stress [9]. Likewise, Siddharth et al. showed that audiovisual media could trigger physiological changes, using EEG and facial expression analysis to monitor emotional reactions [10]. These studies emphasize the complexity of the interaction between auditory stimuli and emotional states and the importance of using multiple physiological and psychological measures when assessing emotional impact.

Building on this body of work, the present study explores auditory and visual stimuli in three categories: auditory-only, combined auditory-visual, and primarily visual stimuli enriched with audio. In our protocol, the "combined auditory-visual" category corresponds to the Sound with Video condition, where the same 30-second everyday sounds used in the Sound condition are paired with simple congruent videos of their source events, whereas the "primarily visual stimuli enriched with audio" category corresponds to the Elicitation Video condition, which consists of 10-second FilmStim movie excerpts presented with their original soundtrack. This approach allows us to investigate how different combinations of sensory inputs impact emotional responses and whether visual components can enhance or alter the emotional effects triggered by auditory stimuli. To achieve these objectives, we employed facial expression recognition to compare emotional reactions across the different types of stimuli. Through this, we aim to provide valuable insights into how different sensory inputs can be used to enhance the user experience, particularly in contexts where multisensory interactions are vital for user satisfaction.

In line with contemporary models of engagement, prior work has shown that emotional engagement is tightly linked to learners' affective reactions (e.g., interest, enjoyment, boredom, frustration) and that these emotions covary with behavioural and cognitive aspects of engagement [11, 12]. Likewise, in affective computing, several

authors have proposed emotion-based engagement models in which engagement levels are explicitly derived from discrete emotions detected by affect recognition systems [13, 14].

1.2 Technologies for engagement evaluation

In immersive and multimodal studies, multiple constructs such as presence/immersion and engagement are commonly assessed via self-report questionnaires and psychophysiological measures via self-report questionnaires/ratings (e.g., presence/immersion scales [15]) and psychophysiological measures such as heart rate/HRV and electrodermal activity, and in some cases EEG-derived indices [1, 6, 8, 9, 10]. While these approaches are well established, they typically require dedicated sensors, calibration, and additional instrumentation that can be intrusive or impractical for rapid, in-situ evaluations. In contrast, camera-based facial-expression recognition (FER) provides a low-intrusive and potentially real-time alternative that can be deployed with minimal setup (e.g., a single camera, offline processing). Here, however, our FER-derived metric targets orienting/threat-related emotional engagement (attentional capture under high arousal), and it is not intended to quantify presence/immersion per se. Since increased arousal/engagement under audiovisual stimulation is already a well-established finding [6, 10], the value of the present work lies in demonstrating, within an immersive and acoustically isolated protocol, that a FER-derived, emotion-based index can discriminate expected condition differences and behaves coherently against a validated strong-elicitation benchmark (FilmStim), supporting its use as a practical decision-support signal to complement (not replace) questionnaire- and sensor-based measures in early-stage evaluation workflows. Noninvasive technologies like facial expression analysis have increasingly been used to evaluate emotional responses. These techniques aim to identify facial movements linked to emotions, with models such as the Facial Action Coding System (FACS) commonly employed to recognize universal emotions like happiness, sadness, anger, and fear [16]. Presently, most facial expression recognition systems leverage deep neural networks, specifically Convolutional Neural Networks (CNNs), to predict emotional states from facial imagery [17, 18]. Compared to other biofeedback tools, such as EEG or galvanic skin response (GSR), facial expression recognition is far less invasive, making it better suited for real-world applications. Advances in deep learning have significantly enhanced the accuracy and reliability of these systems, enabling more detailed emotion recognition by capturing subtle variations in facial expressions.

Attention, another important factor in emotional engagement, can be assessed using gaze tracking techniques. These methods employ computer vision algorithms to analyze gaze direction and identify where the user is focusing their attention [19, 20]. In this study, facial expression analysis was used to provide a comprehensive understanding of customers' engagement with the presented stimuli. In addition to facial expression recognition and gaze tracking [21, 22], other noninvasive methods, such as body posture analysis and vocal analysis, can also be utilized to assess emotional valence and engagement [23, 24]. While these methods were not applied in the present study, they represent promising directions for future research. Integrating multiple noninvasive approaches provides a more comprehensive understanding of emotional responses, which is particularly valuable for applications in product testing, user experience evaluation, and marketing research.

1.3 Research aim

The main objective of this study is to operationalise an emotion-derived engagement proxy from facial expression recognition (FER) outputs and to evaluate its feasibility and sensitivity for assessing user engagement in response to auditory and audiovisual stimuli. By comparing different types of stimuli and using advanced emotion recognition technologies, we aim to identify key factors that drive emotional and cognitive engagement. In particular, given the adopted engagement operationalisation, we target a high-arousal, threat-related facet of engagement (orienting/alerting engagement), and we therefore use aversive versus relaxing stimuli as controlled contrasts for testing the sensitivity of the proposed index. To this end, we conducted a case study where we compared emotional responses to auditory stimuli alone, audiovisual stimuli, and validated visual stimuli enhanced with audio, all within an immersive and acoustically isolated environment. The design addresses this objective by isolating a modality manipulation through the Sound vs. Sound with Video contrast (same audio content and duration), while using validated FilmStim clips as an additional benchmark condition to contextualise response magnitude. Accordingly, the main empirical goal of the case study is a sensitivity/feasibility check of the proposed FER-derived engagement proxy (i.e., whether it discriminates between conditions expected to differ), rather than the demonstration of a novel or sequence-independent multimodal effect.

The knowledge gained from this research can be applied to designing products and experiences that align more closely with users' emotional needs, thereby improving satisfaction and brand loyalty. Moreover, our findings have broader implications for fields such as advertising, virtual reality, and human-computer interaction, where understanding and optimizing user engagement is crucial.

Rather than claiming novelty in the multimodal effect itself, the contribution of this work is to demonstrate, in a design-relevant case study and immersive/acoustically isolated setup, the feasibility and sensitivity of a low-intrusive, camera-based (FER-derived) engagement proxy that can complement questionnaire and sensor-based approaches and be embedded into early-stage evaluation workflows for auditory and audiovisual experiences.

2 Materials and methods

Research into human emotional responses to video and audio stimuli has yielded significant insights. The International Affective Digitised Sounds (IADS) [25] and International Affective Pictures System (IAPS) [26] are widely used datasets as stimuli for experiments about eliciting emotions [27]. compared different emotional stimulation methods, auditory, visual and combined audio-visual utilising IADS and IAPS both individually and in combination. It concluded that auditory stimuli alone could evoke stronger emotional and physiological responses than visual or combined stimuli as measured by physiological signals.

Emotional images and sounds together result in stronger and faster emotional reactions compared to presenting either alone [28], this can be observed comparing video clips and static images eliciting emotional responses, with videos eliciting stronger emotional responses, both subjectively and neurologically [29].

In this study, we employ a software tool based on deep learning and computer vision to extract indicators of engagement during auditory and visual stimulation. For visual

stimulation, video clips are used rather than static images due to their effectiveness. This tool has been utilized to assess participants' reactions during auditory-only and combined auditory-visual stimuli.

2.1 Facial expression recognition tool used for engagement evaluation

For this research, a tool based on a Convolutional Neural Network (CNN) implemented in Python using the Keras and TensorFlow frameworks was developed [30] and validated in a real-world scenario, i.e. in a clothing store to analyse customer experience [31]. The classifier outputs per-frame probabilities over seven classes (neutral, happiness, surprise, sadness, anger, disgust and fear) from aligned 64×64 grayscale face crops. Facial expression recognition models are typically trained on two types of datasets: those created under controlled laboratory conditions with high-quality data and those captured "in the wild" with greater diversity but lower accuracy [32]. In this study, we combined the strengths of both dataset types to create a comprehensive merged dataset. Public datasets CK+ [33], FER+ [34], and AffectNet [35] were used, preprocessed to maintain consistency across images, and filtered to exclude unsuitable data (e.g., images with multiple faces or low resolution).

In the present study, the FER component is therefore not introduced as a novel contribution, but as a previously validated module integrated into our real-time immersive setup: it was selected because it is already fully compatible with the acquisition hardware and control software of the experimental room and because it runs completely offline and on-premise, in line with our data-protection requirements. We report below the complete architecture and training protocol to ensure transparency and reproducibility.

Training was performed from scratch using Keras (v2.2.4) and TensorFlow (v1.13.1). The model was trained on the merged dataset using predefined partitions (train/validation/test file lists and corresponding folders). Model selection was performed via ModelCheckpoint monitoring validation accuracy (save_best_only=True), and the final performance reported in this study (75.48%) refers to the held-out test partition.

We additionally report the confusion matrix and per-class precision, recall, and F1 scores on the held-out test partition (Appendix A.1, Table 5; Fig. 6). Because the engagement proxy in Eq. (1) is directly driven by FER outputs, these class-wise metrics provide an estimate of the signal quality (and noise) underlying the engagement computation. As commonly observed in in-the-wild FER benchmarks, performance is emotion-dependent, especially among visually similar negative expressions. Accordingly, in this study the engagement proxy is interpreted as a time-averaged, relative signal for within-subject condition comparisons rather than as a frame-level ground truth.

To mitigate class imbalance, the training set was balanced via oversampling by replication: for each class, samples were cycled until reaching the maximum class count (i.e., the size of the most represented class). Data augmentation was applied online using an ImageDataGenerator with featurewise centering and standardisation, width/height shifts (0.08), zoom (0.05), rotation (20°), shear (0.05), and horizontal flipping. The loss function was categorical cross-entropy.

Subject-level independence could not be verified uniformly across all merged sources because identity information is not consistently available in the public datasets and the training code relies on predefined partition lists; we therefore report results under

this constraint and explicitly acknowledge it as a limitation of the training/evaluation protocol.

The deployment process involved several key steps:

- Cropping of detected face regions
- Facial alignment and centralization
- Rotation correction
- Resizing the cropped image to 64 × 64 pixels
- Grayscale conversion
- Processing by the CNN model

The deployed CNN follows a VGG13-style configuration with four convolutional blocks (3 × 3 kernels, same padding, ReLU activations) and no batch normalisation. The blocks use filter progressions 64–64, 128–128, 256-256-256, and 256-256-256; each block is followed by 2 × 2 max pooling and dropout (0.25). The classifier head is Flatten > Dense(1024, ReLU) > Dropout(0.5) > Dense(1024, ReLU) > Dropout(0.5) > Dense(7, Softmax). Training used SGD (lr = 0.025, momentum = 0.9, nesterov=True, decay = 0.0005) with a linear epoch schedule lr(epoch) = 0.025·(1 – epoch/100), batch size 128, for 100 epochs, selecting the best checkpoint by validation accuracy (no early stopping).

The network's final layer implements a softmax, producing per-frame probabilities over the Ekman emotions that sum to 100%. Following Altuwairqi et al.'s affective model [14], we mapped the emotion percenteges to 3 levels and assigned equal impact weights within each level. In particular:

- Strong: surprise
- High: anger + fear
- Medium: happiness + sadness + disgust

Low and Disengagement in Altuwairqi's scheme rely on non-Ekman states (e.g., boredom/sleepiness/behaviors) and were not scored here. After that, we obtained a continuous 0-100 engagement score by assigning ordinal weights (Strong = 1, High = 0.5, Medium = 0) and computing:

$$Engagement_{0-100} = \ Surprise\,(\%) \ + \ 0.5\,[Anger\,(\%) + Fear\,(\%)] \tag{1}$$

This scalar preserves Altuwairqi's ordinal emphasis (Strong > High > Medium) while providing a frame-level continuous measure. Because the index is computed directly from softmax outputs, $\text{Engagement}_{0-100}$ should be regarded as a classifier-dependent estimate (pseudo-probability based), whose absolute values may be affected by probability miscalibration and class-specific errors; therefore, in this study we primarily rely on time-averaged, within-subject comparisons rather than frame-level interpretation.

In this formulation, the engagement score specifically targets a high-arousal, threat-related facet of emotional engagement. Surprise, anger and fear are weighted because, in Altuwairqi et al.'s model, they correspond to Strong and High engagement levels and are typically associated with orienting and defensive responses, whereas happiness, sadness and disgust are grouped in a Medium level and therefore do not contribute directly to the scalar index. As a consequence, stimuli that elicit predominantly positive emotions

(e.g., pure happiness with little or no surprise, anger or fear) can yield low/zero Engagement$_{0-100}$ values. For this reason, the index should be interpreted as a context-specific, emotion-derived engagement proxy, rather than as a comprehensive measure of all possible forms of emotional engagement.

Engagement$_{0\text{-}100}$ is first computed at the frame level from FER probabilities and then time-averaged across all frames of each clip to obtain a participant-level engagement value per stimulus. No temporal smoothing (e.g., moving-average filtering) was applied to the per-frame softmax probabilities prior to computing Eq. (1): the index was computed per frame and then averaged across frames of each clip. The weighting coefficients (1.0 for Surprise, 0.5 for Anger/Fear) follow Altuwairqi's ordinal mapping and should be interpreted as a model-based design choice, exploring formal coefficient-sensitivity and probability calibration is left as future work. This index relies on the assumptions that (1) FER probabilities provide a noisy but informative proxy of expressed affect, (2) Altuwairqi's ordinal mapping (Strong > High > Medium) can be operationalised as a continuous score for comparative analyses, and (3) time-averaging mitigates frame-level misclassifications. Surprise, anger, and fear are theoretically linked to attentional orienting and defensive mobilisation in threat-appraisal models of attention, which motivates their role as the highest-weight components of the index for capturing high-arousal, orienting engagement. The index is intended to capture a high-arousal, orienting/threat-related facet of engagement (rather than enjoyment-based engagement), consistent with Altuwairqi's ordinal mapping. In this manuscript, the index is evaluated as a proof-of-concept sensitivity measure. Sensitivity is operationally defined as the index's ability to rank the three a priori ordered conditions (Sound < Sound with Video < FilmStim benchmark) in the expected direction with statistical reliability, while additional analyses verify that the observed ordering is not trivially explained by monotonic time-on-task drift (Sect. 3).

2.2 Immersive audio-visual paradigm for affective assessment

The experimental protocol was designed to evaluate emotional responses effectively through facial expression recognition within an immersive environment. To objectively measure emotions such as anger, fear, happiness, and overall engagement, participants underwent a structured sequence involving distinct auditory-only stimuli followed by combined audiovisual presentations. Importantly, the three stimulus blocks were presented in a fixed order for all participants (Sound Sound with Video - Elicitation Video) to maintain a consistent familiarisation and pairing between the audio-only clips and their corresponding audiovisual counterparts. We presented the Sound block first to obtain an audio-only baseline not contaminated by visual-context carryover, since visual context can prime and influence the identification of environmental sounds (i.e., contextual priming effects) [2]. Each auditory stimulus was deliberately interspersed with calming sounds (e.g., a serene river stream) to ensure emotional neutrality before the subsequent stimulus. This sequencing strategy is consistent with the gold-standard methodologies established in prior research, which demonstrate that alternating emotionally charged stimuli with neutral segments enhances the sensitivity and accuracy of emotional detection [27]. The auditory-only conditions aimed to isolate the emotional impact of sound, acknowledging findings from literature indicating typically lower fear responses in auditory-only scenarios compared to audiovisual conditions [28].

Subsequently, the Sound with Video block was introduced as an expected sensitivity check for the proposed FER-derived engagement proxy (same audio content and duration, with congruent visual information added). Given the fixed block order, this contrast is not interpreted as sequence-independent causal evidence, but as a proof-of-concept demonstration of the index's ability to detect a directionally predictable difference within the adopted protocol.

The immersive setup, featuring dim lighting, controlled acoustics, and advanced visual projection technology, was purposefully developed to minimize external distractions, thereby fostering an environment conducive to authentic emotional reactions.

To minimise potential confounds related to exposure time, the three auditory stimuli used in the Sound and Sound with Video conditions were identical and had a fixed duration of 30 s in both modalities: thus, the main comparison of interest (audio-only vs. audiovisual) is duration-matched at the stimulus level. By contrast, the 10-second clips in the Elicitation Video condition constitute a different stimulus category: validated film excerpts from the FilmStim database [36], included as an additional benchmark for our engagement metric rather than as a direct duration-matched control, and used to provide a strong emotional reference for calibration rather than to isolate modality effects. For all conditions, engagement scores were computed as time-averaged values over the full length of each clip, so that differences in total duration mainly affect the number of analysed frames but not the scale of the engagement index itself.

2.3 Experimental setup

The case study was conducted at the Polytechnic University of Marche, involving a total of 37 participants. These participants were both students and university employees, representing a diverse range of age groups. The composition of the participant pool included 16 females and 21 males, resulting in an almost equal male-to-female ratio and they came from different countries. The average age of participants was 25 years, with a standard deviation of 5.3, indicating a relatively young group but with some variability in ages. Participants were recruited from different departments and had varying levels of familiarity with the content of the study, which provided a balanced and representative sample for our purposes. This approach allowed for a broader understanding of how different demographics might respond to the experimental conditions.

The experimental setup featured a webcam mounted on a tripod, positioned in front of each participant. This specific placement was chosen to ensure that the facial recordings were unobtrusive, minimizing any potential impact on the participants' natural experience of the auditory and visual stimuli. The environment was designed to capture participants' attention and focus it on the content displayed. The tests took place in a dimly lit room, with ambient noise minimized. Participants were provided with headphones and placed in an enclosed environment to prevent external distractions. For visual content, the case study utilized a ProjectionDesign F10 AS3D ZOOM projector, to enhance visual immersion. Upon arrival, participants were given time to acclimate to the testing environment, and they received clear instructions about the testing procedure and their roles. This configuration aimed to capture authentic emotional reactions while maintaining the participants' comfort and minimizing distractions during the experiment.

The study utilized several tools for data collection and analysis. Video streams were recorded using the Windows Camera app, with a Logitech HD C925E webcam

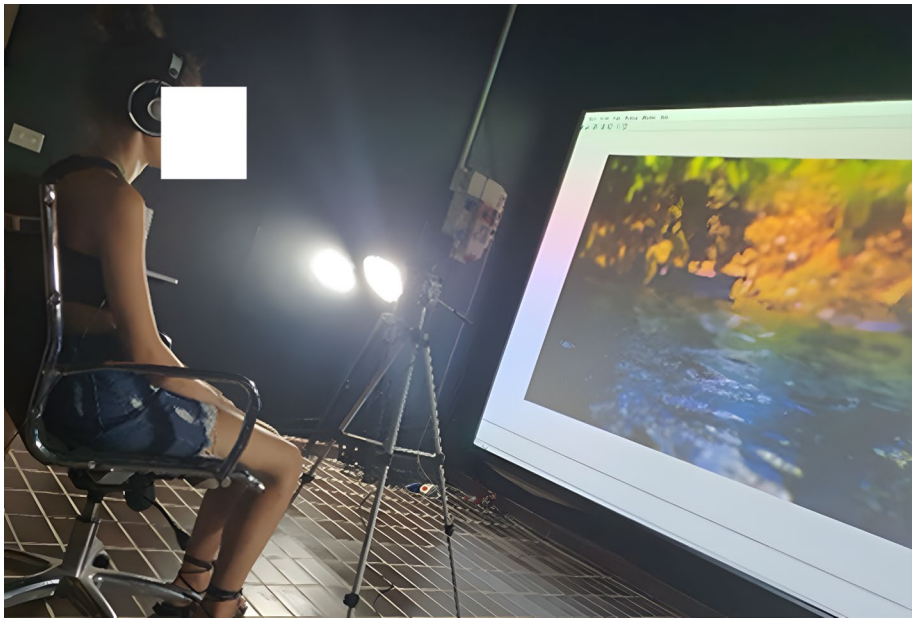

Fig. 1 Test environment setup

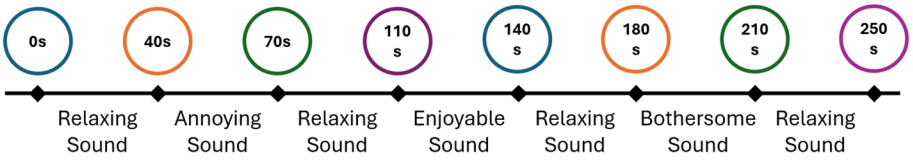

Fig. 2 Sound sequence [0-250] seconds

connected to a PC for capturing participants' facial expressions. To manage the presentation of audio and video stimuli, a custom software application developed in Matlab was used, which also handled the synchronization of stimulus timing with the recording software. For video analysis, a Python-based software, described in paragraph 2.1, was employed to extract relevant data. This software analyzed each frame to determine the probability scores associated with Ekman's six universal emotions, along with timestamps. The collected data were then stored in CSV format for further analysis. Image processing was carried out using a high-performance desktop PC featuring an Intel i9 10,900 CPU, Nvidia Quadro P2200 GPU, and 128GB of RAM, ensuring efficient processing of the recorded data and accurate analysis of facial expressions.

All participants read and signed an Informed Consent Form, ensuring they were fully informed of the study's purpose, the procedures involved, and any potential risks and benefits before agreeing to participate. Figure 2 illustrates the experimental setup used in the case study.

The primary aim of the case study was to investigate whether variations in auditory and visual stimuli could affect participants' emotional responses, particularly focusing on engagement levels as an indicator of participants' reactions to the presented stimuli. To achieve this, participants were exposed to different auditory and visual stimuli, each intended to evoke a range of emotions corresponding to Ekman's expressions as defined by the FACS method. The stimuli were divided into three main categories: "Sound" (Fig. 2), "Sound with video" (Fig. 3), and "Elicitation video" (Fig. 4). Elicitation videos

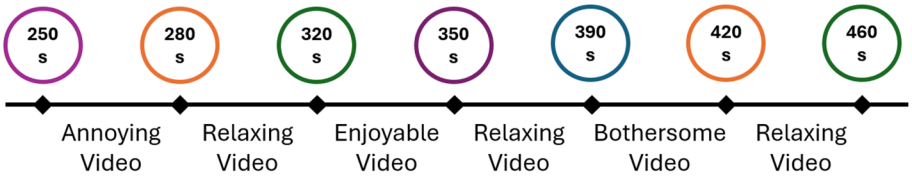

Fig. 3 Sound with Video sequence [210–460] seconds

Fig. 4 Elicitation Video sequence from 420 s to the end

are selected from the FilmStim database validated by Schaefer's studio and collaborators [36], they can elicit specific emotional states.

Figures 2, 3 and 4 present the sequence used in the case study. Initially, in Fig. 2, participants were presented with three distinct 30-second sound clips interspersed with a 40-second relaxing audio segment of a calm spring river to help return the participant to a neutral emotional state. In the subsequent sequence (Fig. 3), the same previous auditory stimuli were repeated with corresponding video clips, which included visuals of a blackboard, clapping, and a mosquito, paired with their respective sounds. Each segment was separated by the same relaxing video. Finally, as shown in Fig. 4, participants viewed six 10-second videos selected from the aforementioned validated dataset, designed to elicit specific emotional responses, with the same 40-second soothing video interspersed between them.

Thus, in this study the combined auditory-visual condition ("Sound with Video") always uses the same 30-second auditory clips as the Sound condition, simply enriched with congruent videos depicting the same source event. By contrast, the "validated visual content with audio" condition ("Elicitation Video") consists of independent 10-second FilmStim movie excerpts, each with its own original audio track and validated emotional profile.

In addition, all stimuli are summarised in Table 1, which reports for each item intended target emotion, duration, and source. In the Sound and Sound with Video conditions, three 30-s auditory clips were used (e.g., a hand scratching a Blackboard, an applauding audience, a mosquito) selected from the Web to elicit anger/disgust/fear and happiness, and more "relaxing" sounds (e.g., a calm river stream) aimed at inducing neutral or low-arousal states. In the Sound with Video condition, these sounds were paired with congruent videos (e.g., a mosquito flying near the ear), with identical durations to the corresponding audio-only clips. The 40-s river segment served as a neutral baseline between stimuli. The Elicitation Video condition consisted of six 10-s clips from the FilmStim database [36], each targeting a specific discrete emotion (happiness, sadness, anger, fear, disgust, or neutral). The FilmStim identifiers and targeted emotions are also reported in Table 1.

Although these stimuli are generic affective elicitors rather than sounds taken from a specific product, they were deliberately selected because they instantiate acoustic qualities, such as harshness, roughness, high-frequency annoyance versus smoothness

Table 1 Summary of stimuli used in the three experimental conditions

Label	Condition	Intended target emotion(s)	Duration (s)	Source
Bothersome sound/video	Sound / Sound with Video	Anger, disgust, fear	30	Blackboard scratching [37]
Enjoyable sound/video	Sound / Sound with Video	Happiness	30	Applause [38]
Annoying sound/video	Sound / Sound with Video	Anger, disgust	30	Mosquito buzzing (combined [39] and [40])
Relaxing sound/video	Neutral filler	Neutral / low arousal	40	Calm spring river [41]
Dramatic video	Elicitation Video	Sadness, Anger	10	FilmStim clip (Schindler's list (1))
Anger video	Elicitation Video	Anger	10	FilmStim clip (Schindler's List (3))
Scary video	Elicitation Video	Fear	10	FilmStim clip (Scream 2)
Funny video	Elicitation Video	Happiness	10	FilmStim clip (There is something about Mary (1))

Table 2 Median engagement scores with 95% CI and IQR for each stimulus category

Macro-category	Median value	95% CI of median	IQR (Q1 – Q3)
Sound	6.9	4.9–8.9	5.0–11.5
Sound + video	9.5	7.3–11.7	6.0–13.2
Elicitation video	11.0	9.2–12.8	8.0–14.0

and calmness, that are directly relevant to product- and sound-design decisions (e.g., warning tones, notification sounds, ambient soundscapes). In this proof-of-concept study, they thus serve as controlled exemplars of design-relevant emotional dimensions (annoyance vs. relaxation, tension vs. comfort) rather than as representations of a single concrete product.

3 Discussion of results

We compared the three stimulus macro-categories using a repeated-measures approach. Normality of the engagement-score distributions was assessed with Shapiro-Wilk tests (Sound: $W = 0.94$, $p = .021$; Sound + Video: $W = 0.91$, $p = .006$; Elicitation Video: $W = 0.88$, $p = .001$), all indicating significant departures from normality. Consequently, a non-parametric Friedman test was applied and revealed a significant overall effect, $\chi^2(2) = 28.10$, $p < .001$ (Kendall's $W = 0.45$). The magnitude of the effect we observed lies in the medium-to-large range and is comparable to the engagement modulation reported in multimodal emotion-induction studies that paired IAPS images with IADS sounds [27], thereby situating our findings within the broader affective-science literature.

Bonferroni-adjusted Wilcoxon signed-rank tests ($\alpha\_adj = 0.017$) showed higher engagement for Sound + Video versus Sound ($Z = -3.10$, $p = .002$, $r = .46$) and for Elicitation Video versus Sound ($Z = -4.19$, $p < .001$, $r = .63$). Engagement was also greater for Elicitation Video versus Sound + Video ($Z = -2.51$, $p = .012$, $r = .37$). Median (IQR) engagement scores were 6.9 (5.0–11.5) for Sound, 9.5 (6.0–13.2) for Sound + Video, and 11.0 (8.0–14.0) for Elicitation Video (Table 2). Within the fixed-sequence protocol adopted in this study, engagement scores were higher in the Sound + Video block than in the Sound block, which is consistent with the expected multimodal modulation; however, because block order was not counterbalanced, this contrast is reported primarily

as a sensitivity check for the proposed engagement proxy within this protocol, rather than as a sequence-independent causal demonstration of a modality effect. We also note that the time-on-task analysis can only rule out simple monotonic drift and cannot fully separate condition effects from sequence effects. Throughout this section, 'engagement' refers specifically to the Eq. (1) orienting/threat-related engagement proxy, not to broader motivational or immersive engagement.

The even stronger responses observed for the Elicitation Video condition reflect the fact that these FilmStim clips are purpose-built and validated to elicit intense emotions and are used here as a benchmark to contextualise the magnitude of the engagement modulation, rather than as a modality-matched comparator for the Sound + Video clips. Although the direction of the audiovisual effect is well established, the contribution of this study is methodological-applied: within an immersive, acoustically isolated protocol, we show that a low-intrusive, camera-based FER pipeline can yield an emotion-derived engagement proxy that differentiates the expected condition contrast (Sound vs. Sound + Video, with identical audio content and duration) and behaves coherently when contextualised against a validated strong-elicitation benchmark (FilmStim). This supports the use of the proposed index as a practical screening and decision-support signal in early-stage evaluation, to be complemented (rather than replaced) by self-report and/or physiological measures when stronger construct validation or theory-driven engagement modelling is required.

As a robustness check for time-on-task effects, we modelled engagement as a function of block position and within-block progression with participant as a random effect, and we additionally tested for monotonic drift in the neutral filler segments. The mixed-effects model on stimuli showed only a borderline negative time trend ($\beta = -3.253$, $p = .054$), whereas the filler-only model did not show any significant drift over the session ($\beta = 0.296$, $p = .239$). This pattern suggests that the observed condition differences are unlikely to be solely explained by a monotonic order-related increase in engagement.

As an exploratory, descriptive complement to the inferential analyses reported above, another qualitative analysis has been conducted.

Figure 5 presents the time-courses of the means and associated variability (± 1 SD) for anger, disgust, fear, and engagement elicited by scary versus relaxing sounds. Anger remained uniformly low for relaxing stimuli, whereas scary stimuli yielded consistently higher mean anger; even the lower standard-deviation bound for anger in the scary category exceeded the mean for relaxing sounds, underscoring their reliably anger-provoking effect. Disgust responses were generally modest, with only slightly higher means for scary than relaxing elicitations, but the markedly larger standard deviation for scary stimuli indicates that aversive sounds occasionally triggered stronger, more heterogeneous disgust reactions. Average fear remained low in both categories and only slightly elevated for scary elicitations; pronounced spikes in the standard deviation for the scary condition reveal transient episodes of intense fear despite the low overall mean, whereas relaxing stimuli produced minimal fear with negligible variance. Finally, engagement remained comparatively high across both categories, yet it was more variable and reached higher peaks for scary stimuli: in some segments, the upper bound for scary sounds exceeded 28 (arbitrary units), indicating brief episodes of very high attentional capture, while relaxing sounds clustered within a narrower, lower range. This figure is

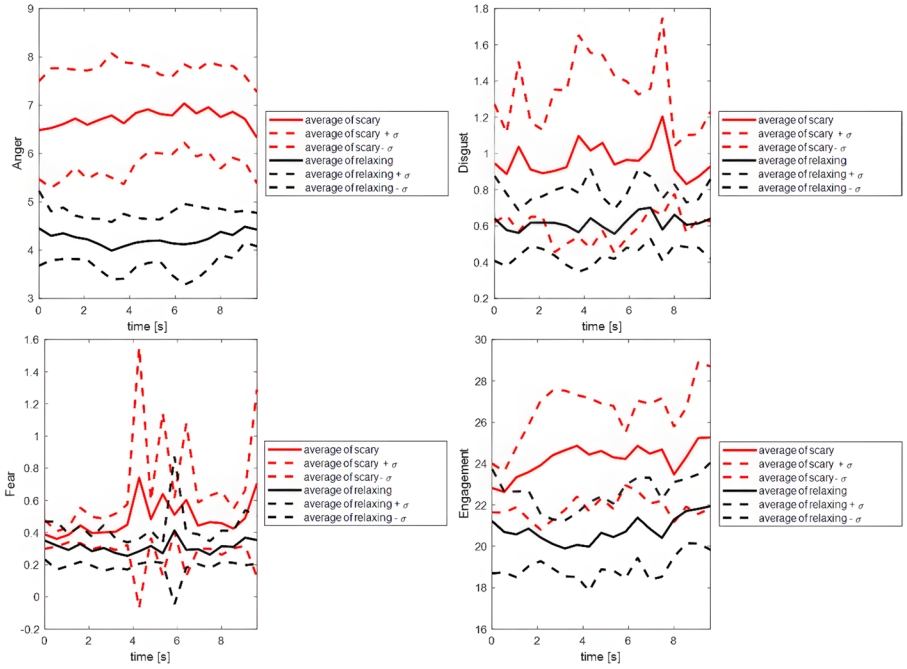

Fig. 5 Time sequence of scary/relaxing sounds mean and standard deviation

intended as an exploratory, descriptive illustration of temporal dynamics and between-participant variability rather than a basis for additional inferential claims.

From a product and sound-design perspective, these results illustrate how the proposed FER-based, emotion-derived engagement index can function as a decision-support tool in early evaluation phases. By exposing users to alternative auditory and audiovisual concepts and aggregating engagement scores across participants, designers can obtain an objective, low-intrusive indication of which variants elicit stronger or weaker emotional involvement. In practice, such quantitative evidence can be combined with self-report and qualitative feedback to guide the selection and refinement of product sounds, notification tones, or branding videos, aligning the final design more closely with the intended emotional experience.

4 Conclusions

This study highlights the potential of facial expression recognition tools as an automated method for evaluating a high-arousal, orienting/threat-related engagement proxy derived from facial expressions under auditory and audiovisual stimuli. The research had two primary objectives. The first objective was to examine how these tools perform in detecting emotional states elicited by auditory stimuli, both in isolation and in combination with visual stimuli, using validated content specifically designed to evoke particular emotions. As expected and in line with prior work on multimodal communication, the FER-derived engagement proxy yielded higher values in the audiovisual block than in the audio-only block; in this manuscript, this pattern is interpreted as proof-of-concept sensitivity evidence for the proposed engagement index within the adopted fixed-sequence protocol rather than as a sequence-independent causal effect. This effect, although not conceptually novel, was objectively quantified through our FER-based pipeline and

confirmed by statistical analysis, which indicated higher values of the FER-derived orienting/threat-related engagement proxy in the audiovisual block.

The second objective of the study was to explore how facial expression recognition tools could aid in understanding participants' levels of engagement. The findings revealed certain limitations of the facial expression recognition software, but at the same time they showed a correspondence between the engagement index derived from facial expressions and the observed emotional responses. In this sense, the contrast between audio-only and audiovisual conditions served as a controlled, "expected" test case to evaluate the sensitivity and feasibility of the proposed engagement metric.

Several constraints temper the generalization of the present findings. First, engagement was inferred exclusively from facial-action-unit analysis, an approach sensitive to camera angle, illumination, and algorithmic bias. In this study, no block-level self-report ratings or physiological/behavioural measures were collected: therefore, convergent validation of the proposed proxy could not be performed on the present dataset. Incorporating converging measures such as galvanic skin response, heart-rate variability, or self-report scales would provide a multi-method validation of the construct and help rule out modality-specific noise. Moreover, because the engagement index is computed from the probabilities produced by a FER model whose overall accuracy on in-the-wild data is moderate (75.48%) and emotion-dependent, it should be regarded as a noisy proxy rather than a ground-truth measure of engagement; accordingly, in this case study we focus on relative, within-subject differences between conditions rather than on absolute values for individual frames. In addition, the absolute scale of the proxy depends on the chosen ordinal weighting scheme and on the calibration of the underlying FER probabilities; hence, results are interpreted at the level of relative, aggregated contrasts. Second, the sequence of stimulus blocks was not counterbalanced: all participants experienced first the Sound, then the Sound with Video, and finally the Elicitation Video condition. While this choice ensured consistent familiarisation with the auditory material before its audiovisual counterparts, it also implies that differences between blocks may partly reflect order-related factors (e.g., learning, anticipation, fatigue, habituation). However, an additional time-on-task analysis on the current dataset did not indicate a systematic monotonic drift in engagement during neutral fillers ($p = .239$), and the overall time effect on stimuli was weak and borderline ($\beta < 0$, $p = .054$).

Finally, although the audio-only and audio-visual stimuli were carefully duration-matched, the validated elicitation clips were shorter (10 s vs. 30 s). While the use of a time-averaged engagement index partly mitigates this issue, future studies should employ strictly duration-matched stimuli across all conditions to fully exclude any residual influence of exposure time on engagement scores.

Moreover, to exploit the full richness of the temporal emotion and engagement trajectories, future work should adopt more sophisticated quantitative approaches (e.g., mixed-effects or time-series models) capable of capturing within- and between-participant variability over time.

Despite the aforementioned limitations, this study has demonstrated the potential of the proposed approach, especially in scenarios where a sufficiently large sample size is available or where specific auditory stimuli are expected to evoke distinct emotional states. The ability to evaluate participants' emotional responses in an automated and objective manner, without introducing bias due to distractions, represents a promising

step toward more scalable evaluation methods. Moreover, this study underscores the importance of integrating emotional considerations into product design and opens new opportunities for incorporating sound quality into the design process, ensuring a more holistic approach that takes into account the emotional responses of customers.

Overall, the study should be regarded as a proof-of-concept demonstrating the practical use of a FER-based, emotion-derived engagement index in multimodal, design-relevant evaluation scenarios.

Appendix

Emotion recognition network architecture

Table 3 VGG13 FER architecture (layer-by-layer)

Block/layer	Configuration
Input	64 × 64 × 1 (grayscale aligned face crop)
Conv Block 1	Conv(64, 3 × 3, padding=same, ReLU) → Conv(64, 3 × 3, padding=same, ReLU) → MaxPool(2 × 2) → Dropout(0.25)
Conv Block 2	Conv(128, 3 × 3, padding=same, ReLU) → Conv(128, 3 × 3, padding=same, ReLU) → MaxPool(2 × 2) → Dropout(0.25)
Conv Block 3	Conv(256, 3 × 3, padding=same, ReLU) → Conv(256, 3 × 3, padding=same, ReLU) → Conv(256, 3 × 3, padding=same, ReLU) → MaxPool(2 × 2) → Dropout(0.25)
Conv Block 4	Conv(256, 3 × 3, padding=same, ReLU) → Conv(256, 3 × 3, padding=same, ReLU) → Conv(256, 3 × 3, padding=same, ReLU) → MaxPool(2 × 2) → Dropout(0.25)
Classifier head	Flatten → Dense(1024, ReLU) → Dropout(0.50) → Dense(1024, ReLU) → Dropout(0.50) → Dense(7, Softmax)
Batch Normalization	Not used
Output classes (index→label)	0 = neutral; 1 = happy; 2 = surprise; 3 = sad; 4 = angry; 5 = disgust; 6 = fear

Table 4 Training protocol and hyperparameters

Item	Value
Framework versions	TensorFlow 1.13.1, Keras 2.2.4
Training initialisation	From scratch (no pretrained weights loaded during training)
Loss function	categorical_crossentropy
Optimizer	SGD(lr = 0.025, decay = 0.0005, momentum = 0.9, nesterov=True)
Learning-rate schedule	Linear per epoch: lr(epoch) = 0.025*(1 - epoch/100)
Batch size	128
Epochs	100
Early stopping	Not used
Model selection	ModelCheckpoint(monitor='val_acc', save_best_only=True) (best validation-accuracy checkpoint retained)
Data augmentation	featurewise_center=True; featurewise_std_normalization=True; width_shift_range = 0.08; height_shift_range = 0.08;zoom_range = 0.05; rotation_range = 20;shear_range = 0.05; horizontal_flip=True
Subject-level independence	Not uniformly verifiable across merged datasets due to inconsistent identity metadata
Brightness/contrast jitter	Not used
Class imbalance handling	Oversampling by replication: for each class, samples are cycled until reaching maxnum (= size of the most frequent class)
Train/val/test split method	Predefined file lists and separated folders
Inference decision rule	Predicted class = argmax(softmax output vector)

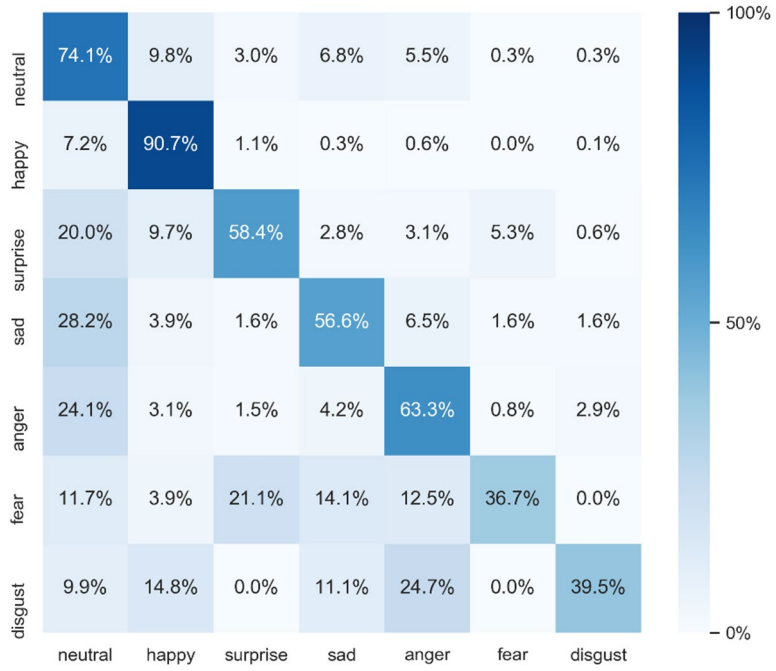

Fig. 6 Confusion matrix

Table 5 Per-class performance of the FER classifier on the held-out test split used to compute the reported overall accuracy (75.48%)

Classe	Precision	Recall	F1-score
neutral	69.84%	74.10%	71.91%
happy	88.79%	90.70%	89.73%
surprise	62.99%	58.40%	60.61%
sad	62.81%	56.60%	59.54%
anger	63.45%	63.30%	63.37%
fear	59.39%	36.70%	45.36%
disgust	49.86%	39.50%	44.08%

Acknowledgements
We would like to thank Dr. Reza Jamali and Dr. Josè Yuri Villafan for their valuable contribution in conducting the experimental tests and data analysis.

Author contributions
System design, software development, data collection, writing review, experimental design, writing original paper, data interpretation A.G.; data analysis, writing review, data interpretation M.Ma.; project coordination, writing review and editing, M.Me., P.C. All authors have read and agreed to the published version of the manuscript.

Funding
No funding was received for conducting this study.

Data availability
The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the Università Politecnica delle Marche (Ancona, Italy), protocol code 0021586 and date of approval February 7, 2023. Informed consent was obtained from all subjects involved in the study.

Consent for publication
Participants consented to the publication of anonymized data and images.

Competing interests
The authors declare no competing interests.

Received: 30 August 2025 / Accepted: 27 March 2026

Published online: 11 April 2026

References

1. Jelinčić DA, Šveb M, Stewart E, A. Designing sensory museum experiences for visitors' emotional responses. Museum Manage curatorship. 2022;37(5):513–30.
2. Özcan E, Cupchik GC, Schifferstein HJN. Auditory and visual contributions to affective product quality. Int J Des. 2017;11(1):35–50.
3. Melzner J, Raghubir P. The sound of music: The effect of timbral sound quality in audio logos on brand personality perception. J Mark Res. 2023;60(5):932–49.
4. Lowe ML, Haws KL. Sounds big: The effects of acoustic pitch on product perceptions. J Mark Res. 2017;54(2):331–46.
5. Zsidó AN. The effect of emotional arousal on visual attentional performance: a systematic review. Psychol Res. 2024;88(1):1–24.
6. Park SH, Lee PJ, Jung T, Swenson A. Effects of the aural and visual experience on psycho-physiological recovery in urban and rural environments. Appl Acoust. 2020;169:107486.
7. Ekman P. (1971). Universals and cultural differences in facial expressions of emotion. In Nebraska symposium on motivation. University of Nebraska Press.
8. Lee YJ, Shin TJ, Lee SK. Sound quality analysis of a passenger car based on electroencephalography. J Mech Sci Technol. 2013;27(2):319–25.
9. Medvedev O, Shepherd D, Hautus MJ. The restorative potential of soundscapes: A physiological investigation. Appl Acoust. 2015;96:20–6.
10. Siddharth S, Jung TP, Sejnowski TJ. Impact of affective multimedia content on the electroencephalogram and facial expressions. Sci Rep. 2019;9(1):1–10.
11. Fredricks JA, Blumenfeld PC, Paris AH. (2004). School engagement: Potential of the concept, state of the evidence. Review of Educational Research, 74(1), 59–109.n based real-time learner engagement detection system in online learning context using deep learning models.
12. Pentaraki A, Burkholder GJ. Emerging evidence regarding the roles of emotional, behavioural, and cognitive aspects of student engagement in the online classroom. Eur J Open Distance E-Learning. 2017;20(1):1–21.
13. Altuwairqi K, Jarraya SK, Allinjawi A, Hammami M. A new emotion–based affective model to detect student's engagement. J King Saud University-Computer Inform Sci. 2021;33(1):99–109.
14. Alyüz N, Okur E, Oktay E, Genc U, Aslan S, Mete SE, Stanhill D, Arnrich B, Esme AA. (2016). Towards an emotional engagement model: Can affective states of a learner be automatically detected in a 1:1 learning scenario? In Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2016) – Extended Proceedings.
15. Witmer BG, Singer MJ. Measuring presence in virtual environments: A presence questionnaire. Presence: Teleoperators Virtual Environ. 1998;7(3):225–40.
16. Ekman P, Friesen WV. Facial action coding system. Environmental Psychology & Nonverbal Behavior; 1978.
17. Li S, Deng W. Deep facial expression recognition: A survey. IEEE transactions on affective computing; 2020.
18. Ge H, Zhu Z, Dai Y, Wang B, Wu X. Facial expression recognition based on deep learning. Comput Methods Programs Biomed. 2022;215:106621.
19. Kim N, Lee H. Assessing consumer attention and arousal using eye-tracking technology in virtual retail environment. Front Psychol. 2021;12:665658.
20. Gheorghe CM, Purcărea VL, Gheorghe IR. Using eye-tracking technology in Neuromarketing. Romanian J Ophthalmol. 2023;67(1):2.
21. Hu, G., Xin, Y., Lyu, W., Huang, H., Sun, C., Zhu, Z., … Seifi, H. (2024). Recent trends of multimodal affective computing: A survey from NLP perspective. arXiv preprint arXiv:2409.07388.
22. Ma Y, Zhang X, Zhang Y, Fjeld M. (2025, April). Measuring User Experience Through Speech Analysis: Insights from HCI Interviews. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1–9).
23. Gerdemann SC, Vaish A, Hepach R. Body posture as a measure of emotional valence in young children: a preregistered validation study. Front Dev Psychol. 2025;3:1536440.
24. Themistocleous C. Linguistic and emotional prosody: A systematic review and ALE meta-analysis. Neuroscience & Biobehavioral Reviews; 2025. p. 106210.
25. Bradley MM, Lang PJ. International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings (Tech. Rep. No. B-2). Gainesville, FL: The Center for Research in Psychophysiology, University of Florida; 1999.
26. Lang PJ, Bradley MM, Cuthbert N. International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual, 2005.
27. Polo EM, et al. Comparative assessment of physiological responses to emotional elicitation by auditory and visual stimuli. IEEE J Translational Eng Health Med. 2023;12:171–81.
28. Gerdes, Antje BM, Matthias J, Wieser, Georg W. Alpers. Emotional pictures and sounds: a review of multimodal interactions of emotion cues in multiple domains. Front Psychol. 2014;5:1351.
29. Romeo Z, Fusina F, Semenzato L, Bonato M, Angrilli A, Spironelli C. Comparison of Slides and Video Clips as Different Methods for Inducing Emotions: An Electroencephalographic Alpha Modulation Study. Front Hum Neurosci. 2022;16:901422. https://doi.org/10.3389/fnhum.2022.901422. PMID: 35734350; PMCID: PMC9207173.
30. Talipu A, Generosi M, Mengoni L, Giraldi. Evaluation of Deep Convolutional Neural Network architectures for Emotion Recognition in the Wild, in 2019 IEEE 23rd International Symposium on Consumer Technologies, 25–27, 2019. IEEE https://doi.org/10.1109/ISCE.2019.8900994
31. Generosi A, Ceccacci S, Mengoni M. (2018, September). A deep learning-based system to track and analyze customer behavior in retail store. In 2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin) (pp. 1–6). IEEE.
32. Gaya-Morey FX, Manresa-Yee C, Martinie C, Buades-Rubio JM. (2025). Evaluating Facial Expression Recognition Datasets for Deep Learning: A Benchmark Study with Novel Similarity Metrics. arXiv preprint arXiv:2503.20428.

33. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. 2010 ieee Comput Soc Conf Comput Vis pattern recognition-workshops. 2010. https://doi.org/10.1109/CVPRW.2010.5543262.
34. Barsoum E, Zhang C, Ferrer CC, Zhang Z. Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution 2016. https://doi.org/10.1145/2993148.2993165
35. Mollahosseini A, Hasani B, Mahoor MH. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput. 2017;10(1):18–31.
36. Schaefer A, et al. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. Cogn Emot. 2010;24(7):1153–72.
37. lovenails. (2025, November 22). ASMR Nails on a Chalkboard [Video]. YouTube. https://www.youtube.com/watch?v=nKdJACJm4qM
38. Panogs LLC. (2025, November 22). TEDx Portland Standing Ovation 360º VR Experience [Video]. YouTube. https://www.youtube.com/watch?v=7d7NyXxEbak
39. The Phantom Camera. (2025, November 22). Slow Motion Mosquito Flight [Video]. YouTube. https://www.youtube.com/watch?v=9uBuKDkyVY
40. HQSoundEffects. (2025, November 22). Mosquito – Sound Effect [HQ] [Video]. YouTube. https://www.youtube.com/watch?v=AMg-lgW7Dzg
41. TheSilentWatcher. (2025, November 22). Relaxing River Sounds – Peaceful Forest River – 3 Hours Long – HD 1080p [Video]. YouTube. https://www.youtube.com/watch?v=IvjMgVS6kng

Publisher's Note