Audio Engineering Society

Convention Express Paper 10267

Presented at the 160th Convention

2026 May 28-30, Copenhagen, Denmark

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

JoyCam: Blending Facial Recognition with Neural Activity measurement for Real-time Estimation of Listener Emotion

Duncan Williams[1]

[1] *Acoustics Research Centre, University of Salford, Manchester UK, M5 4WT*

[2] *Directorate of Psychology and Sport, School of Health and Society, University of Salford, Manchester UK, M5 4WT*

Correspondence should be addressed to Duncan Williams (d.a.h.williams@salford.ac.uk)

ABSTRACT

The ability to objectively measure listener emotion is a critical frontier for adaptive audio systems, healthcare, and personalized music therapy. While music is a powerful driver of affect, traditional self-reporting is often intrusive or inaccessible for users in wellbeing settings who may struggle to articulate their mood. This paper introduces JoyCam, a multimodal system that estimates subtle moments of joyful engagement by blending lightweight brain-wave monitoring (wearable EEG) with facial-expression sensing. By capturing physiological reactions that occur below the threshold of conscious awareness, the system creates a more stable emotional profile than single-modality methods. In our system, Facial joy is estimated via MediaPipe landmark analysis, focusing on normalized mouth-width deviations. Simultaneously, neurological engagement is tracked through Frontal Alpha Asymmetry (FAA) using an OpenBCI Cyton system. To address the sensitivity of EEG to movement, a dynamic artefact index down-weights neural signals during high-frequency interference. The system was tested in a pilot study with five participants. Preliminary results indicate that baseline-corrected physiological scores align closely with self-reported music impact and valence ratings across joyful and sad conditions. These findings suggest that JoyCam offers a robust framework for responsive musical companions that can adjust playlists or production parameters based on a listener's real-time physiological state

1 Introduction

The ability to accurately quantify listener affect in real-time represents a significant advancement for the field of audio engineering, particularly within the domains of adaptive music systems and healthcare. While music is a profound driver of human emotion [1], traditional methods of measuring its impact have historically relied on retrospective self-reporting, such as Likert scales or valence-arousal grids. These methods, while informative, are inherently intrusive

as they interrupt the listening experience, and are often inaccessible for individuals in wellbeing settings who may find it difficult to recognize or articulate changes in their internal state. Recent developments in wearable technology and computer vision have opened new pathways for non-obtrusive affective monitoring. The historical application of electroencephalography (EEG) in affective computing has traditionally relied on high-density, clinical-grade systems characterized by wet electrodes and cumbersome wiring. While these systems provide high signal resolution, their restrictive nature and the requirement for conductive gel create significant barriers to the study of music and emotion in naturalistic settings. The evolution toward wearable EEG technology has been driven by the need to bypass these constraints, moving toward lightweight, wireless, and dry-electrode configurations that prioritize user comfort and mobility [2]. Transitioning away from bulky clinical equipment is particularly necessary for listening situations, as the physical presence of a traditional EEG cap can introduce discomfort and self-consciousness, potentially altering the very emotional states being measured. By utilizing unobtrusive wearables, researchers can capture more authentic physiological correlates of joyful engagement and other affective responses in real-world environments, transforming the EEG from a static diagnostic tool into a transparent interface for personalized audio experiences. However, single-modality systems often struggle with reliability; facial recognition can be fooled by lighting or non-emotional movements, while electroencephalography (EEG) is notoriously sensitive to motion-based noise (such as that created by nodding or tapping along to music when listening). This paper introduces JoyCam, an experimental system designed to overcome these limitations by blending lightweight brain-wave monitoring with simple facial-expression sensing.

The primary motivation for JoyCam is to move toward an "objective listener" model, where the audio system acts as a quiet musical companion. In healthcare and wellbeing contexts, this supports a more stable picture of emotional responses than either method could provide on its own. Specific applications include:

Adaptive Playlisting: Automatically adjusting musical style, tempo, or genre based on a listener's real physiological responses rather than generic preferences.

Healthcare Support: Assisting individuals who struggle to describe their mood by tracking subtle, positive engagement as it happens.

Self-Reflection Tools: Functioning as a long-term tracker to facilitate self-reflection and emotional journaling, similar to popular sleep-tracking devices.

This paper outlines the technical implementation of the JoyCam system and presents results from a five-participant pilot study, where we analyze the alignment of baseline-corrected physiological scores with self-reported music impact across joyful and sad conditions, where participants were asked to bring their own selection of music they found "joyful" or "sad". Neutral music was selected by the experimenter from a "muzak" library.

2 Methods and Experimental Design

The architecture of JoyCam is designed to capture a stable picture of emotional responses by blending two physiological sources. This multimodal approach addresses the inherent limitations of single-source methods, such as the sensitivity of electroencephalography (EEG) to motion and the potential for lighting to disrupt facial recognition.

2.1 Facial pattern recognition

The vision module utilizes the MediaPipe FaceMesh [3] framework to extract a high-density mesh of 468 facial landmarks from standard webcam video. Rather than employing a "black box" emotion classifier that might struggle with camera angles or lighting, JoyCam identifies changes in facial shape that correlate with positive expressions. The system focuses specifically on the geometric relationship between the mouth corners (landmarks 61 and 291). To ensure the metric remains stable regardless of participant scaling or proximity to the sensor, the raw mouth width is normalized by the inter-ocular distance. This geometric approach is more than a simple smile detector as it treats expression as a physiological correlate rather than a binary state. By measuring these facial landmarks, the system captures subtle reactions that participants may not consciously notice [4]. A baseline is established during a neutral calibration period, and deviations are mapped to a continuous joy score between 0 and 1. To minimize frame-to-frame noise and tracking jitter, the score is processed using an exponential moving average (EMA).

2.2 Neurological Signal Capture and Processing

Previous work has shown that emotionally powerful, pleasant music elicits increased left-lateralized frontal alpha asymmetry relative to neutral or unpleasant auditory stimuli, consistent with music-induced positive affective engagement [5]. In Joycam, the neural stream is captured using an OpenBCI Cyton system configured to stream time-series data via the Lab Streaming Layer (LSL) protocol [6]. To ensure a continuous, high-resolution data stream from the OpenBCI Cyton board, the FTDI latency timer for the serial-to-USB interface was manually reduced from the default 16 ms to 1 ms, minimizing jitter in the incoming time-series data.. The electrode configuration focuses on frontal clusters, with the left cluster comprising Fp1 and FT9 and the right cluster comprising Fp2 and FT10. Signals undergo digital filtering, including a 4th-order Butterworth band-pass filter (1 to 45 Hz) and a 50 Hz notch filter to eliminate mains interference.

The system calculates Frontal Alpha Asymmetry (FAA) as the primary neurological correlate of emotion using Welch's method to estimate power spectral density (PSD) [7]. An Individual Alpha Frequency (IAF) calibration routine is implemented during an initial 20 second neutral baseline to identify the user's specific alpha peak [8]. The alpha power is then integrated across a personalized band defined as [IAF−2, IAF+2]∩[6, 14] Hz15. The FAA score is computed as the log-ratio of the mean alpha power from the left and right clusters16:

$$FAA = log(P_{a,l}) - log(P_{a,r}) \quad (1)$$

Because alpha power is inversely related to cortical activation, negative FAA values indicate higher left frontal activation, which is associated with joyful or approach-related states [9]

2.3Combining using dynamic weighting

To create a single estimate of listener emotion, the system fuses the vision and neural scores using a dynamic weighting algorithm. The EEG weight is adjusted based on a dynamic artefact index that monitors high-frequency activity (20 to 40 Hz) sensitive to facial muscle movements, cognitive engagement, and noise. When excessive beta power is detected from blinks or jaw clenches, the system automatically reduces the influence of the EEG signal. The final fusion weight is recomputed at each step as follows:

$$w_{eeg} = max(w_{min}, (W_{eeg} \times EEG_conf) / (W_{eeg} \times EEG_conf + W_c_{am} \times Vision_conf)) \quad (2)$$

This ensures the system relies more heavily on the stable vision metrics during periods of high neural noise, while a floor prevented the EEG signal from being excluded entirely.

Figure 1 shows the system in use

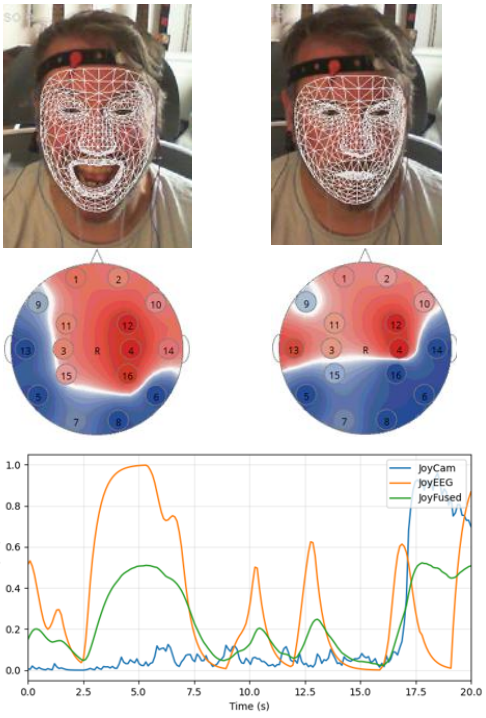


Figure 1. Above: Facial detection in "joyful" and "sad" conditions. Middle: EEG topograph of the same conditions. Bottom: Combined metric from facial recognition, EEG signal, and artifact removal to a single "joy estimate".

2.4Pilot study

The pilot study utilized a within-subjects design with five participants to validate the system's sensitivity to emotional shifts. While a cohort of five participants may appear limited in scale for a traditional listening test, it represents a more substantial data set within the context of high-density EEG research. The technical complexity and time requirements for establishing reliable, low-impedance electrode connections with the OpenBCI Cyton system often necessitate smaller, high-fidelity pilot groups to

ensure signal integrity across all modalities. Participants provided their own selection of music they personally characterized as joyful and music they considered to be the opposite descriptor, specifically sad. The experimenter provided a third category of neutral music sourced from a library of "muzak" to serve as an emotional baseline.

The testing session followed a standardized protocol that began with a 20-second neutral calibration period. During this phase, the system recorded baseline physiological data to establish the individual alpha frequency and median values for facial landmarks. Following this calibration, the joyful, sad, and neutral tracks were presented while the JoyCam system logged synchronized data via the Lab Streaming Layer protocol. To verify the emotional effectiveness of each track and provide a ground truth for sensor validation, participants provided post-stimulus ratings for valence, arousal, and music impact.

3 Results

The results of the pilot study demonstrate the technical feasibility of the JoyCam system and provide an initial validation of its multimodal architecture. While the limited sample size of the pilot restricted the statistical power of the physiological sensors, the alignment between these metrics and the significant subjective reports suggests that the system successfully tracks affective shifts.

The experimental design utilized Music Impact and Perceived Joy as ground truth measures to verify that the musical stimuli successfully evoked the intended emotional states. A paired t-test revealed that Music Impact ratings were significantly higher for the joyful condition, with a mean of 6.2, compared to the sad condition, which had a mean of 3.6, resulting in a p-value of 0.025. Similarly, the baseline-corrected Perceived Joy scores showed a significant distinction with a p-value of 0.025, where joyful tracks produced a positive mean deviation of +1.2 and sad tracks produced a negative mean deviation of -1.4 from the neutral baseline. These significant findings serve as a critical manipulation check, confirming that participants experienced distinct emotional states that align with the sensor data targets.

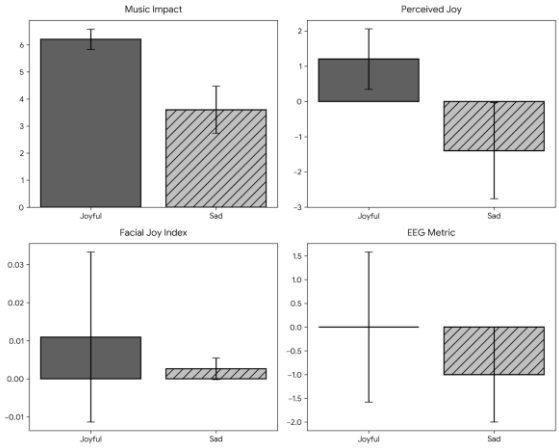


Figure 2. Both Music Impact and Perceived Joy show a statistically significant difference p = 0.025 between the joyful and sad conditions. While the physiological metrics show directional alignment with the subjective reports, the delta for Facial Joy and the EEG Metric did not reach significance error bars (Standard Error of the Mean) and significance brackets for the markers where $p < 0.05$.

The physiological metrics, presented as baseline-corrected deviations, exhibited directional trends that correspond to these significant subjective shifts. The Facial Joy Index resulted in a positive mean deviation of +0.011 in the joyful condition, while the sad condition showed a minimal shift of +0.003. Neural responses measured through the EEG metric showed a mean negative shift of -1.0 for the sad condition relative to the joyful condition mean of 0.0. The higher p-values for these specific metrics, recorded at 0.721 and 0.704 respectively, are attributed to the small sample size rather than a lack of directional sensitivity in the JoyCam hardware.

Metric	Joyful Condition (Mean)	Sad Condition (Mean)	p-value (Paired)
Music Impact	6.2	3.6	0.025*
Perceived Joy	+1.2	-1.4	0.025*
Facial Joy Index	+0.011	+0.003	0.721
EEG Metric	0	-1	0.704

Table 1. summarizes mean and paired p-values across subjective and physiological metrics, *

indicates significance

4 Discussion

Please The statistical significance achieved for both Music Impact and Perceived Joy confirms that the experimental protocol effectively induced the intended joyful and sad emotional states among participants. While the physiological sensors did not reach the threshold for statistical significance in this small-scale cohort, the directional trends observed in the facial and neural data align closely with these significant subjective markers. This alignment suggests that the system is successfully capturing the physiological underpinnings of music-induced affect, with higher statistical power likely to be achieved in future studies with larger participant groups.

A critical technical finding of the pilot was the stability of the fused joy score during periods of physical movement. The system utilized a dynamic artefact index to effectively identify high-frequency beta surges between 20 and 40 Hz caused by blinks and minor facial adjustments. During these events, the confidence-weighted fusion recomputed the weights, reducing the EEG contribution to a floor of 0.2. This allowed the system to maintain a reliable joy estimate based on the facial landmark stream, preventing the noise-induced fluctuations that often compromise single-modality EEG systems. This multimodal resilience is essential for real-world audio applications where pure neurological monitoring is frequently interrupted by muscular noise or environmental interference.

The implementation of baseline correction was fundamental to the system's ability to standardize data across participants with varying resting states. By focusing on the delta relative to an initial neutral calibration, the system isolated the specific impact of the musical stimuli from individual baseline facial geometry or neurological activation levels. This approach is particularly relevant for longitudinal affective monitoring and therapeutic applications, as it enables the detection of subtle shifts in engagement that might otherwise be obscured by inter-subject variability. The data indicates that even in cases where raw sensor values were divergent, the baseline-corrected scores moved in a consistent direction for the joyful tracks compared to the sad tracks.

5 Conclusions and further work

This paper has presented JoyCam, a multimodal system designed for the real-time estimation of listener emotion through the integration of facial landmark analysis and electroencephalography. The technical architecture successfully demonstrated that blending vision-based geometric proxies with Frontal Alpha Asymmetry provides a more resilient measure of affect than single-modality approaches. By implementing a dynamic confidence-weighted fusion, the system maintained stability during muscular artefacts that typically compromise wearable EEG monitoring. The pilot study with five participants confirmed the effectiveness of the stimulus selection and baseline-correction protocols, as evidenced by the statistically significant differences in reported Music Impact and Perceived Joy between the joyful and sad conditions.

The findings from this research offer a promising foundation for the development of non-intrusive wellbeing and healthcare tools. Because JoyCam captures physiological engagement that may occur below the threshold of conscious awareness, it holds significant potential for individuals who find it difficult to recognize or describe changes in their internal state. In clinical or domestic settings, the system could act as a long-term reflection aid to help users identify patterns in their emotional landscape. Furthermore, the principles of real-time affect sensing enable the design of personalized musical companions that automatically adjust audio parameters based on a listener's actual physiological responses. Such tools could learn which specific genres or tempos consistently evoke signals of joy, functioning as a responsive and adaptive interface for stress reduction or focus.

The roadmap for future research involves scaling the pilot study to a larger and more diverse participant group to achieve greater statistical power for the physiological sensors. Expanding the dataset will allow for the refinement of the multimodal weighting algorithm and the exploration of more complex facial landmarks beyond mouth width. The system already allows for an additional functionality not evaluated here – outputting control data via OSC and MIDI CC. To facilitate real-time interactivity, the system maps the fused joy score to OSC and MIDI CC outputs. This allows the affective data to directly modulate synthesis parameters or track volume within a Digital Audio Workstation (DAW), enabling a closed-loop system where the music responds instantly to the listener's physiological state. Therefore, future

iterations of the system will focus on integrating these affective estimates into closed-loop audio production environments, where production parameters can be modulated in real time to sustain a desired emotional trajectory for the listener. These developments represent a step toward a new generation of empathetic audio systems that are capable of understanding and responding to the subtle physiological nuances of human emotion.be placed either to the top or the bottom of the page.

References

[1] P. N. Juslin and D. Västfjäll, “Emotional responses to music: The need to consider underlying mechanisms,” Behavioral and Brain Sciences, vol. 31, no. 05, Oct. 2008, doi: 10.1017/S0140525X08005293.

[2] A. J. Casson, “Wearable EEG and beyond,” Biomedical engineering letters, vol. 9, no. 1, pp. 53–71, 2019.

[3] C. Lugaresi et al., “Mediapipe: A framework for building perception pipelines,” arXiv preprint arXiv:1906.08172, 2019.

[4] P. Ekman and W. V. Friesen, “Facial action coding system,” Environmental Psychology & Nonverbal Behavior, 1978.

[5] H.-A. Arjmand, J. Hohagen, B. Paton, and N. S. Rickard, “Emotional Responses to Music: Shifts in Frontal Brain Asymmetry Mark Periods of Musical Change,” Frontiers in Psychology, vol. Volume 8-2017, 2017, doi: 10.3389/fpsyg.2017.02044.

[6] C. Kothe and others, “Lab streaming layer (LSL),” 2014.

[7] J. A. Coan and J. J. Allen, “Frontal EEG asymmetry as a moderator and mediator of emotion,” Biological psychology, vol. 67, no. 1–2, pp. 7–50, 2004.

[8] W. Klimesch, “EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis,” Brain research reviews, vol. 29, no. 2–3, pp. 169–195, 1999.

[9] L. A. Schmidt and L. J. Trainor, “Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions,” Cognition & Emotion, vol. 15, no. 4, pp. 487–500, 2001.