Emotion and Expressivity in Music Performance: A Multimodal Approach

Natalia Kotsani
nkotsani@corelab.ntua.gr
National Technical University of Athens
Athens, Greece

Spyridon Kantarelis
spyroskanta@ails.ece.ntua.gr
National Technical University of Athens
Athens, Greece

Vassilis Lyberatos
vaslyb@ails.ece.ntua.gr
National Technical University of Athens
Athens, Greece

Edmund Dervakos
eddiedervakos@islab.ntua.gr
National Technical University of Athens
Athens, Greece

Giorgos Stamou
gstam@cs.ntua.gr
National Technical University of Athens
Athens, Greece

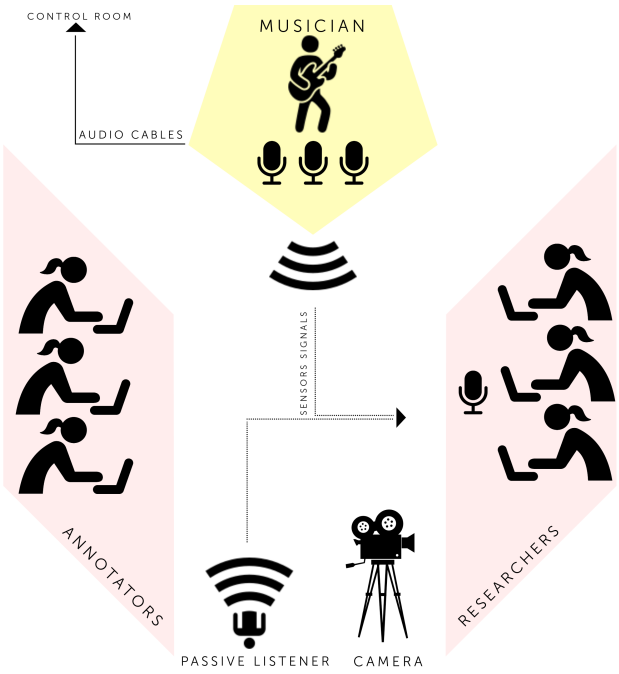


Figure 1: Diagram showing roles and connections during data collection.

Abstract

The relationship between musical performance, perceived, intended, and induced emotion by audience and musician is a complex topic that has been studied for centuries. In the era of Artificial Intelligence (AI), this topic becomes particularly important for the development of applications that assist in the study and creative aspects of musical performance, enhancing both the interpretative process and emotional expression. In this paper we present a protocol for creating datasets that contain live performances of music compositions, continuous and categorical emotion annotations from the audience and the performing musician, in addition to an array of biosignals recorded from performers and listeners. The protocol is designed to contain a variety of musical contexts, from rigid etudes simulating practice to free improvisation emulating pure expression. It is replicable and agnostic of instrumentation, and its main purpose is to facilitate the development of AI applications that will enhance musical expression. We provide the design and details of the protocol, preliminary results from its implementation with professional musicians and we discuss limitations and potential future research directions.

Keywords

Musical Expressivity, Multimodal Dataset Protocol, Biosignal Analysis, Emotion Annotation, Artificial Intelligence in Music

NIME '26, June 23–26, 2026, London, UK

1 Introduction

In recent years, the expanding use of AI applications in the fields of cultural industries and artistic creation has led to a growing interest in the development of new datasets centered on multimedia, particularly audio and visual data. Specifically in the field of

music, despite the immense interest in exploring the connection between human emotions and the art of musical expression and creation, as well as the creative mechanisms and expressive tools employed by artists to evoke these emotions, very few datasets have been developed to provide a fertile ground for conducting in-depth scientific research on these topics [12].

This study outlines the design and implementation of a novel protocol, serving as a guide for constructing comprehensive, open-source datasets along with preliminary results of its implementation. The proposed protocol defines various musical tasks to be performed at different levels of expressivity while capturing biosignals and emotion annotations from the audience and the performer. Research indicates that physiological responses, including heart rate, skin conductance, and brain activity, are closely tied to emotional experiences, whether they are consciously perceived or not [26]. In our implementation, the resulting dataset includes recordings of electroencephalogram (EEG), electrocardiogram (ECG), galvanic skin response (GSR), multi-source audio, scores, video, as well as real-time continuous valence-arousal annotations, categorical emotion labels and performers' interviews. While existing literature includes a broad range of datasets derived either from music collections or live recordings of pre-existing compositions that do not follow any specific recording protocol [20, 34], the organizational challenges and high resource demands of live recording (especially when an audience is involved) have resulted in a limited number [14, 16] of proposed recording protocols that focus on specific research hypotheses.

The choice of live performance, with dual annotations from both a small audience and the performer, despite the complexity of the data collection process, allows for capturing various parameters of expressivity and emotions in a context where it is more fluid and influenced by the performer [34]. It also enables the repeated performance of musical works with varying degrees of expressivity, the study of the emotions evoked in relation to each section of the musical piece, mode, or scale, as well as the examination of the emotional function of music in the contexts of score reading, personal repertoire, or an improvisational setting. While other works focus on listener-perceived emotions using categorical or continuous emotion annotations in a particular context (usually recorded compositions) [10, 19, 27, 30], our proposed approach captures both listener and performer emotions, across different contexts and modalities, with a particular interest in the correlation or the incongruity between them [9]. We argue that the resulting dataset will facilitate a deeper understanding of the link between musical expressivity and emotional response by integrating subjective annotations and biosignals in live performance settings. It will also enable AI applications that will enhance the process of performing music (e.g., modulating effects [7]), instead of applications focused on listening to music (e.g., conditional generative music).

Several emotion-labeled datasets exist that use music stimuli and biosignals. The *DEAP* dataset [15] records EEG and peripheral physiological responses from participants watching music videos, with self-reported valence and arousal ratings. Similarly, *DECAF* [1] incorporates magnetoencephalography (MEG), ECG, and other biosignals to analyze emotional responses to both music and movie stimuli. The *EEGLife* dataset [3] focuses on EEG-based emotion recognition using specifically composed musical pieces designed to evoke happiness, sadness, calmness, and anger. Additionally, studies utilizing the *MediaEval* dataset [21] have implemented machine learning approaches for music emotion recognition (MER), refining classification models to improve accuracy in valence-arousal space. These datasets serve as valuable resources for studying the relationship between biosignals and emotional responses to music. Unlike these datasets, which use pre-recorded stimuli or controlled compositions, our dataset captures multimodal biosignals and emotional responses in a live performance setting, where expressivity naturally emerges through performer-audience interaction.

To ensure the quality and usefulness of collected data, in addition to reproducibility and extensibility, the proposed protocol (described in Section 2) facilitates the participation of musicians across musical genres, instruments, and experience, while its rigidity enables capturing emotion and expressivity related data in a structured way. This results in a diverse, but structured dataset, supporting interdisciplinary research across different aspects of music and emotions. One of the primary challenges in constructing such datasets is ensuring alignment and synchronization across modalities, as each data type presents different temporal and structural characteristics [4, 22], in addition to ensuring diversity in musical genres and metadata completeness while addressing missing data and annotation inconsistencies [25]. Furthermore, the data collection conditions can have a significant impact on the validity of the dataset, especially given the sensitive nature of biosignals, thus when implementing the protocol it is imperative to ensure the minimization of stress. Our specific implementation that addresses these challenges is described in section 3. Finally, given the wealth of information available in such a dataset, there is a large variety of statistical and machine-learning methods that might be used to extract or infer potential patterns, correlations, or other interesting information [18]. As our data collection process is still ongoing, in Section 4 we provide some higher-level statistics for the data that we have gathered so far, while in Section 5 we discuss current limitations.

2 Data collection protocol

Our proposed protocol for data collection is designed to ensure consistency, alignment, and a well-balanced dataset. It addresses a specific research gap by capturing simultaneous physiological and emotional responses from both performers and listeners in

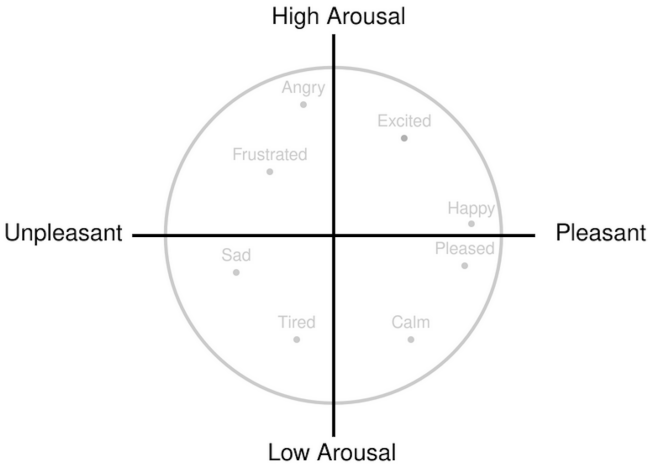


Figure 2: Russell's circle used for real-time valence-arousal annotation by the annotators.

live settings. To mitigate bias due to fatigue and annotation drift, the phases are executed in a randomized order.

2.1 Participants

The protocol includes performers, whose biosignals are recorded while they provide categorical emotion annotations, and audience members, who act either as annotators (providing both continuous and categorical annotations) or as passive listeners (whose biosignals are recorded and who provide categorical annotations).

2.2 Protocol's Recording Phases

Phase 0: Reception and Calibration

Phase 0 (PH0) serves as the preparatory stage, including participant reception, consent signing, and sensor calibration. Audience empathy is assessed using the Toronto Empathy Test [28], and their emotional state is measured with the PANAS scale [32]. Equipment and instruments are logged, participants are reminded that the session will be recorded, and sensors are attached with biosignals verified. Calibration tasks follow: PH0_T0 records resting-state biosignals for two minutes in darkness, while PH0_T1 assesses sensor response to performer movement as they play a single note. Performers may then take an optional three-minute warm-up (PH0_T2).

Phase 1: Common Repertoire

Phase 1 (PH1) comprises 9 tasks: 7 Etudes (PH1_T0-PH1_T6) composed by our team's professional jazz musician, each based on a different diatonic mode (Ionian, Dorian, Phrygian, Lydian, Mixolydian, Aeolian, and Locrian), selected for their systematic variation and their established association with distinct emotional characters in music perception research [29]. It also contains two renditions of Greensleeves; non-expressive (PH1_T7) and expressive (PH1_T8). This selection encompasses both major and minor modes, with Greensleeves featuring a minor-to-relative-major transition. Although cultural familiarity with Greensleeves may influence listener responses independently of performer expressivity, its widespread recognition offers a consistent reference point across participants. Musicians receive the sheet music in advance and perform each task for two minutes in their preferred key and tempo. This standardized approach facilitates cross-performer analysis and provides valuable data for AI applications and comparative studies.

Phase 2: Personal Choice Repertoire

In Phase 2 (PH2), performers record four tasks (PH2_T0–PH2_T3), performing two self-selected public-domain pieces each in non-expressive (PH2_T0, PH2_T2) and expressive (PH2_T1, PH2_T3) conditions for 2–3 minutes. They are instructed to choose pieces of varying difficulty. This phase enables informed expressive variation [11] through familiarity with the repertoire.

Phase 3: Improvisation

Phase 3 (PH3) consists of two tasks: free solo improvisation (PH3_T0) and improvisation over a sustained single-note drone selected by the performer[1]. Improvisation allows for unconstrained expressive variation and is commonly used to study creative musical behavior in less structured contexts [2]. This phase also enables analysis of differences in improvisational approaches across performers with varying levels of improvisation experience.

[1]No stylistic constraints were imposed; performers were free to improvise in any style or idiom.

2.3 Interviews

Short interviews are conducted after each performance phase (PH1_Q1, PH2_Q2, PH3_Q3), with a longer interview (Q4) at the end of the session to gather musically-driven annotations, evaluate proposed methods, and assess the musician's experiences. These interviews, recorded in audio and video, explore expressive techniques, stress or boredom, and provide insights on improvisation, repertoire choices, and familiarity and perspectives of AI in music performance.

2.4 Annotations

To investigate the correlation between expressivity and emotion, we included continuous and categorical annotations. Continuous annotations, based on Russell's circumplex model [24], capture moment-to-moment emotional changes. Categorical annotations, using the GEMS-9 scale (tension, power, joy, wonder, tenderness, transcendence, peacefulness, nostalgia, sadness) [8], provide a broader emotional assessment, with an optional neutral label.

2.5 Instructions to participants

The protocol provides clear instructions for participants, ensuring understanding of upcoming phases and tasks, and proper data collection. Performers were instructed to perform the non-expressive rendition with minimal dynamic variation, a steady tempo, and neutral articulation, functioning essentially as a mechanical reading of the score. The expressive rendition was left to the performer's interpretation. Participants are informed they may request breaks and are encouraged to approach the session like a live performance, without stopping for mistakes. A relaxed atmosphere is maintained, and volunteers are reminded they can leave at any time.

3 Implementation

Each recording session lasts approximately three hours. The first hour is allocated to briefing, soundcheck, sensor placement, and testing, while the remaining two hours are used for all experimental phases and interviews with scheduled breaks. Sessions take place in a university recording studio comprising a control room and a main recording room (Fig. 1). The control room hosts the sound engineer, while the recording room includes performers, audience members, and research staff. Data collection is supervised by three researchers handling biosignals and video recording. The audience consists of three annotators and one passive listener equipped with biosensors.

3.1 Data Collection

The data collection includes a recording studio with audio-visual equipment, laptops and smartphones for emotion annotation, and wearable sensors for physiological measurements. All data streams are synchronized through universal timestamps and specific audio cues during each performance session.

The recordings are conducted in a multi-channel format, utilizing high-quality A/D and D/A converters, a digital **audio** interface supporting MADI to ADAT conversion, and digital audio workstation (DAW) software. The preamplifiers used are high-gain, low-noise solid-state models, accommodating up to seven channels with precise gain control and transparent sound quality. In the recording room, each musician is assigned between one

and five microphones, depending on the directivity and acoustic characteristics of the instrument. The microphones are selected and positioned to optimize sound capture within the signal chain. Various polar patterns, including cardioid, hypercardioid, and omnidirectional, are used, with a frequency response of 20 Hz–20 kHz and a sensitivity of -45 dB. Each audio channel records in uncompressed mono WAV format, with 24-bit PCM encoding, a 48 kHz sampling rate, and a bitrate of 1152 kbps. Musicians each have a dedicated microphone for interviews, while a separate microphone is used by a researcher for control room communication and interview recording.

For **video** recording, a mirrorless camera with a 20.9 MP APS-C sensor is employed, capable of capturing 4K UHD video at 30fps. Mounted on a tripod, the camera is positioned to clearly capture the musicians' faces, hands, instruments, and any additional equipment during the performance, ensuring high-quality visual documentation of the session.

For the real-time valence-arousal **emotion** annotations by the audience, we developed a web application[2], where the user can track their emotions continuously using their mouse. The position of the mouse is sampled every 1 second, and is mapped to valence-arousal values. The app has three additional screens, a login/signup page, a page where the task/phase is chosen (before the continuous annotation starts), and a page with the categorical labels for the participant to choose from, in addition to the two Likert scales (enjoyment, familiarity).

Biosignals are collected using a headband EEG device with four dry electrodes placed at O1, O2, T3, and T4 locations, according to the International 10–20 System [3]. The device includes a reference electrode and a common sensor, sampling at 250 Hz. A separate wearable ECG sensor monitors breathing, and heart rate, featuring an electrostimulation function with a frequency of 1000 Hz. The sensor includes two disposable electrodes for body attachment. The GSR sensors are integrated into the same device as the ECG sensor for comprehensive physiological data collection, with data collected at a sampling rate of 500 Hz. All biosignals are transmitted via Bluetooth Low Energy to dedicated computers, which capture and store the data in JSON files, recording the transmitted values along with their corresponding timestamps.

3.2 Data organization

We developed a domain-specific ontology to organize the dataset as a semantic knowledge graph. This ontology defines hierarchical classes for human participants, data artifacts, and recording components, interconnected through object properties and enriched with contextual data properties. It also incorporates detailed equipment metadata and enables logical reasoning via RDF schema implementation, with provisions for future integration with existing music ontologies.

4 Preliminary Results

We validated the protocol using data from 16 recording sessions, which include physiological signals, musical features, and emotional annotations across different expressivity conditions, along with performers' demographic information. Sessions involved 16 unique solo performers performing grand piano, electric and classical guitar, electric bass, double bass, tenor saxophone, accordion, percussion, ney, cretan lyra, lute, violin.

4.1 Data Preprocessing

In the data preprocessing step, we applied various techniques to process and extract features from the ECG, audio, EEG, and GSR data. ECG preprocessing involved detecting R-peaks to calculate RR intervals and heart rate, with missing values in the RR intervals filled using linear interpolation. Based on the interpolated RR intervals, we computed the heart rate variability (HRV) and the Baevsky Index, resulting in two values at a frequency of 1 Hz. For audio feature extraction, we used Essentia [5], analyzing the signal with a frame size of 4096 samples and a hop size of 2048 samples, extracting Spectral Flux and Attack Slope features. For the EEG data, we extracted power from the Alpha (8-13 Hz), Beta (13-30 Hz), and Theta (4-8 Hz) bands and derived two features: the proportion of Alpha power to the total power of all three bands, and the proportion of Beta power to the total power of all three bands. For the GSR data, we calculated the Area Under the Curve (AUC) by integrating the GSR signal over time and determined the peak frequency by counting how often peaks occur in the GSR signal.

4.2 Phase Comparison

In Figure 3[4], we compare feature averages between listeners and performers, using PH1 as baseline. Both groups show similar patterns, particularly in alpha power, which peaks during personal repertoire (PH2) and improvisation (PH3), aligning with previous findings [17]. HRV also shows parallel trends, with higher values (indicating calmer states) during repertoire and improvisation phases.

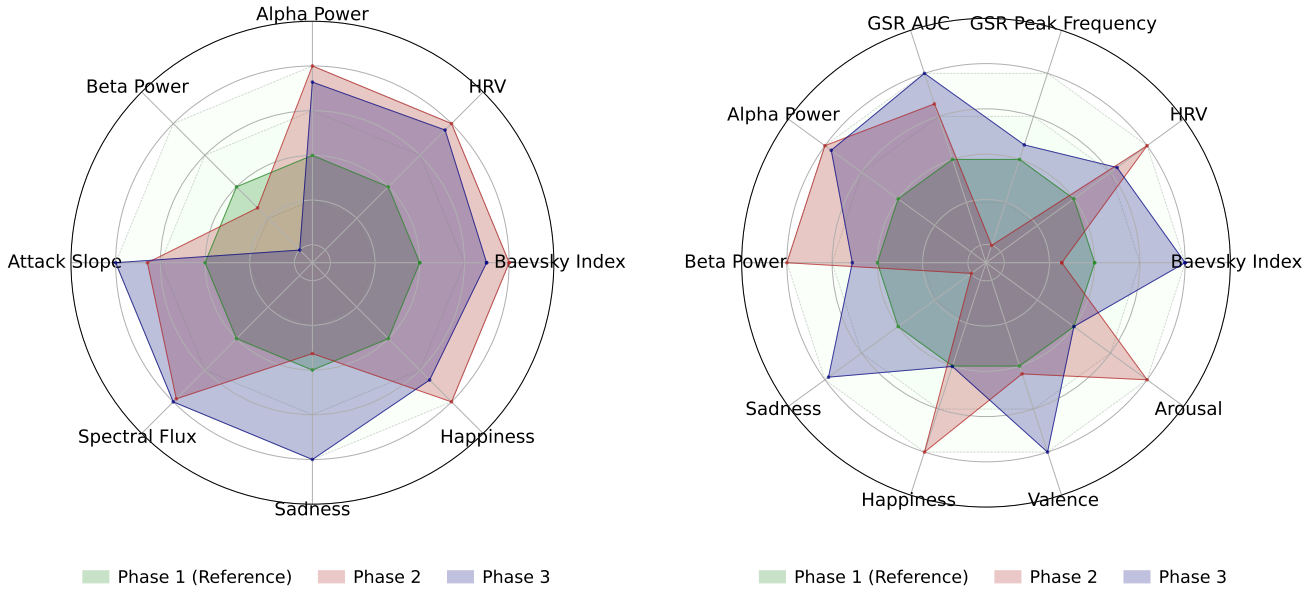
Conversely, some features differ between performers and listeners. Beta power is highest during PH1 for performers but lowest for listeners, suggesting different engagement patterns between unfamiliar pieces (PH1), personal repertoire (PH2), and improvisation (PH3). The Baevsky stress index also varies: performers show the lowest stress during the simple PH1 pieces, while listeners experience peak stress during improvisation sessions, possibly reflecting the musical tension-resolution patterns.

Another interesting difference between performer and listener appears in the categorical emotion labels "Happiness" and "Sadness". These were chosen to display as the two overall most frequent labels. Self-reported happiness appears more frequently in PH2 both for musician and listener, whereas for improvisation "Happiness" appears more frequently in performer responses when compared to listener responses. This could be the result of listeners choosing emotions such as "Wonder" or "Transcendence" instead, which the musicians might be reluctant to. Similarly "Sadness" appears very sparsely in listener responses, where they preferred to use labels such as "Nostalgia", whereas for musicians the most common label given to their own improvisation was "Sadness'.'

[2]https://witheflow.ails.ece.ntua.gr/annotator/

[3]Dry-electrode EEG systems present known signal quality limitations compared to gel-based research-grade devices, particularly in low-frequency bands [13, 23, 33]. The electrode placement used here (O1, O2, T3, T4) does not cover frontal sites associated with valence-related alpha asymmetry, limiting EEG-based inferences primarily to arousal-related features.

[4]In Figure 3, left and right panels correspond to performers and listeners, respectively. Feature values are averaged within each phase after per-participant normalization, providing a high-level comparison across experimental conditions; we acknowledge that this approach does not explicitly control for task duration differences or temporal structure within phases, and therefore results should be interpreted as phase-level descriptive trends. While multimodal physiological and emotional measures are presented jointly, their relationships are interpreted as exploratory co-variation rather than as outputs of a formal statistical model, with no causal claims implied.



(a) Performers. **(b) Audience Members.**

Figure 3: Comparison of the mean normalized values of multimodal features across different phases, using PH1 as the reference.

Figure 3a also depicts audio features averaged for the different phases, in particular Spectral Flux and Attack Slope. Spectral flux measures the magnitude spectrum of consecutive frames, where higher values indicate more changes in frequency content, where as expected values are higher for repertoire and improvisation when compared to the etudes. Similarly attack slope, that measures the steepness of the attack, where higher values indicate more percussive sounds and can relate to more rhythmic qualities of music, exhibits similar behaviour.

Finally, Figure 3b also depicts GSR features that were only captured by listeners. In particular, GSR peak frequency can be interpreted as the frequency of arousal events, whereas GSR AUC can be interpreted as the cumulative arousal across the signal. For both of these features the highest values appear for PH3, indicating similarly to the Baevsky index a more intense experience. On the other hand, PH2 has a higher value of GSR AUC and a lower value of GSR peak frequency, indicating a more constant state of arousal as opposed to improvisation where arousal seems to appear more in bursts.

4.3 Expressivity Comparison

In Figure 4, we analyze the multimodal features comparing expressive (PH1_T8, PH2_T1, PH2_T3) versus non-expressive (PH1_T7, PH2_T0, PH2_T2) musical performances. The physiological responses reveal distinct patterns between performers and listeners. For performers, Beta Power increases during non-expressive performances, suggesting heightened cognitive load, while the Baevsky Index shows reduced stress levels during expressive playing, indicating a more natural state of musical flow. In contrast, passive listeners exhibit elevated GSR measurements (both Peak Frequency and AUC) and increased HRV levels during expressive pieces, suggesting enhanced emotional arousal and engagement. Their emotional dimensions also show clear differentiation, with Valence and Transcendence measures displaying notable increases during expressive performances.

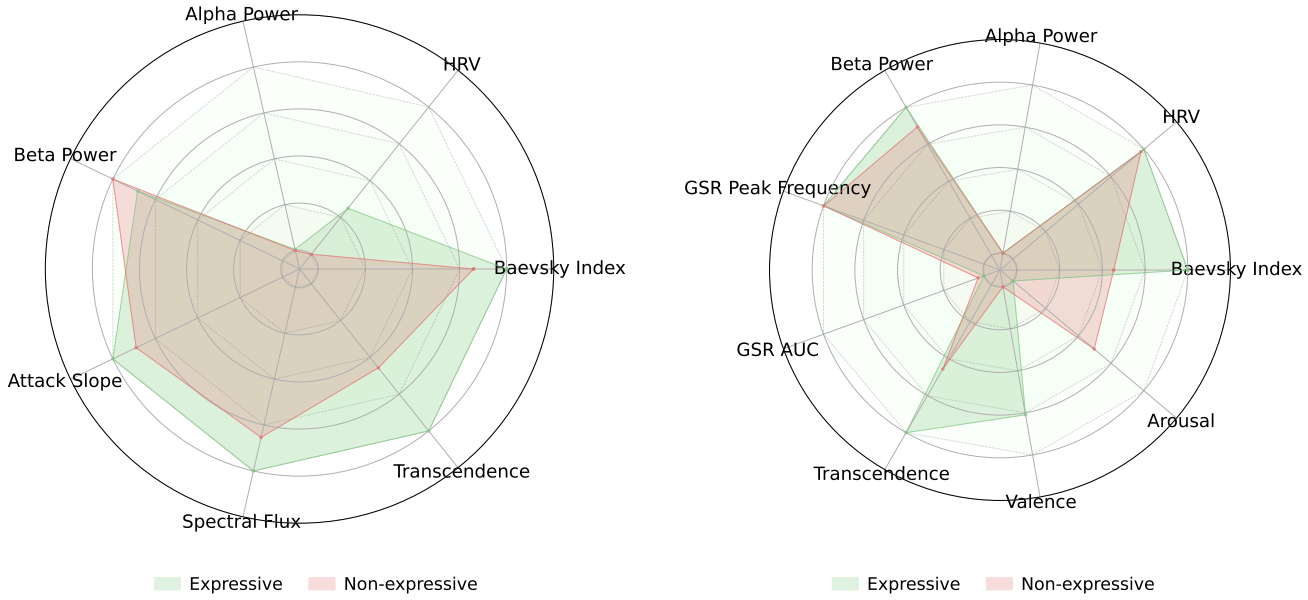
The acoustic analysis provides additional context through audio features that characterize the musical content itself. Expressive performances demonstrate consistently higher values in both Spectral Flux and Attack Slope, indicating richer timbral variations and more pronounced rhythmic articulation. These audio characteristics align with the physiological and emotional responses observed, suggesting that the increased musical expressivity manifests both in the acoustic properties of the performance and in the biological responses of both performers and listeners.

4.4 Demographics

The demographic data reveal a highly educated group of participating musicians, with 69.2% holding a master's degree and 15.4% holding a PhD. The participants' musical education backgrounds show diverse training paths: 23.1% studied at a conservatory, 23.1% at a university, and 38.5% attended both institutions, meaning 84.7% of participants received a formal musical education. The remaining 15.4% had no formal institutional training. Regarding musical capabilities, 61.5% have training in improvisation, and only 7.7% have no experience in reading music scores. The participants' years of experience in music varied across age groups, with the highest concentration of experience found in the 34-41 age range. Additionally, their familiarity with AI showed varying levels on a 1-5 scale, with most participants indicating moderate to high familiarity (levels 2-4), while fewer reported very low (level 1) or very high (level 5) familiarity.

5 Limitations

We treat expressivity as a performer-controlled variable rather than a theoretically defined construct, acknowledging that this operationalization captures intended expressivity rather than a universally defined attribute. The binary framing serves as a methodological simplification to enable comparison, while the



(a) Performers. **(b) Audience Members.**

Figure 4: Comparison of the mean normalized multimodal feature values across expressivity conditions.

continuous annotations and interviews are intended to capture the nuance that a binary label necessarily flattens.

The protocol has inherent Western music biases, evident in the choice of etudes and repertoire. While some Middle-Eastern/East Mediterranean music is included, most participating musicians have Western music training. Due to varying levels of familiarity with the seven modes specific compositions-etudes were created which would be played in repetition, and where the majority of the musicians read the etude from the score. While free modal improvisation with time-annotated transformations would be ideal, this was not feasible due to classical musicians' limited improvisation experience. Furthermore, in our implementation, the data we have collected so far exhibits some clear demographic biases of the musicians (sex skews male, age skews high), and of the audience members. Another important limitation of our implementation of the protocol is the quality of the biosensors that we use. These are commercial-grade sensors with a limited number of electrodes and resulting fidelity and are not medical grade, which makes the collected data less suitable for biomedical or other research that may require higher fidelity signals. Additionally, not all aspects of our experimental setting are fully controlled for as they would be in a medical study[5]. Finally, however, while our analysis suggests the validity of the collected data, we are unable to draw stronger conclusions at this stage.

6 Conclusion and Future Work

This work introduced a data collection protocol for multimodal datasets, including audio, video, scores, biosignals, and emotion annotations from both audience and performers. We implemented it with professional musicians in a recording studio and discussed preliminary findings. We are also adapting the protocol for ensembles, home-studio settings with amateur musicians, and different musical cultures, such as makams in Middle Eastern music and ragas in Indian classical music. Finally, we are exploring applications of these multimodal datasets for developing AI systems that could support expressive performance. Potential applications include automatic effect modulation based on real-time physiological and performance data streams [6, 31]. Such systems would utilize our multimodal emotion recognition framework to make interpretable inferences, while maintaining performer control over the augmentation parameters. This approach aims to enhance rather than constrain the natural aspects of musical expression, ensuring musicians retain agency over their performance.

7 Ethics Statement

This study adheres to ethical guidelines for human research, ensuring informed consent, data privacy, and transparency in data handling. Ethical approval was obtained from the local ethics committee of the National Technical University of Athens. Performers were appropriately compensated for their participation, while audience members contributed on a voluntary basis. To address copyright concerns, all recorded music is either in the public domain, original compositions by the research team, or performer improvisations, for which participants have agreed not to claim copyright, ensuring no copyright restrictions. These measures uphold ethical integrity and support the open accessibility of the resulting dataset. Biosignals are treated as personal health information, and they are captured, stored and transferred taking into consideration best practices for data security and transparency in handling. The societal impact of this work could be significant, facilitating the development of new tools that will enhance expressivity during musical performance. However this would entail that if such tools were developed commercially, some entity would have access to sensitive data of musicians, and without proper safeguards could be used for alternative purposes, or even maliciously.

[5]Physiological signals such as HRV, GSR, and EEG band power are non-specific and should therefore not be interpreted as direct measures of emotional or expressive state; the analyses are descriptive and exploratory.

Acknowledgments

The research project was implemented within the framework of the H.F.R.I. call "Basic Research Financing (Horizontal Support of All Sciences)" under the National Recovery and Resilience Plan "Greece 2.0," funded by the European Union – NextGenerationEU (H.F.R.I. Project No. 15111: Emotional Artificial Intelligence in Music Expression).

References

[1] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. 2015. DECAF: MEG-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing* 6, 3 (2015), 209–222.

[2] Roger E Beaty. 2015. The neuroscience of musical improvisation. *Neuroscience & Biobehavioral Reviews* 51 (2015), 108–117.

[3] Ferney Beltran-Velandia, Jonatan Gómez, Miguel Suarez, Andrés Ojeda, and Elizabeth León. 2022. Classification of Music-Evoked Emotions from EEG signals using Self-Organizing Maps. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. IEEE, 1–6.

[4] D. Bogdanov, N. Wack, and E. Gómez et al. 2019. MTG-Jamendo: Dataset and Baseline Experiments for Automatic Music Tagging. *IEEE Transactions on Multimedia* (2019). https://doi.org/10.1109/TMM.2019.2913661

[5] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Ricardo Zapata, Xavier Serra, et al. 2013. Essentia: An audio analysis library for music information retrieval.. In *ISMIR*, Vol. 13. 493–498.

[6] Edmund Grigoris Dervakos, Spyridon Kantarelis, Vassilis Lyberatos, Jason Liartis, and Giorgos Stamou. 2025. Go witheFlow: Real-time Emotion Driven Audio Effects Modulation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Creative AI Track: Humanity*.

[7] Giorgos Filandrianos, Natalia Kotsani, Edmund G Dervakos, Giorgos Stamou, Vaios Amprazis, and Panagiotis Kiourtzoglou. 2020. Brainwaves-driven Effects Automation in Musical Performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 545–546.

[8] Peer-Ole Jacobsen, Hannah Strauss, Julia Vigl, Eva Zangerle, and Marcel Zentner. 2024. Assessing aesthetic music-evoked emotions in a minute or less: A comparison of the GEMS-45 and the GEMS-9. *Musicae Scientiae* (2024), 10298649241256252.

[9] Javier Jaimovich, Niall Coghlan, and R Benjamin Knapp. 2010. Contagion of Physiological Correlates of Emotion between Performer and Audience: An Exploratory Study. In *International Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications*, Vol. 2. SCITEPRESS, 67–74.

[10] Javier Jaimovich, Niall Coghlan, and R Benjamin Knapp. 2012. Emotion in motion: A study of music and affective response. In *International Symposium on Computer Music Modeling and Retrieval*. Springer, 19–43.

[11] Kelly Jakubowski, Rainer Polak, Martín Rocamora, Luis Jure, and Nori Jacoby. 2022. Aesthetics of musical timing: Culture and expertise affect preferences for isochrony but not synchrony. *Cognition* 227 (2022), 105205.

[12] Jaeyong Kang and Dorien Herremans. 2024. Are we there yet? a brief survey of music emotion prediction datasets, models and outstanding challenges. *arXiv preprint arXiv:2406.08809* (2024).

[13] Daria Kleeva, Ivan Ninenko, and Mikhail A Lebedev. 2024. Resting-state EEG recorded with gel-based vs. consumer dry electrodes: Spectral characteristics and across-device correlations. *Frontiers in Neuroscience* 18 (2024), 1326139.

[14] R Benjamin Knapp, Javier Jaimovich, and Niall Coghlan. 2009. Measurement of motion and emotion during musical performance. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–5.

[15] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[16] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. 2018. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia* 21, 2 (2018), 522–535.

[17] Charles J Limb and Allen R Braun. 2008. Neural substrates of spontaneous musical performance: An fMRI study of jazz improvisation. *PLoS one* 3, 2 (2008), e1679.

[18] Vassilis Lyberatos, Spyridon Kantarelis, Edmund Dervakos, and Giorgos Stamou. 2024. Perceptual musical features for interpretable audio tagging. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 878–882.

[19] Vassilis Lyberatos, Spyridon Kantarelis, Eirini Kaldeli, Spyros Bekiaris, Panagiotis Tzortzis, Orfeas Menis-Mastromichalakis, and Giorgos Stamou. 2024. Crowdsourcing as a Pedagogical Tool in Computer Science Higher Education: a Case Study. *Human Computation* 11, 1 (2024), 47–70.

[20] Esteban Maestre, Panagiotis Papiotis, Marco Marchini, Quim Llimona, Oscar Mayor, Alfonso Pérez, and Marcelo M Wanderley. 2017. Enriched multimodal representations of music performances: Online access and visualization. *Ieee Multimedia* 24, 1 (2017), 24–34.

[21] Yesid Ospitia Medina, José Ramón Beltrán, and Sandra Baldassarri. 2022. Emotional classification of music using neural networks with the MediaEval dataset. *Personal and Ubiquitous Computing* 26, 4 (2022), 1237–1249.

[22] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters. 2018. DALI: A Large Dataset of Synchronized Audio, Lyrics, and Notes. In *ISMIR*. https://doi.org/10.5281/ZENODO.1492443

[23] Elena Ratti, Shani Waninger, Chris Berka, Giulio Ruffini, and Ajay Verma. 2017. Comparison of medical and consumer wireless EEG systems for use in clinical trials. *Frontiers in human neuroscience* 11 (2017), 398.

[24] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[25] M. Schedl, E. Gómez, and J. Urbano. 2014. Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval* 8, 2–3 (2014), 127–261. https://doi.org/10.1561/1500000042

[26] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A review of emotion recognition using physiological signals. *Sensors* 18, 7 (2018), 2074.

[27] M. Soleymani, E. M. Schmidt, Y. Yang, and B. P. Knoll. 2013. The MediaEval 2013 Brave New Task: Emotion in Music. In *Proceedings of the MediaEval Workshop*.

[28] R Nathan Spreng*, Margaret C McKinnon*, Raymond A Mar, and Brian Levine. 2009. The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of personality assessment* 91, 1 (2009), 62–71.

[29] David Temperley and Daphne Tan. 2012. Emotional connotations of diatonic modes. *Music Perception: An Interdisciplinary Journal* 30, 3 (2012), 237–257.

[30] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. 2008. Semantic Annotation and Retrieval of Music and Sound Effects. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 2 (2008), 467–476. https://doi.org/10.1109/TASL.2007.913750

[31] Yichen Wang and Charles Patrick Martin. 2025. Ai see, you see: Human-ai musical collaboration in augmented reality. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[32] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[33] Sarah N Wyckoff, Leslie H Sherlin, Noel Larson Ford, and Dale Dalke. 2015. Validation of a wireless dry electrode system for electroencephalography. *Journal of neuroengineering and rehabilitation* 12, 1 (2015), 95.

[34] Simin Yang, Courtney N Reed, Elaine Chew, and Mathieu Barthet. 2021. Examining emotion perception agreement in live music performance. *IEEE transactions on affective computing* 14, 2 (2021), 1442–1460.