IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 August 2025, accepted 28 August 2025, date of publication 1 September 2025, date of current version 5 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3604631

RESEARCH ARTICLE

Soundy: A Multimodal Audio Interface for Educational Music Production

PIETRO BUCCELLATO, (Graduate Student Member, IEEE), ANDREA BIANCO, (Senior Member, IEEE), AND CRISTINA ROTTONDI, (Senior Member, IEEE)
Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy
Corresponding author: Pietro Buccellato (pietro.buccellato@polito.it)

This work was supported in part by the Project PNRR-NGEU under Grant MUR-DM 118/2023; in part by the "Musical Metaverse: An Inclusive Extended Reality Platform for Networked Musical Interactions" Project, funded by European Union–Next Generation EU within the PRIN 2022 Program (D.D. 104-02/02/2022 Ministero dell'Università e della Ricerca) under Grant 2022CZWWKP; and in part by the MUSMET Project funded by the EIC Pathfinder Open Scheme of the European Commission under Grant 101184379.

ABSTRACT Multimodal interaction has emerged as a promising approach to enrich user experience and foster engagement in Educational Music Production (EMP) environments. Traditional audio interfaces and Digital Audio Workstations (DAWs), while powerful, often present engagement challenges, particularly for beginners, and tend to limit opportunities for spontaneous creativity and hands-on exploration during the learning process. To overcome these limitations, this paper introduces Soundy, a multimodal audio interface that integrates an on-board, lightweight web-based DAW. The system is accessible via a standard browser through a Local Area Network (LAN) and allows users to control audio production tasks using predefined and custom voice and facial commands. A comparative within-subject study was conducted involving 20 participants, each of whom tested both Soundy and a standard configuration composed of a Behringer UMC404HD audio interface and the Reaper DAW. The evaluation focused on two key dimensions, usability and engagement, measured using two standardized and validated instruments: The Post-Study System Usability Questionnaire (PSSUQ) and the User Engagement Scale - Short Form (UES-SF), complemented by open-ended participant feedback. Results revealed a trade-off between configurations: the standard setup ensured higher usability and task efficiency, while Soundy promoted greater engagement, creativity, and exploratory behavior. These findings suggest that embedded multimodal solutions based on voice and facial interaction hold strong potential for enhancing student experience in EMP and support future developments in adaptive command recognition and integrated hardware design.

INDEX TERMS Audio interface, educational music production, multimodal, usability, engagement.

I. INTRODUCTION

Music production (MP) is a structured process that transforms a musical idea into a finalized product, ready for distribution or performance. This process is typically divided into three main phases: writing, production, and release (Fig. 1). This study addresses aspects related to the *production* phase, which includes tracking, mixing, and mastering activities. These tasks heavily rely on both software and hardware tools [1], including audio interfaces, which are essential for capturing and manipulating sound. In the context of Educational Music Production (EMP), the *production* phase offers a particularly rich opportunity for students to engage with music technology in a hands-on, creative, and immersive way. Its pedagogical value has been widely recognized: production-oriented activities foster creativity, deepen engagement, and help bridge the gap between classroom learning and students' real-world musical experiences [2], [3]. Furthermore, research in music education emphasizes that experiential approaches, commonly referred

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

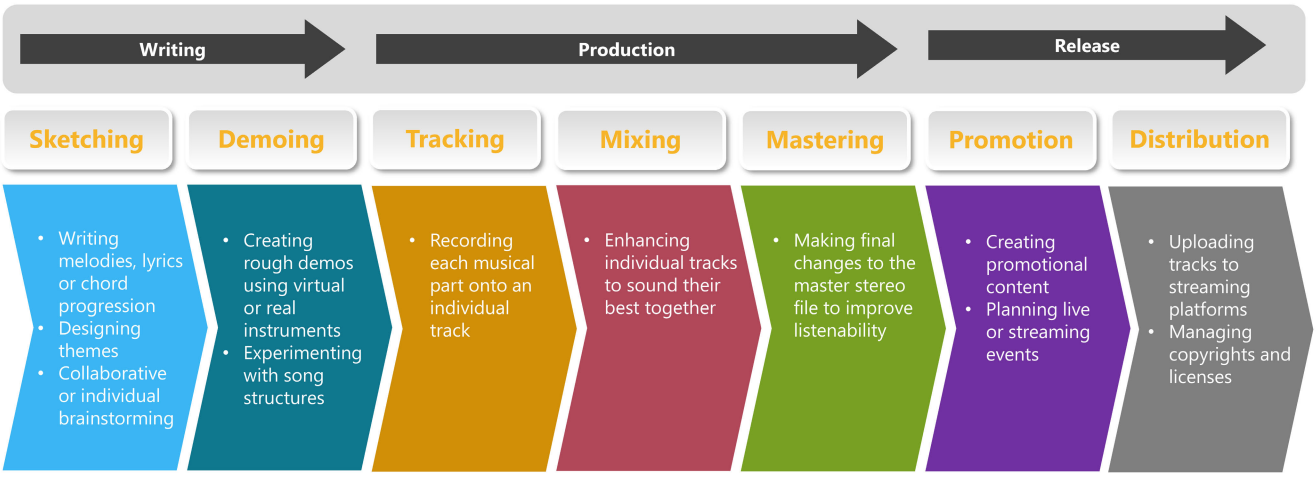

FIGURE 1. The schematic presents the music MP workflow as a hierarchical and sequential process, starting from the top with macro-level tasks and progressively decomposing each into its corresponding micro-tasks. The process begins with the writing phase, which includes sketching and demoing, representing the creative ideation and initial structuring of musical ideas. It then proceeds to the production phase, where tracking, mixing, and mastering are performed to technically refine and shape the audio content. Finally, the release phase includes distribution and promotion, aimed at delivering the final product to the audience and maximizing its visibility. This representation reflects a typical end-to-end pipeline in contemporary MP environments.

to as ''learning-by-doing'', are highly effective in supporting the development of technical skills, artistic expression, and long-term motivation [4]. This emphasis on experiential, student-centered approaches naturally aligns with the adoption of multimodal learning tools. By integrating auditory, visual, and tactile channels, such tools have been shown to improve user engagement and enhance understanding in educational music environments [5], [6]. Moreover, studies in Human-Computer Interaction (HCI) have demonstrated how multimodal systems enrich music experiences even for users with sensory limitations [7], [8] suggesting their broader potential in inclusive and dynamic learning environments. While many widely used commercial audio interfaces, such as the MOTU 828,[1] IK Multimedia iRig Pro Quattro I/O,[2] and Audient EVO 4,[3] offer high-quality signal processing capabilities, they are primarily designed for professional use and do not include features intended to support educational applications based on multimodal interaction. Furthermore, from a research perspective, existing studies on multimodal interaction in educational music technology often address isolated aspects of learning or focus on specific interaction modes in controlled settings, resulting in a lack of integrated solutions designed to support the entire *production* workflow.

Within this framework, and to address these challenges, Soundy is introduced as a multimodal audio interface specifically designed for EMP contexts, aiming to explore how such interfaces can enhance engagement as a pathway to meaningful and sustained musical learning throughout the *production* process. To assess the effectiveness and perceived value of the final system, a representative group of students was selected to participate in hands-on testing and complete validated evaluation questionnaires focused on usability and engagement, through a comparative analysis between Soundy and a standard setup representative of commonly used technologies for performing production tasks. Both configurations were well accepted by participants and proved suitable for completing the assigned tasks. However, notable differences emerged in how students interacted with each system. While the conventional setup was generally favored for its familiarity and ease of use, Soundy promoted higher levels of engagement. A music teacher was involved throughout the design, development, and experimental phases, providing continuous input on user experience, pedagogical relevance, and contributing to the planning and supervision of the evaluation sessions.

The remainder of this paper is organized as follows. Section II reviews related work on multimodal interaction in EMP. Section III presents the system design methodology and outlines the experimental validation protocol. Section IV describes the system architecture, including hardware, firmware, and software components. Section V reports the results of the user study. Section VI discusses the findings in depth, addressing limitations and future research directions. Finally, Section VII presents the conclusions.

II. RELATED WORK

A growing body of research has explored how multimodal interaction can enhance EMP experiences. These systems extend traditional audio production interfaces by integrating alternative input and output modalities, in order to accommodate diverse user needs and learning profiles, foster creativity, and promote greater learner engagement. The following subsections examine the primary modalities

[1] https://motu.com/en-us/
[2] https://www.ikmultimedia.com/
[3] https://evo.audio/

adopted in the design of such multimodal tools within educational audio contexts. The emphasis lies on how and where these multimodal tools are applied, highlighting the diversity of learning scenarios rather than categorizing the individuals involved. Students are considered broadly, regardless of age, physical condition, or formal enrollment status. Following this overview of the main interaction modes, the discussion will shift to identify current limitations across these approaches. These insights will lay the foundation for introducing Soundy.

A. TANGIBLE MODE

Tangible user interfaces (TUIs) allow users to interact with digital audio systems through the manipulation of physical objects. These systems often include knobs, sliders, blocks, or custom hardware that directly map to sound parameters. By externalizing digital processes into the physical world, TUIs facilitate hands-on control over musical structure, effects, and sequencing. One example is provided by [6], which describes the ImproviSchool project. In this study, primary school students engaged in a composition activity using a TUI composed of colored wooden blocks placed on an interactive surface. Each block was associated with specific musical parameters such as rhythm, texture, or intensity, and students used these elements to create soundtracks for narrative episodes presented as illustrated storyboards. The system provided real-time auditory and visual feedback as students manipulated the physical components, enabling a direct and intuitive mapping between gesture, material interaction, and musical expression.

B. HAPTIC MODE

Haptic interaction involves the use of vibrotactile or force-feedback devices to deliver tactile information in response to sound events or control actions. Wearable or surface-based systems can encode rhythm, dynamics, or spatial cues into vibration patterns. This modality enables perception of musical elements through touch and supports non-visual or silent operation. For example, [9] presents a system where a haptic interface converts MIDI data into tactile vibrations applied to the palm. Students were able to perceive melodic and rhythmic structures solely through vibrotactile feedback. Another perspective is offered in [10], which explores a wearable-based approach through the development of haptic prototypes co-designed with participants to enhance synchronization during ensemble performance. The system was evaluated in real-time musical settings and demonstrated that vibrotactile cues can effectively support precise timing and coordination between performers.

C. VISUAL MODE

Systems that translate musical structures into intuitive visual elements, such as blocks, shapes, or color-coded components, can help students grasp abstract concepts like harmony, scales, and chord progressions more effectively. Embodying these ideas, ArchiTone, a LEGO-inspired gamified system grounded in constructivist principles, was designed to make music theory more accessible and engaging for students [11]. ArchiTone visualizes fundamental music theory concepts through symbolic shapes: squares for tonic chords (stability), triangles for dominant chords (tension), and circles for subdominant chords (smooth transitions). These shapes are used to build musical ''blocks'' that represent chords and ''buildings'' that illustrate harmonic progressions. The system features two modes: Learning Mode, where users explore theory through familiar compositions and reconstruct them using blocks; and Creation Mode, where students freely combine blocks to create new harmonic structures.

D. VOICE-BASED MODE

Voice-based interaction uses speech recognition to issue commands, trigger functions, or input values within music software. It is typically implemented through natural language processing or predefined command vocabularies. Voice input allows direct control of actions such as playback, recording, track selection, or effect toggling without manual input. An example is provided in [12], which presents a voice-enabled music training system for middle school students. The proposed platform uses speech recognition to identify sung melodies and assess vocal performance, offering immediate, automated feedback on pitch accuracy and melodic correctness. The system supports repeated, self-guided practice and allows teachers to monitor student progress through performance records and evaluation data.

E. MOTION-BASED MODE

Motion-based interaction enriches music education by introducing a multimodal experience that actively engages students' bodies in the learning process. Various systems based on gesture recognition or motion tracking have been developed to exploit this paradigm. These systems leverage the connection between sensorimotor activity and cognitive development, allowing students to internalize theoretical content through embodied interaction with sound. By using gesture-driven interfaces, students can manipulate musical parameters such as rhythm, pitch, and dynamics through natural movements. A recent example is Music Corner [13], a gesture-based rhythm game designed to democratize access to music education through full-body interaction. By drawing on solfège hand signs and embodied movement, the system aims to enhance users' rhythmic awareness and timing skills, particularly supporting students who benefit from physical engagement with musical concepts.

F. VIRTUAL AND AUGMENTED REALITY MODE

Immersive environments based on virtual reality (VR) have been proposed to support experiential learning in music education. These systems simulate studio environments or musical contexts in three-dimensional space, allowing

students to interact with virtual instruments, control signal chains, and explore spatial audio through embodied interaction. One example is the VR-based Music Education Experience (VR-MEE) system [14], which combines VR with artificial intelligence models to analyze and enhance the educational process. The system allows students to actively engage with music in a virtual environment, where they could perform tasks like playing instruments, singing, or moving to rhythm. While students interacted with the VR setting, the system automatically observed and analyzed their actions, such as how accurately they kept rhythm or followed a melody. Based on these interactions, it provided personalized feedback to help them improve specific aspects of their musical performance, like pitch, timing, and expressiveness.

Differently, Augmented Reality (AR) offers a context-aware enhancement to physical learning environments by overlaying digital information in real time. Unlike VR's fully immersive spaces, AR enables students to engage with physical musical instruments and environments enhanced by interactive visual aids, such as gesture-based cues, score visualizations, and real-time feedback projected onto tangible surfaces. In [15], the use of AR was examined in a secondary school music technology classroom, where students used mobile devices to scan images and access overlaid animations illustrating the roles and connections of components in a public address system such as microphones, mixing desks, amplifiers, and speakers. The augmented content supported students in visualizing how these elements function together within a live sound setup.

G. BRAIN-COMPUTER MODE

Recent developments in Brain-Computer Music Interfaces (BCMI) have opened new possibilities for integrating neurophysiological data into EMP workflows. In this context, electroencephalography (EEG) is used as a means of interaction and as both a creative and educational tool, enabling students to engage with music composition through the direct modulation of musical parameters via their own brain activity. One example involves an educational model in which high school students are guided through the design and implementation of a system that maps EEG signals to musical output [16]. Students engage directly with the construction of the hardware, the programming of signal processing routines, and the generation of musical elements based on their own brain activity, exploring relationships between cognitive states, such as relaxation and concentration, and musical structure.

In a separate study, EEG data was used to drive the generation of polyphonic compositions within a framework that incorporates machine learning models for mental state classification [17]. Students trained the system to recognize patterns in their EEG signals corresponding to focused and relaxed states, which were then mapped to different rhythmic and dynamic parameters in the music generation process. This form of interaction enabled users to influence aspects of musical composition using real-time neural input, offering a multimodal and introspective approach to creative production.

Based on the presented key interaction modes, several gaps remain unaddressed in current approaches to multimodal EMP:

1) **Fragmented production pipeline**: several multimodal systems focus on isolated tasks within the *production* phase without supporting a coherent, end-to-end creative workflow that includes tracking, editing, and mastering.
2) **Low adaptability to creative preferences**: most systems offer limited configurability, making it difficult for students to align the tool's behavior with their individual styles, techniques, or evolving skill levels.
3) **Restricted spatial and situational flexibility**: current solutions often require a fixed, studio-based setup, limiting creative opportunities for users who need mobility, work across multiple spaces, or engage in collaborative or remote learning environments.
4) **Elevated cost of entry**: many systems require specialized or custom hardware, making them financially inaccessible for widespread use in educational settings and limiting opportunities for creative experimentation.

Soundy addresses the four core limitations discussed above in existing multimodal systems for EMP. Architecturally, it is conceived as an embedded system composed of three main components:

- **Hardware**: provides the physical interface for connecting musical input sources, as well as network and power connectivity;
- **Firmware**: manages low-level control of the hardware and handles the interaction logic;
- **Software**: represented by a lightweight web-based Digital Audio Workstation (DAW) hosted directly in the device's memory.

Soundy introduces two main contributions. The first one is the integration of multimodal interactions, allowing users to control different stages of the audio *production* process using voice and facial commands. These commands can be fully customized within the on-board web-based DAW, enabling users to adapt the interface to their own creative methods and preferences. This flexible input mechanism directly overcomes the fragmentation of the production pipeline by enabling seamless control over tracking, mixing, and mastering within a unified environment, and supports adaptability to individual creative preferences by fostering inclusive interaction and allowing for greater expressive autonomy.

The second contribution lies in the system's self-contained, microcontroller-based architecture, in which the entire DAW is locally hosted on the device and made accessible through a secure local area network (LAN) connection. This design enhances spatial and situational flexibility, enabling Soundy to function in mobile or collaborative educational settings.

At the same time, it reduces the cost of entry by leveraging off-the-shelf technologies and hardware components, as well as open-source software, ensuring financial accessibility and scalability for educational institutions.

Collectively, these design choices position Soundy as a comprehensive and practical solution for supporting multimodal interaction in EMP.

III. METHODS

This section presents the methodological approach adopted for the design and validation of the Soundy system. The first part explains all the motivations behind the definition of the design methodology for the project. The second part outlines the validation strategy, including the experimental setup, participant selection, and evaluation instruments used to assess usability and user engagement.

A. SYSTEM DESIGN

Soundy's design methodology was driven by high-level system requirements, defined to address the gaps outlined in Section II and categorized into functional and non-functional domains.

1) FUNCTIONAL REQUIREMENTS

- **Audio signal acquisition, digitization, and storage**: the system shall support the acquisition of analog audio signals from external sources such as musical instruments or microphones. It shall perform real-time analog-to-digital conversion of the incoming signals, and the resulting digital audio data shall be temporarily stored in the internal memory of the microcontroller for further processing and streaming.
- **Network connectivity and communication**: the system shall support Ethernet connectivity to enable integration within a LAN. It shall operate as an embedded HTTPS server, providing access to the user interface from any client device connected to the same local network, including smartphones, tablets, and personal computers.
- **Development and integration of the web-based DAW**: a web-based DAW shall be hosted directly on the embedded platform. The DAW shall provide real-time audio manipulation features, including gain adjustment, equalization, track recording, and mixing. It shall also support mastering operations and export of the final audio to standard uncompressed formats (e.g., WAV) through a standard web browser, without requiring any external software installation. Furthermore, the DAW shall incorporate multimodal interaction mechanisms, including predefined and customizable voice commands as well as facial gesture controls.

2) NON-FUNCTIONAL REQUIREMENTS

- **Low-cost system**: the system shall be implemented using low-cost hardware components and open-source software in order to ensure affordability and broad accessibility.
- **Scalability and extensibility**: its architecture shall be designed to support future extensions, including the addition of new input modalities, audio processing features, or increased numbers of analog audio inputs, without requiring major architectural modifications.

The emphasis on customizable inclusive interaction and remote usability necessitated a development approach capable of accommodating evolving specifications and frequent user-in-the-loop testing. For these reasons, an iterative and incremental (I&I) development model was adopted, upon which the Soundy design methodology (Fig. 2) was based. Each increment targeted a specific functionality derived from high-level system requirements, and was developed through a series of iterations structured into four phases: Requirements, Analysis & Design, Testing, and Evaluation [18]. In particular, all iterations of the third increment were carried out with the involvement of a music teacher, who provided continuous feedback on user experience and pedagogical relevance, while also taking into account insights from HCI studies to ensure that the interface meets the practical needs of students in real-world conditions. This approach enabled the progressive development of functional modules alongside continuous integration of prototyping, user testing, and feedback collection cycles. This was particularly critical in an HCI-driven context, as it reduced the risk of integration mismatches across input modalities and optimized the user experience in terms of control, accessibility, and responsiveness.

B. VALIDATION

The objective of the validation process was to assess whether the two core contributions of Soundy, introduced in Section I, effectively fostered user engagement in the context of EMP. This goal was pursued through the implementation of a structured evaluation protocol based on comparative testing, autonomous task execution, and the collection of both quantitative and qualitative feedback.

1) PARTICIPANTS

A total of 20 non-disabled high school and university students, aged between 18 and 25, were recruited for the study. The sample comprised 8 female and 12 male participants. All individuals reported at least a basic interest in music creation or audio production. This inclusion criterion was adopted to avoid collecting biased or disengaged feedback from users uninterested in the musical domain. The participant group reflected a range of musical backgrounds and experience levels, consistent with the diversity typically observed in EMP contexts.

2) SETUP AND PROCEDURE

Each participant took part in two separate 30-minute sessions, each corresponding to one of the two system configurations: the standard setup, composed of a Behringer UMC404HD audio interface paired with the Reaper DAW,

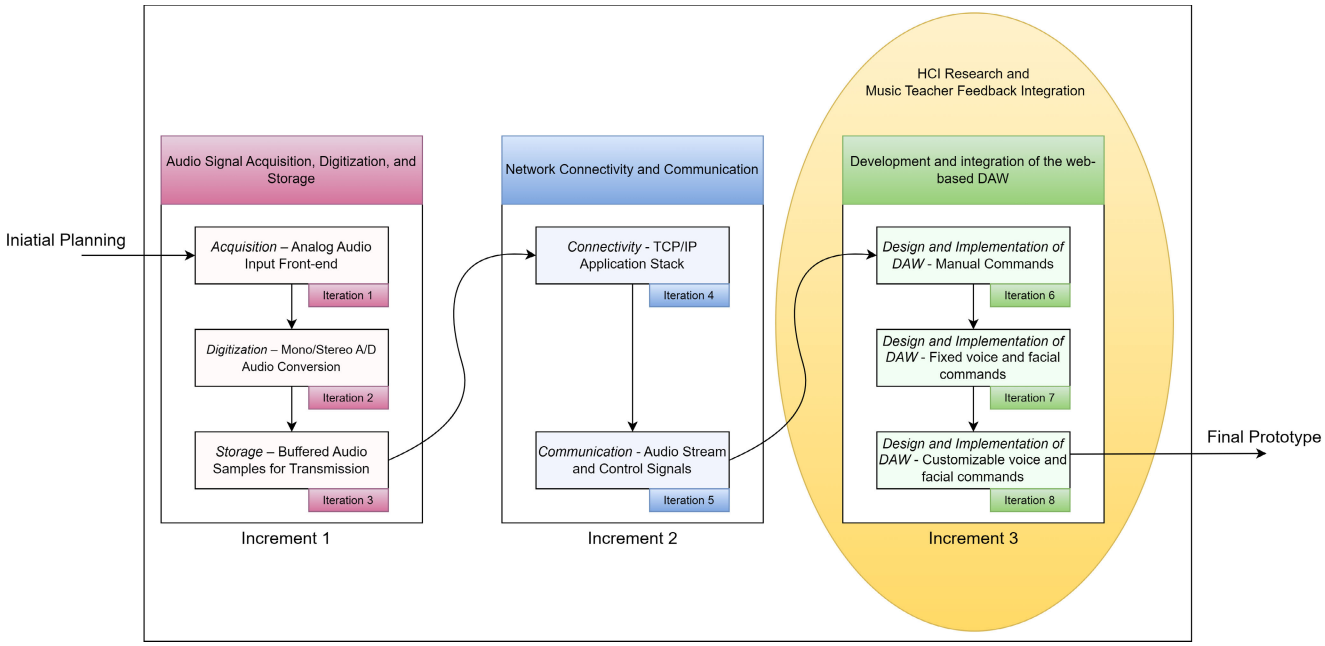

FIGURE 2. The image illustrates the Soundy design methodology, based on an I&I-based development model.

and Soundy, the proposed multimodal system. The standard setup was selected as a representative benchmark, reflecting configurations commonly employed in entry-level EMP and self-guided learning environments. The order of the sessions was counterbalanced across participants to mitigate sequence effects. In both sessions, participants were asked to independently complete the full setup and usage flow, including hardware connection, software initialization, and a *production* task which included audio recording, basic mixing, and exporting a stereo audio file. To emulate a realistic classroom scenario while preserving the integrity of individual evaluations, a music teacher was present during the sessions. The teacher was available to provide clarification regarding the task objectives or specific technical aspects when explicitly requested by the participant. A 10-minute break was introduced between sessions to reduce the impact of fatigue on performance and subjective evaluation.

3) EVALUATION INSTRUMENTS

To evaluate the impact of the system design on user experience, two instruments were employed: the Post-Study System Usability Questionnaire (PSSUQ) [19] and the User Engagement Scale – Short Form (UES-SF) [20]. The PSSUQ, a 16-item instrument, was used to assess system usability across three main factors: System Usefulness (SU), Information Quality (IQ), and Interface Quality (IFQ). An additional item, referred to as the Overall PSSUQ, captured the user's overall satisfaction with the system, offering a global evaluation based on the three factors. The UES-SF, a 12-item instrument, was used to assess user engagement across four experiential factors: Focused Attention (FA), Perceived Usability (PU), Aesthetic Appeal (AE), and Reward and Involvement (RW). This instrument was designed to capture users' experiential and emotional involvement while interacting with the system. Although the UES-SF already includes a usability dimension, the PSSUQ was also employed to provide a more granular and structured assessment of system usability. This approach enabled a deeper understanding of the system's performance and responsiveness within a realistic educational context, as well as further insight into how specific usability aspects may influence user engagement. Both questionnaires utilized a 7-point Likert scale ranging from 1 (strongly agree) to 7 (strongly disagree), with lower scores indicating more favorable responses. This format was selected not only to ensure consistency across instruments, but also based on prior evidence suggesting that 7-point scales offer an optimal balance between reliability, discriminative power, and user preference [21]. Each participant completed both questionnaires immediately after each session, allowing for direct within-subject comparison of the two setups. Additionally, participants provided open-ended feedback after each session to qualitatively contextualize their responses and suggest possible areas of improvement.

IV. SYSTEM DESCRIPTION

Translation of high-level requirements into concrete technical specifications characterized the design phase. These specifications informed both the structural organization of the system and the selection of hardware, firmware, and software components, along with supporting technologies.

Fig. 3 presents the conceptual framework of the Soundy system, highlighting the key functional blocks and their interactions, from analog audio acquisition to multimodal

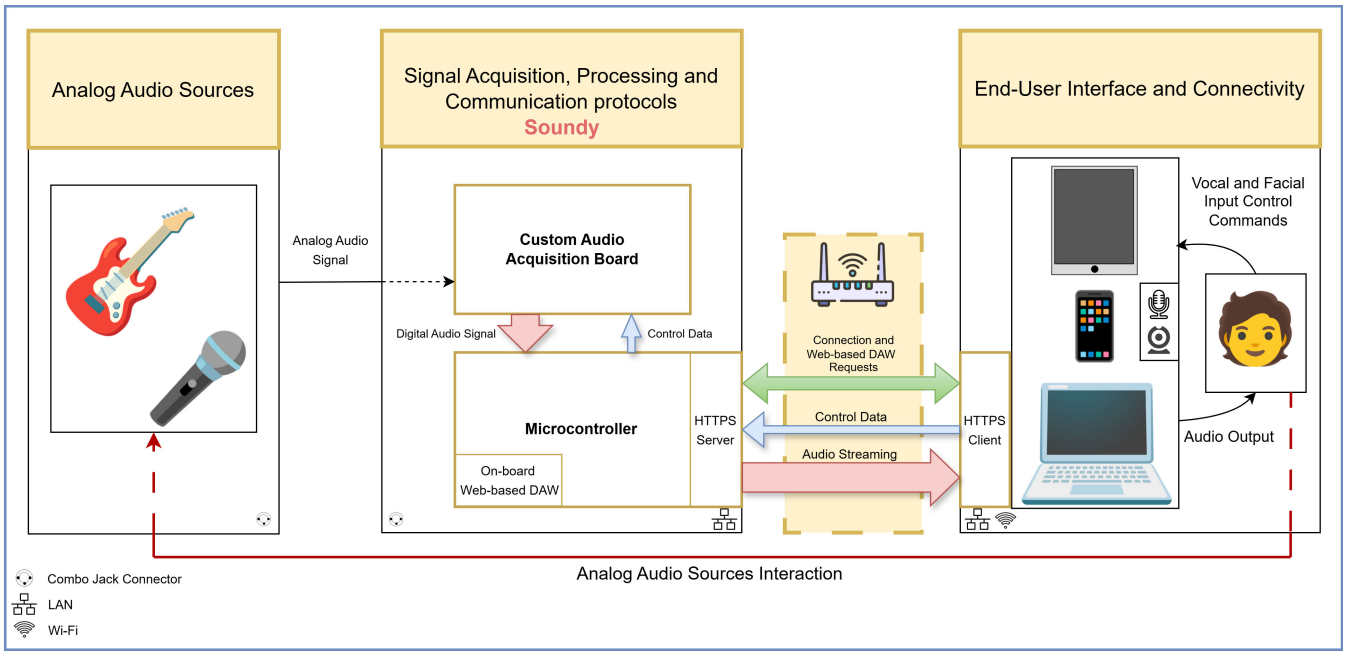

FIGURE 3. The image illustrates the conceptual framework of the Soundy system, structured into three main blocks arranged from left to right. The first block represents analog audio sources connected to the system. The central block corresponds to Soundy itself and comprises a custom audio acquisition board, responsible for capturing, filtering and digitizing the analog audio signal, and a microcontroller that handles processing, control, and communication. The third block focuses on connectivity and user interaction, allowing client devices on the same network to access the web-based DAW. Interaction is enabled through voice and facial commands, leveraging the device's built-in microphone and camera.

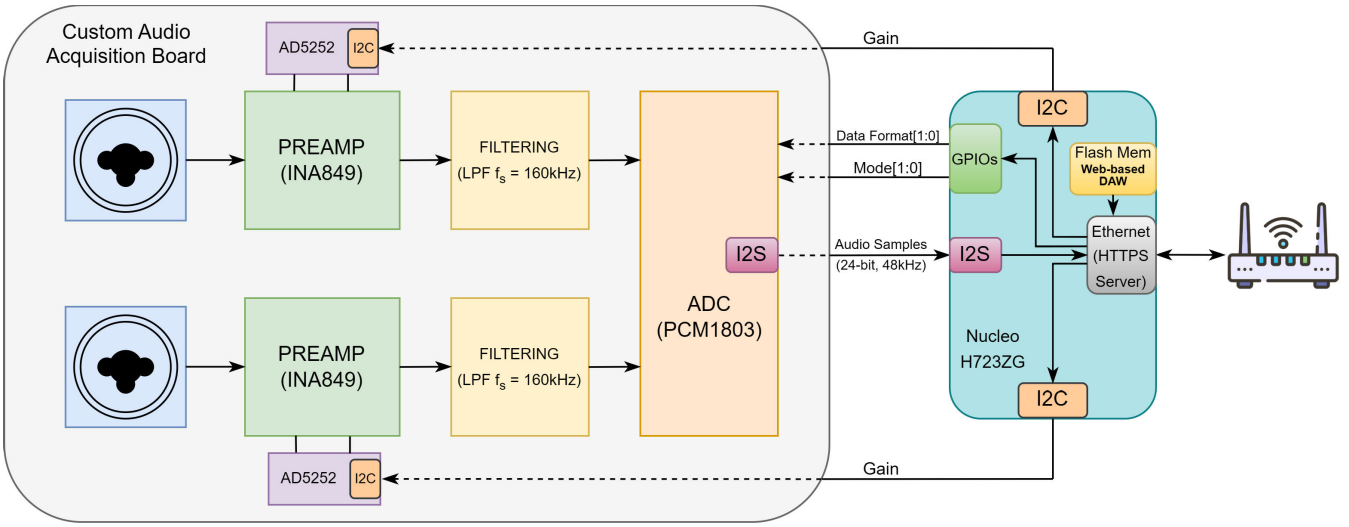

FIGURE 4. Soundy high-level hardware block diagram.

user interface. This framework serves as a reference for the detailed system description provided in the following subsections.

A. HARDWARE

The structure is based on two main hardware components, whose interaction is outlined in the high-level block diagram shown in Fig. 4. The diagram offers an overview of the audio signal path, emphasizing how control parameters are handled throughout the system. This abstraction is intended to provide a clear conceptual framework for later introducing the firmware and software layers, where the actual development of the multimodal interaction logic takes place.

The first hardware block is a custom-designed multi-input audio acquisition board developed by our research group and previously employed in other research [22]. In the context of this work, only two input channels are utilized, and the board is used exclusively as a audio capture

unit. The signal path is organized into three key stages: a preamplification stage, a filtering stage, together forming the analog front-end, and a digitization stage. Audio sources are connected via combo jack connectors, which offer high flexibility by accepting both XLR and 1/4'' TRS inputs. This enables support for microphones as well as musical instruments. The incoming signal first passes through an ultra-low noise instrumentation amplifier (INA849), which is chosen specifically for its low-noise performance. It is paired with a digitally programmable potentiometer (AD5252), selected for its suitability in audio applications; which acts as a user-configurable variable resistor to adapt the gain according to the input source type, accommodating the differences in signal level between microphones and instruments. This stage ensures that the signal amplitude is optimized to fully exploit the dynamic range of the analog-to-digital converter (ADC). Next, the signal undergoes an analog low-pass filter (LPF) with a 160kHz cutoff frequency, serving as an antialiasing filter, as recommended by the PCM1803 ADC datasheet [23]. The ADC carries out a sequence of key internal operations, including oversampling, digital filtering, and decimation. It is selected for its ability to handle stereo signals and transmit them as a single data stream, as well as for its suitability in hi-fi audio applications. The resulting 24-bit, 48kHz digital audio signal is then transmitted via an I2S interface to the second hardware block: an evaluation board based on an STM32H723ZG microcontroller [24]. This board is responsible for managing the incoming digital stream, performing further processing, sending control commands, and initiating HTTPS client-side streaming. Since the board was an off-the-shelf evaluation platform, no custom hardware development is required. The main design effort for this component is focused on firmware and software development.

B. FIRMWARE

The firmware layer is developed as a bare-metal implementation directly on top of the hardware. It is responsible for defining and controlling the system's behavior, effectively translating the static capabilities of the hardware into a dynamic, responsive audio interface. The firmware is organized into three core components.

1) SIGNAL CONDITIONING, DIGITIZATION, AND PACKETIZATION

The first segment of the firmware is responsible for interfacing with the analog front-end and preparing the audio signal for streaming. Gain control is managed by converting user-defined input values, corresponding to desired dB levels, into digital resistance settings for the programmable AD5252 potentiometer via the I2C peripheral. This ensures that the gain is appropriately matched to the characteristics of each input source. The digitized audio stream is read from the PCM1803 ADC via the I2S peripheral, using direct memory access (DMA) to achieve efficient data transfer. Audio samples are stored in a circular buffer and organized into packets of 256 24-bit samples, chosen as a trade-off to reduce latency without compromising buffer transmission time.

2) NETWORK CONNECTIVITY AND COMMUNICATION HANDLING

The second component of the firmware handles networking tasks, enabling the STM32H723ZG to act as a standalone IP-reachable device. Leveraging the Mongoose networking library, a lightweight TCP/IP stack is integrated to support both HTTPS and WebSocket (WS) protocols. Through this stack, the board is able to establish its presence on a LAN and serve content directly via its assigned IP address. Control parameters such as gain, streaming start, data format, and mode are exchanged via HTTPS: sent by the client to update the STM32H723ZG in real time, and served by the STM32H723ZG to initialize the DAW on page load. In parallel, the digital audio stream is continuously transmitted over a WS connection, ensuring efficient, low-latency delivery independent of the control channel.

3) EMBEDDED WEB SERVER VIA PACKED FILE SYSTEM

The final component of the firmware implementation resides in the file system management module. This component, also provided by the Mongoose library, enables the STM32H723ZG microcontroller to host a fully functional Web-based DAW interface, whose source code has been converted into an ASCII-encoded byte array embedded within the firmware image and stored in flash memory. As soon as the user enters the board's IP address in a browser, the microcontroller serves this static content to the client over HTTPS, enabling seamless interaction without the need for any external server or storage medium. This approach significantly enhances the accessibility of the system, making it entirely self-contained, and thus more immediate in its usage.

Together, these three firmware layers form a tightly integrated and cooperative runtime environment (Algorithm 1). Signal acquisition and packetization work in concert with real-time control via network commands, while the embedded Web-based DAW interface provides intuitive access to parameters and audio streaming.

C. SOFTWARE

As previously introduced, the software layer of Soundy features a built-in lightweight web-based DAW (Fig. 5), acting as the primary user interface. Developed using HTML, CSS, and JavaScript, it allows users to interact with audio tracks in a browser-based environment without requiring any additional software installation. It integrates a set of functions corresponding to the typical phases of the *production* process:

- **Tracking:** The user can record audio via one or both input channels of the Soundy interface, assigning the acquired signal to any of the available track slots. During the recording process, input gain can be adjusted in

Algorithm 1 This pseudocode summarizes the core logic of Soundy's firmware, detailing the inclusion of libraries, initialization of hardware and services, and the handling of HTTPS requests for configuration, audio streaming, and web-based DAW access

// Include Libraries

1: Include STM32H723ZG HAL[1] libraries.
2: Include Mongoose networking and file system management libraries with TLS[2]/SSL[3].
3: Include PCM1803 and AD5252 libraries.

// Initialize System

4: Initialize STM32H723ZG (e.g., system clock, peripherals, memory regions).
5: Initialize Mongoose manager and configure TLS/SSL for secure connections.
6: Initialize PCM1803 and AD5252.

// Main Loop

7: Main Loop:
  8: Poll Mongoose event handler for HTTPS events.
  9: On HTTPS request received:
    9.1: If URI[4] matches `/api/get/data`: STM32H723ZG responds with a JSON[5] payload to the client containing current control parameters.
    9.2: Else if URI matches `/api/send/data`: the STM32H723ZG receives commands from the client by parsing the JSON payload to update control parameters (PCM1803 and AD5252).
    9.3: Else if URI matches `/api/stream/data`:
      9.3.1: If *start streaming* requested: upgrade to WS connection and set streaming active.
        9.3.1.1: Capture audio samples from PCM1803.
        9.3.1.2: Buffer and prepare audio data for WS transmission.
        9.3.1.3: Send audio packets.
      9.3.2: Else if WS connection closed: stop streaming and reset streaming flags.
    9.4: Else: serve web-based DAW content.
  10: Error event handler.

[1] HAL — Hardware Abstraction Layer
[2] TLS — Transport Layer Security
[3] SSL — Secure Sockets Layer
[4] URI — Uniform Resource Identifier
[5] JSON — JavaScript Object Notation

real-time using the "Volume 1" and "Volume 2" sliders, which transmit control commands directly to the Soundy device. Furthermore, a digital equalizer is available to modify the frequency content of each track in real time.

- **Mixing:** The DAW includes four independent track slots that can be manipulated post-recording. Users can move and trim audio clips, perform stereo panning, and adjust individual track volume, enabling a flexible mixing process.
- **Mastering:** Upon exporting the final audio, users can activate the "Assisted Mastering" mode. This feature performs global equalization and applies a limiter to optimize the overall dynamic range and spectral balance of the output.

By default, the "Multimodal Interaction" checkbox is disabled, allowing users to interact with the DAW manually, as with any traditional interface. Once enabled, browser permissions for microphone and webcam access are requested. From that moment, the web-based DAW enters a passive listening state, awaiting specific awake commands designed to ensure robustness and protocol adherence. For voice-based mode, the awake/sleep command is triggered by the user saying "start" or "stop." For facial-based mode, it is triggered by keeping both eyes closed for 2 seconds. Both awake/sleep commands are fully customizable by the user. It is possible to activate only one inclusive mode at a time, either voice or facial, and switching to the other requires returning first to manual mode. Once the awake command has been issued, Soundy becomes responsive to inclusive commands relevant to the selected mode. A detailed description of the multimodal interactions is provided below.

1) VOICE MODE

The voice-based mode is a feature designed to allow system control through voice commands. This mode is structured into two main levels:

- **Static level:** At the core of the voice mode lies a static dictionary that maps a set of predefined keywords to specific user interface controls. For example, terms like "bass", "treble", or "volume" are associated

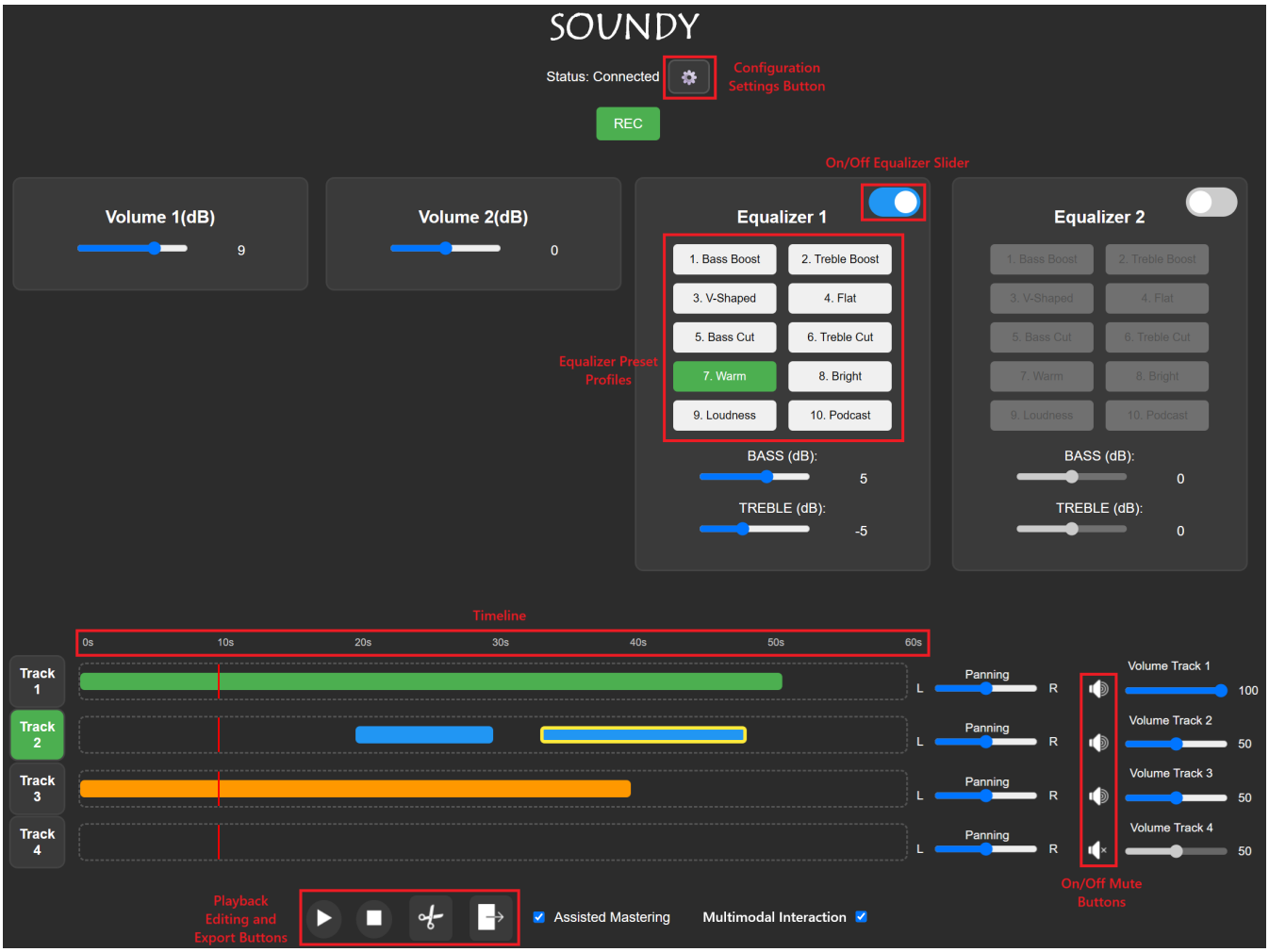

FIGURE 5. The image shows the Soundy web-based DAW. At the top, the system status is shown along with a gear icon used to access configuration settings such as ADC control parameters and the customization for voice and facial commands. Below, two volume sliders control input gain, followed by two equalizer panels (one per channel) offering both preset profiles and manual bass/treble adjustment. The timeline view allows visual arrangement of audio clips across four tracks, each configured with dedicated panning, and volume adjustment. At the bottom, playback, editing, and export controls are available along with the "Assisted Mastering" and "Multimodal Interaction" checkboxes.

with their respective sliders, while commands such as "equalizer on" or "record start" activate toggle-based functionalities. The dictionary also includes common phonetic variants and synonyms to improve robustness in voice recognition, such as "bus", "best", or "base" for "bass", and "trouble" or "travel" for "treble". Words are recognized using the Web Speech API[4] and mapped dynamically via a parser that normalizes recognized phrases. The selection of these predefined voice commands was directly informed by established principles from voice interface and HCI research. As highlighted by [25], commands that are semantically clear, short, and easy to remember help reduce user frustration and cognitive load. In addition, [26] emphasize that in multitasking or hands-busy scenarios, such as playing a musical instrument during a production session, concise and well-structured voice commands are strongly preferred by users due to their low interaction cost and immediacy.

- **Dynamic level:** In addition to static mappings, the system allows users to customize the activation keyword for each interactive element. Through the settings interface (accessible via the settings icon), users can specify alternative words that they prefer to use in place of the default commands. This customization is performed via manual input and supports any word present in the English dictionary, as long as it does not conflict with existing predefined entries. For example, a user may manually assign the word "boo" to replace "bass" within the settings menu. From that moment on, saying "boo five" would adjust the bass level to 5 dB. These mappings are stored locally and persist across sessions. This approach allows the interface to progressively adapt to the user's preferred vocabulary.

[4]https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API

TABLE 1. Comparison between Soundy and general full-fledged DAW.

Category	Feature	Soundy	General full-fledged DAW
Core Workflow	Multitrack audio recording	Supported	Supported
	Playback and transport controls	Supported	Supported
	Basic mixing (e.g., volume, pan)	Supported	Supported
	Export audio file (e.g., WAV, MP3)	Partially supported	Supported
Audio Effects	FX presets (e.g., equalization, reverb)	Partially supported	Supported
	Plugin support (e.g., VST, JSFX)	Not supported	Supported
Advanced Editing	MIDI sequencing	Not supported	Supported
	Clip editing (e.g., split, trim)	Partially supported	Supported
File Management	Project saving/loading	Not supported	Supported
	Import external audio files	Not supported	Supported
Accessibility and Inclusivity	Web-based/installation-free DAW	Supported	Not supported
	Predefined voice/facial commands	Supported	Not supported
	Custom voice/facial commands	Supported	Not supported

The system provides both auditory and speech synthesis feedback to confirm each new command execution, enhancing both usability and transparency of interaction.

2) FACIAL MODE

The face-based mode enables users to control the interface using facial gestures. At the core of the facial mode lies the MediaPipe FaceMesh[5] solution, which processes video input from the user's webcam and outputs a set of 478 three-dimensional facial landmarks per frame. These landmarks include key positions around the eyes, mouth, nose, and jawline, and are updated continuously in real time. The framework operates entirely on the client side, ensuring low latency and preserving user privacy. The raw landmark data is then processed by a custom algorithm that extracts high-level geometric features used to detect deliberate gestures. For example, the system computes the Eye Aspect Ratio (EAR) to detect eye closures (winks), calculates yaw angle deviation to detect horizontal head tilts, and evaluates mouth geometry to identify smile expressions. Just like with voice commands, the facial mode is composed by two levels:

- **Static level:** At this level, facial interaction is organized into four main classes of predefined commands, each serving a distinct role in the user interaction flow. Within each class, specific facial gestures map to actions that the user performs to interact with the system:
  1. *Awake commands* manage the activation and deactivation of facial mode. As anticipated, the user can enter or exit facial mode by default by closing both eyes for two seconds.
  2. *Navigation commands* allow the user to cycle through a predefined sequence of interactive interface elements (e.g., buttons, sliders, toggles). The user tilts their head to the left to move to the previous element, and to the right to advance to the next one.
  3. *Confirmation and adjustment commands* let the user interact with the currently selected element using a wink gesture. For binary controls (e.g., a button or toggle), a wink with either eye activates the command, as the system does not consider direction. For continuous controls (e.g., volume or EQ sliders), the eye used to wink determines the adjustment: a right eye wink increases the value, while a left eye wink decreases it.
  4. *Shortcut commands* are designed to enhance usability. For example, the user smiles to start or stop audio recording, allowing them to stay focused on track-level controls without needing to navigate back to the record button. This reduces effort and streamlines the workflow.

  The selection of facial gestures was informed by previous studies, such as [27], highlighting the importance of intentionality, low muscular effort, reliable recognition, and social acceptability in gesture design. In addition, [28], who explored gesture-based control in immersive musical environments, emphasized the importance of natural and metaphorically meaningful gestures. Their findings support the idea that intuitive mappings between facial gestures and musical actions, such as using a right wink to increase and a left wink to decrease the volume of a slider, can enhance both learnability and user satisfaction.
- **Dynamic level:** Unlike the static case, where gestures are predefined and recognized using fixed rules (e.g., Eye Aspect Ratio), the dynamic approach adopts a different strategy, as the specific facial command cannot be known in advance. Through the settings interface, user can initiate the creation of a new facial command (Fig. 6). The process begins with a facial calibration phase, during which the system ensures that the user's face is properly aligned and stable. This step creates consistent conditions for the accurate acquisition of facial gestures. After calibration, the user selects the class and command to which the new gesture will be assigned, specifying which existing gesture will be replaced. The user then performs the desired facial gesture, such as a distinct blink sequence, a head tilt, or a

[5]https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker

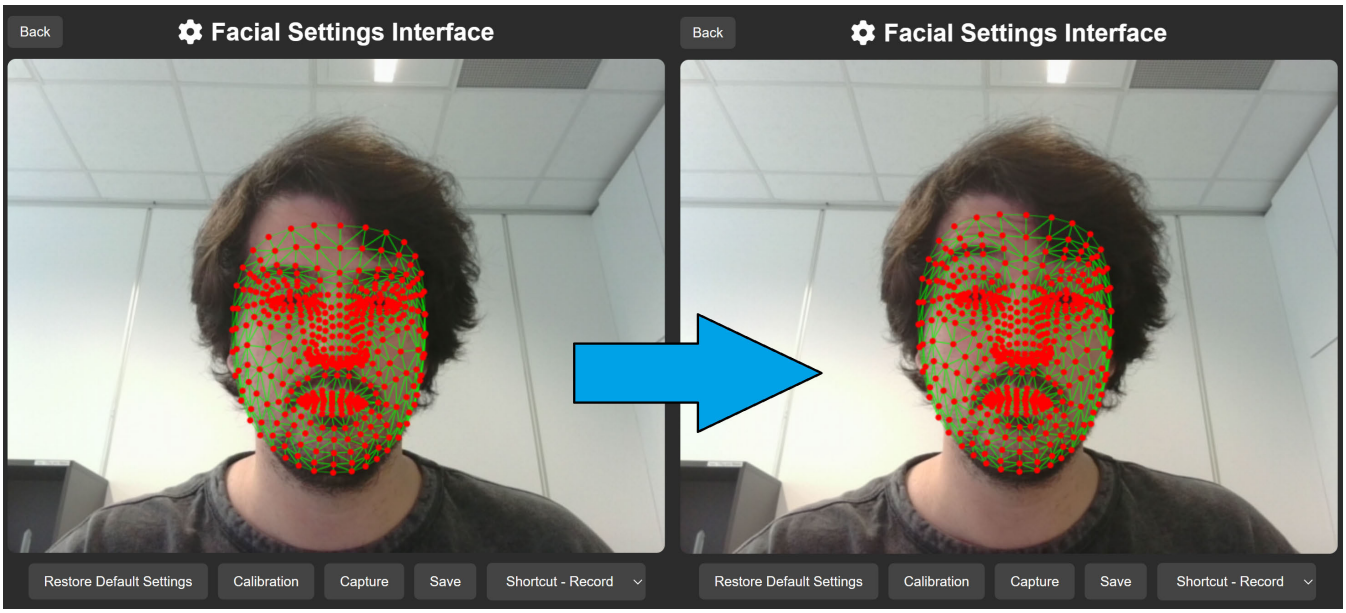

FIGURE 6. The image shows a user during two phases of selecting a custom facial gesture. On the left, the face is shown at rest with the associated landmarks, ready for calibration. On the right, the user is performing a new facial expression, specifically raising both eyebrows, which is selected as the command to start and stop recording under the *Shortcut* class. The procedure also allows restoring the default settings to the predefined ones.

specific expression. The system instantly captures the facial landmark configuration corresponding to the user-defined gesture. Once acquired, the gesture is analyzed to determine which facial regions exhibit the most significant deviation from the user's calibrated neutral expression. Based on this analysis, a subset of relevant landmarks is dynamically selected. Landmarks showing the greatest spatial variation are assumed to encode the core expressive components of the gesture. This approach is supported by findings in facial expression recognition research, where selecting geometric features based on spatial variance has been shown to improve performance [29]. Rather than storing raw video frames or the full set of landmark coordinates, the system saves only the extracted subset of landmark positions, representing the gesture in a compact, discriminative, and privacy-preserving format. This stored subset serves as the reference template for the selected command, effectively completing the registration process.

During real-time interaction, the system extracts from the current input frame the same subset of landmarks identified during registration, and performs a landmark-wise comparison with the previously stored configuration to determine whether the gesture has been successfully reproduced. Since the user who records the gesture is the same as the one interacting with the system, interpersonal variation is eliminated, allowing for a more direct and reliable matching process.

For both the static and dynamic levels, the detection sensitivity of facial commands, defined through configurable gesture thresholds, is adjustable via the settings interface. This reliability is further supported by the robustness of the underlying MediaPipe FaceMesh model, which leverages deep neural networks trained to produce stable and normalized landmark configurations across moderate variations in lighting, facial orientation, and camera placement. Additionally, cooldown timers, which are preset to 0.5 seconds, can be configured to prevent unintentional activations and reduce noise; they effectively act as a debouncing mechanism by enforcing a minimum time interval between successive gesture recognitions, thus avoiding false positives caused by rapid or involuntary repetitions.

It is important to note that Soundy is not intended to replace professional DAWs, but to serve as an educational tool for EMP, supporting accessibility research, inclusivity, and user engagement. For clarity, Table 1 compares its feature set with that of typical full-fledged DAWs, such as Reaper and Ableton Live. This is not meant as a performance benchmark, but rather as a reference to understand which commonly expected features are available, limited, or absent in Soundy's web-based DAW.

V. RESULTS

All participants successfully completed the experimental procedures outlined in Section III. A comprehensive comparative analysis was conducted on the main factors based on the collected data, consisting of quantitative responses from the PSSUQ and UES-SF questionnaires. To enrich the quantitative analysis, qualitative insights derived from open-ended participant feedback were subsequently used to inform the discussion.

A. PSSUQ

The PSSUQ was used to assess the system's usability. Table 2 reports the calculated scores for all PSSUQ factors. For each one, in both the Standard and Soundy setups, mean values (M) and standard deviations (SD) were computed. In addition, internal consistency for the full questionnaire was measured using Cronbach's alpha (C-α), confirming the reliability of the PSSUQ by assessing how consistently participants responded to items across the scale. Finally, to evaluate statistical differences between setups, paired sample t-tests were performed for each factor, reflecting the within-subject design in which participants experienced both interface setups. A two-tailed test with a significance level of $\alpha = 0.05$ was applied, testing the null hypothesis that the mean difference between the two setups was zero. For each factor, the p-value is reported to determine whether the observed differences reached statistical significance. Fig. 7 illustrates a comparative view of the score percentage distributions across the three core dimensions for the two setups. Both configurations received overall positive usability evaluations, with consistently low PSSUQ scores across all factors, indicating high user satisfaction in absolute terms. Building on these overall positive results, the comparative analysis reveals consistently more favorable evaluations for the Standard setup, whose lower average scores reflect an even higher degree of satisfaction. The most pronounced difference emerged in the SU factor, where the Standard setup was clearly perceived as more effective and easier to use. This suggests a stronger alignment with user needs during task execution. In contrast, differences in IQ and IFQ were modest. The relatively small score gaps in these dimensions suggest that the Soundy setup still offered a solid experience in terms of information clarity and interface comfort, even if slightly less effective than the Standard version. The reliability of the responses was confirmed by high C-α values in both conditions, supporting the internal consistency of the collected data. The paired-sample t-test revealed a statistically significant difference only for SU. In contrast, the non-significant p-values for IQ and IFQ factors indicate that the data do not provide enough evidence to reject the null hypothesis of equal means.

B. UES-SF

The UES-SF was used to assess user engagement. Table 3 reports the calculated scores for all UES-SF factors. As with the PSSUQ, M, SD, C-α and p-values were computed. Fig. 8 illustrates a comparative view of the score percentage distributions across the components for the two setups. Both configurations received positive engagement evaluations, with consistently low UES-SF scores across all factors, indicating high user engagement in absolute terms. Given these positive outcomes, the comparative analysis reveals consistently more favorable evaluations for the Soundy setup, whose lower average scores in FA, AE, and RW reflect higher immersion, aesthetic appeal, and perceived reward during the interaction. The only clear exception was found in PU, where the Standard setup received more favorable evaluations, suggesting a higher level of perceived usability. To compute an overall UES-SF score, it was necessary to first reverse-score the PU items for both setups. These items were negatively worded (e.g., "I felt frustrated while using the system"), meaning that a high score indicated strong disagreement, which actually reflects a positive experience. Without this adjustment, these items would have been misaligned with the rest of the scale, where lower scores already indicate more favorable responses. The reliability of the responses was confirmed by high C-α values in both conditions, supporting the internal consistency of the collected data. The paired-sample t-test showed that FA, RW and AE factors all reached statistical significance, though for AE this is less evident since its p-value lies just below the α threshold. By contrast, PU did not achieve significance, so the null hypothesis of equal means cannot be rejected; nevertheless, its p-value is relatively close to the α threshold, suggesting a possible, albeit weak, trend toward higher perceived usability in the Standard setup.

VI. DISCUSSION

In this section, users' open-ended responses are employed to contextualize and explain the quantitative findings presented in Section V. These insights are discussed across two main thematic areas: multimodal interaction and the embedded web-based DAW. A third subsection addresses general hardware, inclusivity, and accessibility considerations, which are based on the authors' technical evaluations and design reflections rather than direct user feedback. Each subsection highlights system strengths and criticalities, linking user perceptions to the core technical contributions. Limitations and directions for future work are integrated within the relevant discussions to provide a comprehensive evaluation of the current system and its development potential.

A. MULTIMODAL INTERACTION

The comparative analysis of the two interaction setups highlights a fundamental trade-off between usability and engagement. While the Standard configuration achieved superior usability scores, as reflected in the PSSUQ results, the Soundy setup proved more effective in fostering user engagement, as captured by the UES-SF and confirmed by qualitative feedback. User responses pointed out that although individual facial and voice commands in the multimodal system were executed with low latency, the overall time required to complete tasks was significantly longer than with the Standard setup. This inefficiency was largely attributed to recognition inaccuracies, particularly for custom commands. Predefined commands, both facial and voice-based, exhibited high recognition accuracy; however, custom facial gestures occasionally showed inconsistency, and custom voice commands lacked an adaptive misunderstanding-handling mechanism (e.g., dictionary as in

TABLE 2. Comparative PSSUQ results for standard and Soundy setups.

Factor	M (Standard)	M (Soundy)	SD (Standard)	SD (Soundy)	p-value	C-α (Standard)	C-α (Soundy)
SU	1.77	2.61	0.51	1.31	$\ll 0.050$	0.884	0.939
IQ	1.90	2.10	0.64	0.85	0.523		
IFQ	1.66	1.86	0.49	0.61	0.262		
Overall PSSUQ	1.79	2.40	0.53	1.06	–		

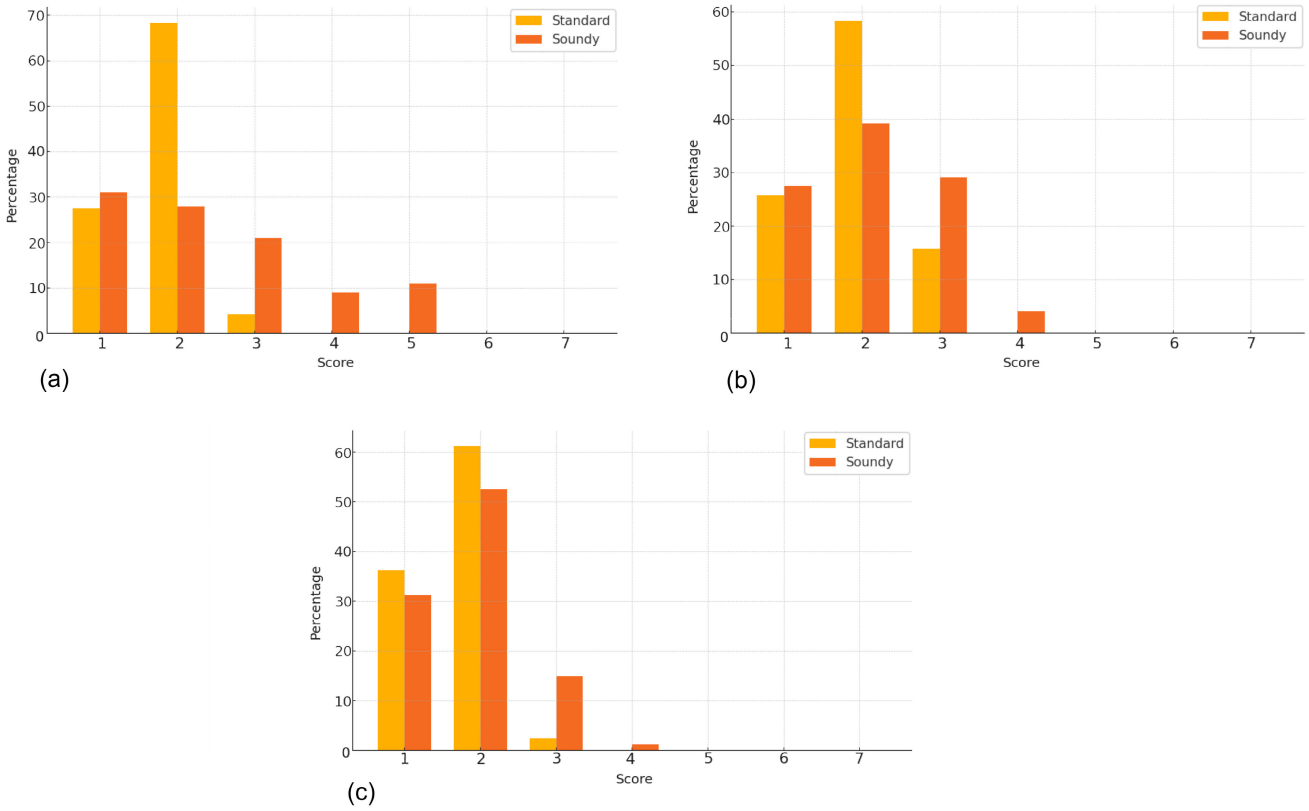

FIGURE 7. The figure shows the percentage distribution of PSSUQ factor scores comparing the two systems (Standard vs. Soundy) with respect to the three main factors: *(a)* SU, *(b)* IQ, and *(c)* IFQ. Lower PSSUQ scores, based on a 7-point Likert scale, indicate better system's usability (i.e., 1 = strongly agree, 7 = strongly disagree).

the static case), due to the fact that the system could not anticipate in advance which word the user would choose. These technical limitations negatively affected the perception of usefulness and task efficiency in Soundy, aligning with the higher SU scores observed in the PSSUQ. Despite these challenges, Soundy clearly stimulated user curiosity and exploration. The ability to freely create and assign custom commands led to a playful and personalized experience that several users described as enjoyable and even entertaining. This experiential richness, although technically less robust, resulted in higher perceived engagement, consistent with the lower FA, AE, and RW scores in the UES-SF, which indicate stronger immersion and reward. Users expressed a willingness to tolerate recognition errors in exchange for the creative freedom offered by the multimodal interface. In contrast, the Standard setup benefited from users' familiarity with traditional manual inputs. Every function was reliably executed, and task flow was generally faster and more predictable. However, this static interaction model failed to stimulate user interest or motivation to explore beyond the necessary steps. Some participants characterized it as "repetitive" or "uninspiring," underscoring a disengagement that was reflected in comparatively higher UES-SF scores in engagement-related dimensions. These findings suggest that, although the Soundy setup successfully achieved high levels of user engagement in EMP contexts, there is still room for technical improvement to optimize its usability and enhance its overall quality.

TABLE 3. Comparative UES-SF results for Standard and Soundy setups (PU reversed values in parentheses used for Overall UES-SF computation).

Factor	M (Standard)	M (Soundy)	SD (Standard)	SD (Soundy)	p-value	C-α (Standard)	C-α (Soundy)
FA	2.58	1.62	0.83	0.52	$\ll 0.050$	0.891	0.881
PU	5.83 (2.17)	5.45 (2.55)	0.72	0.77	0.082		
AE	2.43	1.88	0.83	0.66	0.048		
RW	2.08	1.53	0.89	0.62	0.024		
Overall UES-SF	2.32	1.89	0.82	0.65	–		

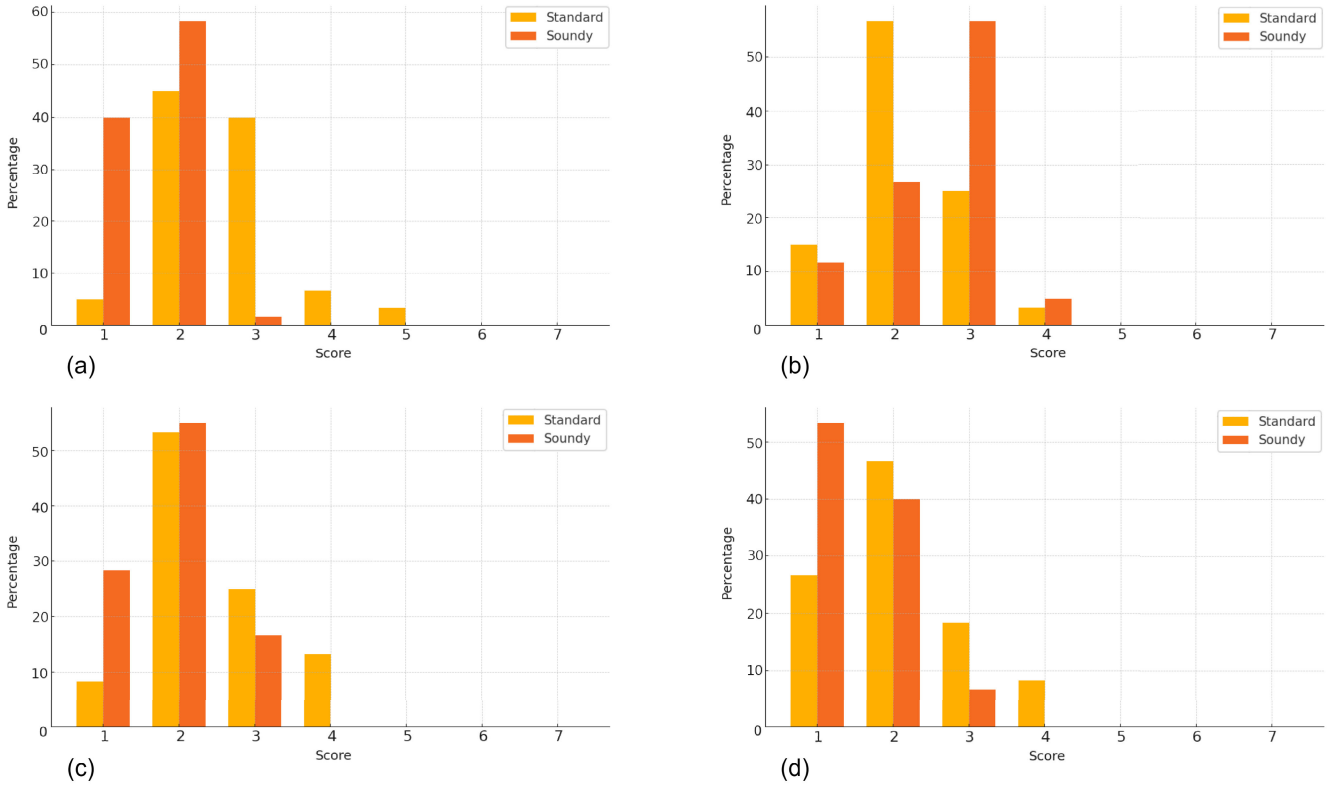

FIGURE 8. The figure shows the percentage distribution of UES-SF factor scores comparing the two systems (Standard vs. Soundy) across the four engagement dimensions: (a) FA, (b) PU, (c) AE, and (d) RW. Note that PU scores are shown in their reversed form, consistent with how they were treated in the overall UES-SF computation. Lower UES-SF values, based on a 7-point Likert scale, indicate stronger user engagement (i.e., 1 = strongly agree, 7 = strongly disagree).

To build upon these results and address the identified limitations, future work should focus on both the technical refinement of the system and its broader evaluative validation. A key direction for future work involves enhancing the robustness of custom command recognition, particularly through the integration of adaptive feedback loops or improved strategies and algorithms focused on command memorization and evaluation. For instance, reliability could be improved by asking users to perform the same facial gesture multiple times during the customization phase. These repeated samples can serve as a lightweight form of training data. By aligning the captured landmark sets and computing average positions (or discarding landmarks with high spatial variance), the system can derive a more robust geometric representation of the gesture. Although gesture evaluation during real-time interaction would still follow a single-shot comparison, the reference configuration used for matching would no longer correspond to a single frame, but rather to a consolidated representation derived from multiple frames. This approach, while not involving traditional machine learning, helps mitigate variability caused by slight changes in facial orientation or lighting conditions, resulting in a more stable and reliable matching process, at the cost of requiring additional time and effort from the user during the customization phase. In parallel, future developments will also explore the introduction of explicit feedback for

unsuccessful recognitions, complementing the existing cues already shown when commands are correctly executed. Visual and auditory indicators will inform users when a command is not recognized, helping them adjust or retry their input. Additionally, for custom facial commands, an auto-tuning mechanism for the gesture sensitivity threshold will be implemented to adapt over time based on recognition performance. Meanwhile, for custom voice commands, the implementation of automated word similarity matching, based on the user-defined input, will be considered to build an adaptive dictionary that improves recognition robustness by accounting for minor pronunciation differences and input variability. Together, these enhancements aim to improve recognition reliability and strengthen user trust without requiring manual recalibration. Although several iterative tests were conducted during the development phase, a dedicated evaluation of the effectiveness of custom commands using a larger and more diverse user base is warranted. Additionally, while the current study benefitted from the guidance of a music teacher during development and testing, future work will include the administration of targeted questionnaires, such as the Course Experience Questionnaire (CEQ), specifically aimed at music educators. This will provide further insight into pedagogical expectations and practical requirements for integrating multimodal systems in EMP.

B. EMBEDDED WEB-BASED DAW

The second core contribution of this research involved evaluating the integration of a web-based DAW directly within the hardware platform. As reflected in user feedback, this solution was well-received in terms of accessibility and responsiveness, especially in relation to the ease of setup and immediacy of access: participants reported no issues with connectivity, audio streaming, or command configuration during their sessions. Entering the static IP address, labeled on the device, directly in the browser was described as intuitive and problem-free. These observations align with the low PSSUQ scores recorded for the ''Information Quality'' (IQ) and ''Interface Quality'' (IFQ) dimensions, confirming that users found the interface clear and the interaction efficient.

However, some users expressed concerns regarding the limited feature set: although it enabled basic tasks related to the *production* phase, several participants noted that it lacked the depth and flexibility typically expected in a creative music environment. By contrast, the Standard setup, despite being described as ''confusing'' and featuring a steeper learning curve, was praised for its greater customizability and expressive potential. These contrasting perceptions emphasize different user needs: while simplicity and speed were appreciated, more experienced users expressed a desire for enhanced creative control. Future work should therefore explore how to expand the DAW's capabilities while maintaining its accessible and compact design.

Potential improvements include a higher degree of equalizer customization (e.g., additional frequency bands), as well as the addition of new effects such as compression, delay, and reverb. These enhancements would better support more advanced production workflows without compromising usability, while also improving opportunities for creative expression. At the same time, the Soundy's limited internal flash memory presents a constraint for large-scale functional expansions. To overcome this, future versions may require external flash storage to accommodate more advanced features while preserving performance.

Cross-device accessibility also remains an open direction: although the current system was successfully tested on desktop browsers via LAN, its behavior across different devices, such as smartphones and tablets, has not been systematically assessed. Broader compatibility testing will be essential to ensure consistency across platforms.

C. GENERAL CONSIDERATIONS

Although the core contributions of this work primarily focused on the development of the firmware and the embedded web-based DAW, the decision to start from a custom hardware foundation proved essential to explore the system as a fully dedicated and self-contained solution from the earliest stages. While much attention has been given to the evaluation of interaction modalities and the embedded web-based DAW, it is important to reflect on the hardware platform itself and the broader design implications emerging from this study. From a technical standpoint, the hardware setup employed during testing demonstrated reliable performance and system stability, both in functional evaluations and throughout the experimental sessions involving questionnaire-based feedback. The microcontroller-based platform was able to sustain real-time audio streaming, command processing, and DAW access without functional disruptions. However, as expected for a prototype, the current physical implementation presents limitations in terms of size, power consumption, and portability. The device is relatively bulky and requires a constant power supply, which limits its suitability for mobile scenarios or portable use. These constraints are not due to architectural inefficiencies, but rather to the developmental nature of the hardware.

To address these limitations, future iterations of the system will explore the design of a fully custom PCB that integrates all necessary components on a single board. This approach would significantly reduce the physical footprint, improve energy efficiency, and enable the integration of a battery-powered solution. Additionally, incorporating a Wi-Fi communication module would allow the system to operate without an Ethernet cable, making it possible to connect directly to musical instruments or microphones. This would remove the need for both power and data cables, thereby enabling a fully wireless and self-contained device that is more easily deployable in a variety of educational and performance contexts.

In parallel, the theme of accessibility and inclusivity emerged as a relevant consideration for the long-term vision of the system. Although the inclusion of participants with disabilities was outside the scope of this study, the current modular and web-based architecture provides a solid foundation for future inclusive adaptations. Investigating how users with diverse physical or cognitive needs interact with the system could inform the design of universally accessible tools, particularly in the EMP context. Future research should therefore aim to evaluate the system with a more diverse participant base and explore accessibility-oriented extensions that support equitable interaction for all users. Finally, although neither Soundy's firmware nor software is currently open-source, making both publicly accessible is actively being considered for the future. The present focus is on improving system robustness and usability, key aspects identified through this research that require further refinement before the platform can be released to the broader research and education communities.

VII. CONCLUSION

This study introduced and evaluated Soundy, an embedded audio interface designed to enable multimodal interaction and provide direct access to a web-based DAW for EMP contexts. The overall system was evaluated using a comparative approach against a standard setup, based on feedback from users who completed two standardized and widely validated questionnaires measuring usability (PSSUQ) and engagement (UES-SF). Findings revealed a distinct trade-off between the two configurations. While the standard setup proved more effective in terms of usability and task efficiency, Soundy promoted higher levels of engagement, creativity, and user exploration. Despite the limitations identified, upon which future work will focus with the ultimate goal of optimizing the student experience, the study demonstrated that a multimodal system can effectively enhance engagement during a learning activity within an EMP context.

DATA AVAILABILITY

To maintain conciseness, the raw responses to the PSSUQ and UES-SF questionnaires have not been included in this paper. All reported results are derived from these raw data. Anonymized questionnaire responses are available upon request. Inquiries can be directed to the corresponding author.

ACKNOWLEDGMENT

The authors acknowledge Riccardo Peloso for the development of the custom audio acquisition board used in this work. This manuscript reflects only the views and opinions and the Ministry, the European Union or the European Innovation Council cannot be considered responsible for them.

REFERENCES

[1] B. Owsinski, *The Music Producer's Handbook*, 2nd ed., Madison, WI, USA: Hal Leonard Corporation, 2016.

[2] M. Clauhs, B. Franco, and R. Cremata, "Mixing it up: Sound recording and music production in school music programs," *Music Educators J.*, vol. 106, no. 1, pp. 55–63, Sep. 2019.

[3] A. P. Bell, *The Process of Production | The Production of Process: The Studio As Instrument and Popular Music Pedagogy*, R. Wright, Ed., Scarborough, ON, Canada: Canadian Music Educators' Association, 2017.

[4] M. Jiang, "The role of experiential learning in transforming music appreciation education and pedagogical practices," *Pacific Int. J.*, vol. 7, no. 5, pp. 157–162, Oct. 2024.

[5] Y. Wu, N. Bryan-Kinns, and J. Zhi, "Exploring visual stimuli as a support for novices' creative engagement with digital musical interfaces," *J. Multimodal User Interface*, vol. 16, no. 3, pp. 343–356, Sep. 2022.

[6] G. Palaigeorgiou and C. Pouloulis, "Orchestrating tangible music interfaces for in-classroom music learning through a fairy tale: The case of ImproviSchool," *Educ. Inf. Technol.*, vol. 23, no. 1, pp. 373–392, Jan. 2018.

[7] T. B. McHugh, A. Saha, D. Bar-El, M. Worsley, and A. M. Piper, "Towards inclusive streaming: Building multimodal music experiences for the deaf and hard of hearing," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–6.

[8] M. Bremmer, C. Hermans, and V. Lamers, "The charmed dyad: Multimodal music lessons for pupils with severe or multiple disabilities," *Res. Stud. Music Educ.*, vol. 43, no. 2, pp. 259–272, Jul. 2021.

[9] R. V. Moora and G. Prabhakar, "Tactile melodies: A desk mounted haptics for perceiving musical experiences," 2024, *arXiv:2408.06449*.

[10] L. Turchet, D. Baker, and T. Stockman, "Musical haptic wearables for synchronisation of visually-impaired performers: A co-design approach," in *Proc. ACM Int. Conf. Interact. Media Experiences*, Jun. 2021, pp. 20–27.

[11] J. Yu, T. Zhang, S. Wu, X. Wu, T. Wu, Y. Chen, and K. Zhang, "ArchiTone: A LEGO-inspired gamified system for visualized music education," 2024, *arXiv:2410.15273*.

[12] J. Mao, Y. Wang, X. Wang, L. Yang, and Y. Ding, "The application of speech recognition in education and teaching," *EAI Endorsed Trans. e-Learn.*, vol. 9, no. 4, 2023.

[13] F. A. Rana, Y. L. Tsang, and T. W. Yip, "Music corner—A feasibility study for creating a gesture-based rhythm game for music education inspired by solfege hand signs," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, May 2024, pp. 1–4.

[14] F. Sun, "Analysis of virtual reality-based music education experience and its impact on learning outcomes," *Scalable Comput., Pract. Exper.*, vol. 25, no. 6, pp. 4755–4762, Oct. 2024.

[15] M. J. Cook, "Augmented reality: Examining its value in a music technology classroom. Practice and potential," *Waikato J. Educ.*, vol. 24, no. 2, pp. 23–38, Nov. 2019.

[16] P.-C. Hu, P.-H. Chen, and P.-C. Kuo, "Educational model based on hands-on brain-computer interface: Implementation of music composition using EEG," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2018, pp. 982–985.

[17] T. Colafiglio, C. Ardito, P. Sorino, D. Lofù, F. Festa, T. Di Noia, and E. Di Sciascio, "NeuralPMG: A neural polyphonic music generation system based on machine learning algorithms," *Cognit. Comput.*, vol. 16, no. 5, pp. 2779–2802, Sep. 2024.

[18] C. Larman and V. R. Basili, "Iterative and incremental development: A brief history," *IEEE Comput.*, vol. 36, no. 6, pp. 47–56, Jun. 2003.

[19] J. R. Lewis, "Psychometric evaluation of the post-study system usability questionnaire based on data from five years of use," *Int. J. Hum.-Comput. Interact.*, vol. 14, nos. 3–4, pp. 463–488, 2002.

[20] H. L. O'Brien, P. Cairns, and M. Hall, "A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form," *Int. J. Hum.-Comput. Stud.*, vol. 112, pp. 28–39, Apr. 2018.

[21] C. C. Preston and A. M. Colman, "Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences," *Acta Psycholog.*, vol. 104, no. 1, pp. 1–15, Mar. 2000.

[22] D. Bert, N. Domini, R. Peloso, L. Severi, M. Sacchetto, A. Bianco, and C. Rottondi, "FPGA-based low-latency audio coprocessor for networked music performance," in *Proc. 4th Int. Symp. Internet Sounds*, Oct. 2023, pp. 1–8.

[23] *PCM1803A Stereo Audio Analog-to-Digital Converter Data Sheet*, Texas Instruments, Dallas, TX, USA, Oct. 2015. [Online]. Available: https://www.ti.com/lit/ds/symlink/pcm1803a.pdf

[24] *NUCLEO-H723ZG Development Board*, User Manual UM2592, Rev. 4, STMicroelectronics, Geneva, Switzerland, Oct. 2021. [Online]. Available: https://www.st.com/en/evaluation-tools/nucleo-h723zg.html
[25] C. Myers, A. Furqan, J. Nebolsky, K. Caro, and J. Zhu, "Patterns for how users overcome obstacles in voice user interfaces," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2018, pp. 1–12.
[26] N. A. N. Ch, D. Tosca, T. Crump, A. Ansah, A. Kun, and O. Shaer, "Gesture and voice commands to interact with AR windshield display in automated vehicle: A remote elicitation study," in *Proc. 14th Int. Conf. Automot. User Interface Interact. Veh. Appl.*, Sep. 2022, pp. 171–182.
[27] K. Masai, K. Kunze, D. Sakamoto, Y. Sugiura, and M. Sugimoto, "Face commands–user-defined facial gestures for smart glasses," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Nov. 2020, pp. 456–467.
[28] H. Y. Leng, N. M. Norowi, and A. H. Jantan, "A user-defined gesture set for music interaction in immersive virtual environment," in *Proc. 3rd Int. Conf. Hum.-Comput. Interact. User Exper. Indonesia*, Apr. 2017, pp. 44–51.
[29] V. Perez-Gomez, H. V. Rios-Figueroa, E. J. Rechy-Ramirez, E. Mezura-Montes, and A. Marin-Hernandez, "Feature selection on 2D and 3D geometric features to improve facial expression recognition," *Sensors*, vol. 20, no. 17, p. 4847, Aug. 2020.

PIETRO BUCCELLATO (Graduate Student Member, IEEE) received the master's degree in electronic engineering (embedded systems) from Politecnico di Torino, Italy, in October 2023, where he is currently pursuing the Ph.D. degree in electrical, electronics, and communications engineering with the Department of Electronics and Telecommunications (DET). His research interests include the design and development of inclusive and accessible audio solutions.

ANDREA BIANCO (Senior Member, IEEE) received the Ph.D. degree in telecommunications from Politecnico di Torino, Italy, in 1993. He is currently a Full Professor and the Vice Rector of Internal Affairs of Politecnico di Torino. He has co-authored more than 230 papers published in international journals and presented at leading international conferences in the area of telecommunication networks. His main research interests include the design of all-optical networks, switch architectures for high-speed networks and data centers, SDN, and networked music performance.

CRISTINA ROTTONDI (Senior Member, IEEE) received the Ph.D. degree in information engineering from Politecnico di Milano, Italy, in 2014. From 2015 to 2018, she had a research appointment with the Dalle Molle Institute for Artificial Intelligence (IDSIA), Lugano, Switzerland. She is currently an Associate Professor with the Department of Electronics and Telecommunications, Politecnico di Torino. She is the co-author of more than 100 scientific publications in international journals and conferences. Her main research interests include optical networks planning and networked music performance.