Toward Multimodal Audio Interfaces in Educational Music Production: A Microcontroller-Based Embedded System with Voice and Facial Control

Pietro Buccellato, Cristina Rottondi
Dep. of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy
pietro.buccellato@polito.it, cristina.rottondi@polito.it

***Abstract*—Multimodal interaction represents a promising direction to enhance accessibility and inclusivity in Educational Music Production (EMP) environments. This work reports on the ongoing design, development, and feasibility evaluation of a microcontroller-based embedded system designed to extend standard audio interfaces with multimodal interaction capabilities. The proposed solution is architecturally composed of two main elements: an external stereo Analog-to-Digital Converter (ADC), specifically the PCM1803AEVM, which digitizes analog audio signals originating from standard audio interfaces; and a NUCLEO-H723ZG evaluation board, based on an STM32 microcontroller, which receives the digital audio stream, manages buffering, and enables data transmission over Ethernet. Once connected to a Local Area Network (LAN), the system operates as a self-contained server, hosting an onboard, browser-accessible Web Audio Interface (WAI) which allows the user to perform core production tasks through voice or facial commands, besides the standard manual ones. A functional evaluation with six users was conducted to assess the accuracy of the multimodal commands across diverse vocal and facial profiles. Results indicate promising recognition accuracy, supporting the case for further validation in representative EMP scenarios.**

***Index Terms*—Educational Music Production, Multimodal, Inclusivity, Accessibility, Microcontroller-Based**

I. Introduction

Educational Music Production (EMP) has emerged as a powerful approach to engaging students with music technology in a hands-on, creative, and immersive way. Several studies have highlighted its potential to bridge the gap between formal classroom instruction and students' everyday musical practices [1], [2]. By supporting alternative interaction methods that go beyond standard manual control, EMP environments enhance user engagement and deepen conceptual understanding in music education settings. [3], [4]. Furthermore, research in Human-Computer Interaction (HCI) suggests that multimodal systems can significantly enrich musical experiences for diverse learners, including those with sensory impairments [5], [6], pointing to the inclusive and adaptable potential of EMP in contemporary educational contexts.

While standard audio interfaces (e.g., Behringer UMC404HD) and Digital Audio Workstations (DAWs) (e.g., Reaper) offer extensive control over the music production process, they often rely heavily on manual input and complex visual interfaces. This can present significant barriers for beginners and users with cognitive or physical impairments, even though operating systems and DAWs often include accessibility features, such as OSARA for Reaper, which improves screen reader compatibility and enables keyboard-based control. While these tools enhance accessibility, they still fall short of delivering a fully integrated and inclusive approach to music production. Moreover, most research on multimodal interaction in EMP (see Section II) tends to focus on isolated aspects of the learning process or investigates single input modalities. As a result, there remains a lack of unified, multimodal systems that support the EMP in an inclusive and flexible way, bringing together diverse technologies to accommodate a wider range of users.

To bridge this gap, this paper presents a prototype of a microcontroller-based embedded system, designed to extend standard audio interfaces with multimodal control, targeting accessible and inclusive use in EMP environments. The system consists of two main components: an external Analog-to-Digital Converter (ADC), which digitizes the analog output signal from standard audio interface and transmits the resulting audio samples via I2S protocol, and a microcontroller that receives the digitized samples and manages audio buffering, packetization, network transmission, as well as the hosting of the Web Audio Interface (WAI). Once connected to the Local Area Network (LAN), the prototype can be accessed by entering its IP address in a web browser. Users can then interact with the loaded WAI either manually or through predefined voice and facial commands. The predefined multimodal commands are used to control interactive elements within the WAI, such as buttons and sliders, which, when triggered, execute essential production tasks including starting and stopping recordings, adjusting input gain, applying equalization, and exporting the edited audio file.

To assess system reliability, a functional validation was conducted with six users, balanced by sex to ensure diversity in facial geometry and vocal characteristics. Each participant was asked to complete a sequence of basic tasks using WAI. The goal was to evaluate the accuracy of the predefined multimodal commands under natural variability. Preliminary feedback indicates good overall performance across users.

The remainder of the paper is organized as follows: Section II presents related work on multimodal interaction in EMP. Section III details the system architecture. Section IV reports the user test results. Section V concludes the paper.

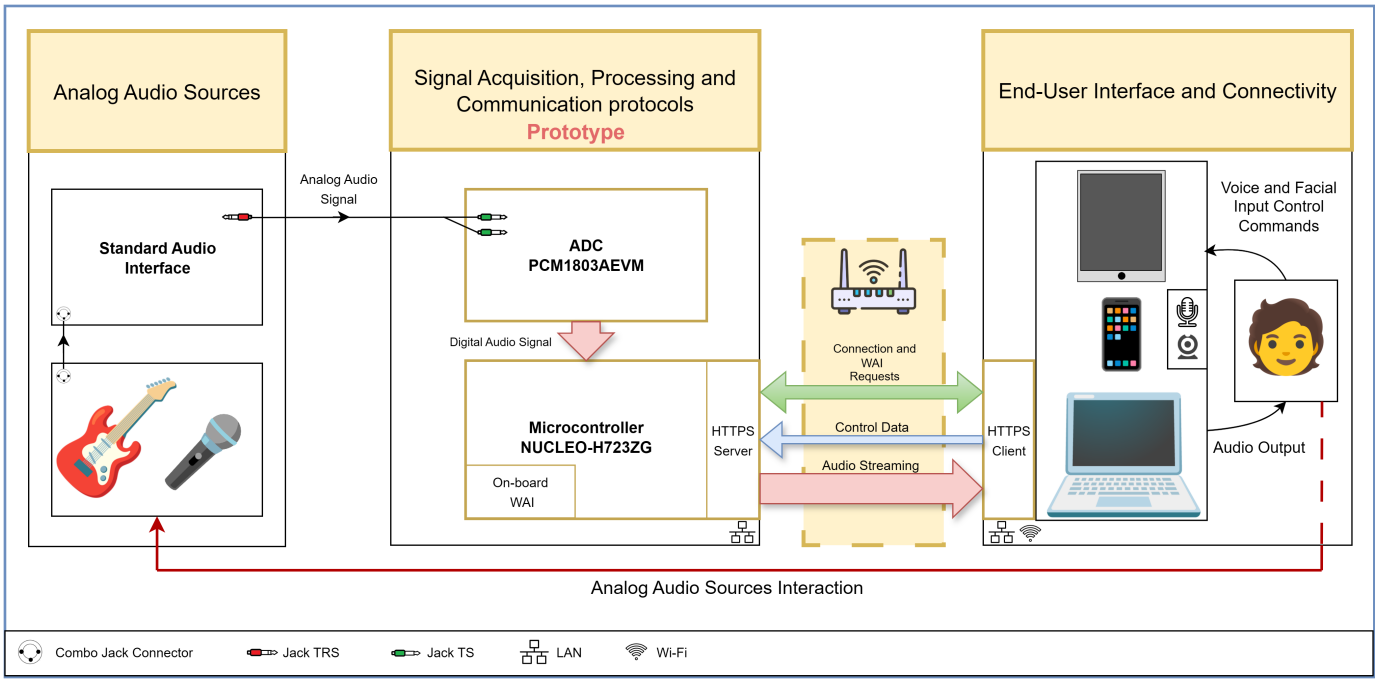

Fig. 1. Conceptual framework of the prototype, divided into three main blocks from left to right. The first block represents the analog audio source chain, including musical instruments and microphones connected to a standard audio interface. The central block corresponds to the prototype and includes an external ADC (PCM1803AEVM) that digitizes the analog signal, and a NUCLEO-H723ZG microcontroller-based evaluation board responsible for buffering, control, and network-based audio streaming via an on-board WAI. The third block illustrates the end-user interface and connectivity layer, where client devices, such as PCs, tablets and smartphones, access the WAI via HTTPS and interact using predefined voice and facial commands through built-in microphones and cameras.

II. Related Work

A wide range of interaction modalities have been explored in the context of inclusive and accessible EMP. These modes reflect different design philosophies, sensor technologies, and levels of immersion, each offering unique strengths for specific user needs.

Among these, *tangible* interaction systems allow users to engage with sound through the physical manipulation of dedicated objects or surfaces that are tightly coupled with musical functions. Unlike standard manual controls, these systems often rely on spatial arrangements and embodied gestures to map real-world artifacts, such as tokens, blocks, or physical grids, to abstract musical parameters like harmony, rhythm, or timbre [4].

Similarly, *haptic* interaction employs vibrotactile feedback systems that deliver tactile sensations in response to audio events or user actions, and can also be used to convey musical content, such as rhythm, dynamics, or harmonic structure, through the body, enhancing multisensory engagement with sound [7].

Visual interaction involves the use of graphical representations to control or interpret musical structures. Systems based on visual interaction encode elements such as pitch, harmony, rhythm, or dynamics into visual features like geometric shapes, color gradients, spatial positioning, or animated transitions, allowing musical processes to be externalized in a visual domain [8].

Voice-based interaction enables control over musical systems through spoken language, typically relying on automatic speech recognition to map vocal commands to specific software functions [9].

Motion-based interaction, involving gestures such as hand movement, head orientation, eye motion, or full-body displacement, enables users to influence musical parameters in expressive and dynamic ways through inertial measurement units (IMU) or video tracking systems [10].

Virtual and *Augmented Reality* (VR and AR) technologies have also been investigated for their potential to transform music interaction by introducing spatial and immersive control paradigms. VR enables users to operate within fully simulated environments, offering three-dimensional representations of instruments and controls that can support novel creative workflows [11]. In contrast, AR superimposes interactive musical elements onto the physical environment, allowing users to engage with virtual controls while maintaining situational awareness [12]. Although less immersive than VR, AR provides greater contextual integration and is potentially more adaptable to classroom and collaborative settings.

Finally, *brain–computer* interaction represents an additional modality in which neurophysiological signals, typically electroencephalography (EEG), can also be employed to interact with musical systems [13], [14].

III. System Description

The overall architecture of the prototype is based on a conceptual framework (see Fig. 1) that defines the main

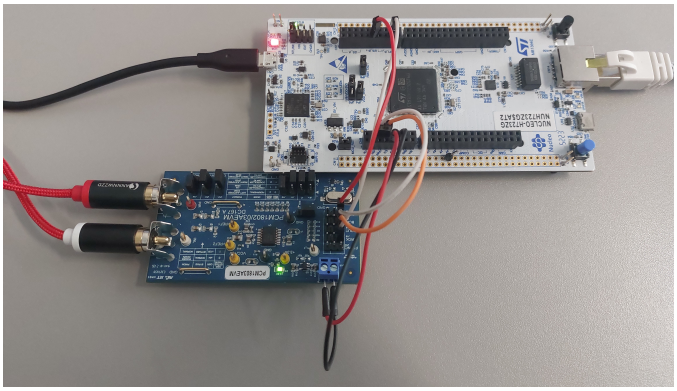

Fig. 2. The image shows the complete embedded hardware configuration used during the system prototyping phase: the blue board corresponds to the PCM1803AEVM, and the white board to the NUCLEO-H723ZG.

functional components and their interactions, from analog audio acquisition to the multimodal interface. To realize this architecture, the prototype is organized into three integrated layers: hardware, firmware and software.

A. Hardware and Firmware Layers

The prototype's core functionalities are realized through a tightly coupled hardware–firmware system. This integrated architecture consists of two evaluation modules representing the hardware (see Fig. 2) and a bare-metal firmware stack that enables audio acquisition, buffering, packetization, and WAI hosting.

The first evaluation board is the PCM1803AEVM, a stereo ADC used to digitize analog signals at a sampling rate of 48 kHz with 24-bit resolution, ensuring high-fidelity audio quality for each channel. The resulting digital stream is then transmitted via the I2S protocol to the microcontroller. The second evaluation board is the NUCLEO-H723ZG, featuring an STM32H723ZG microcontroller. It receives the I2S stream using Direct Memory Access (DMA) and stores the audio data in a circular buffer. The firmware handles packetization into 256-sample blocks, preparing the stream for network transmission. In parallel, the firmware transforms the board into a self-contained LAN server, leveraging the Mongoose networking stack to support both HTTPS and WebSocket protocols. This enables client devices to access the WAI hosted directly from the STM32's internal flash memory. The WAI source code is embedded in the firmware as an ASCII-encoded byte array, eliminating the need for external storage and a file system.

B. Software Layer

As anticipated, the software layer consists of a WAI, which serves as the main point of user interaction. Entirely implemented using standard web technologies, HTML, CSS, and JavaScript, it enables audio manipulation directly within the browser, eliminating the need for external software installation.

The WAI supports basic tracking functionality by allowing the user to record audio coming from the standard audio interface output. During recording, input gain adjustment is possible through the use of the "Volume" slider. Additionally, the interface includes a two-band equalizer, with bass and treble sliders that can be manually adjusted to shape the frequency content of the audio signal. Once the recording is complete, the track can be immediately downloaded in *wav* format.

By default, the WAI operates in manual mode through traditional graphical elements. However, when predefined activation cues are detected, it switches to multimodal mode, allowing for hands-free interaction via voice or facial input. Specifically, voice mode can be activated by saying "listen" and deactivated with the command "stop". Facial mode, on the other hand, is toggled on or off by keeping both eyes closed for at least two seconds.

Voice mode enables control of interface elements using a set of predefined words. These words are mapped to interactive components such as sliders and toggle buttons. For instance, saying "record start" or "equalizer on" activates the corresponding controls. To ensure robust performance, the system accounts for common phonetic variations (e.g., "base" or "bus" for "bass"). Voice input is processed using the Web Speech API[1] and a parser dynamically maps recognized words to their associated actions. The voice command design follows principles from HCI literature, favoring brevity, clarity, and ease of recall [15], [16].

Facial mode allows users to operate the interface through facial gestures. This functionality is built upon the MediaPipe FaceMesh[2] framework, which analyzes real-time video input from the user's webcam and extracts 478 three-dimensional facial landmarks per frame. The system processes this data entirely on the client side, preserving privacy and ensuring low latency. A custom algorithm extracts high-level geometric features from the landmarks to identify gestures. For instance, the system derives the Eye Aspect Ratio (EAR) to recognize intentional eye closures (winks), analyzes variations in jaw angle to detect head tilts, and assesses mouth geometry to recognize smiling gestures.

Unlike voice commands, facial gestures used to control the WAI are not mapped directly to specific interface functions. This is due to the inherent limitations of facial expressiveness compared to the extensive vocabulary of spoken language. As a result, the predefined facial commands are grouped into three functional categories, each associated with a distinct role in the interaction flow:

1) *Navigation* commands allow the user to sequentially browse through interface elements (e.g., buttons, sliders, toggles). A head tilt to the left selects the previous element, while tilting to the right advances to the next.
2) *Confirmation and adjustment* commands enable interaction with the currently focused element via wink gestures. For binary controls (such as buttons or toggles), a wink with either eye confirms the action. For continuous parameters (like volume or equalizer), the eye used to

[1]https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API
[2]https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker

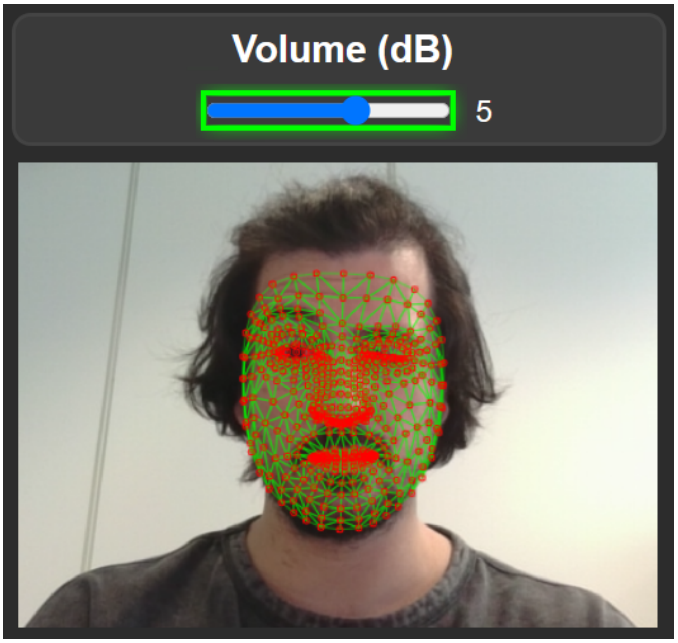

Fig. 3. The image shows a user interacting with a volume control interface using facial gestures. A facial landmark mask is overlaid, highlighting the tracking system that detects specific expressions (in this case, a right-eye wink used to increase the volume level).

wink determines the direction of adjustment (see Fig. 3): right eye to increase, left eye to decrease.

3) *Shortcut* commands provide rapid access to frequently used actions. For instance, smiling initiates or stops audio recording, allowing the user to stay engaged with track-level controls without navigating away from the current context.

The facial gesture design was guided by findings from HCI research, prioritizing intentionality, low muscular effort, and metaphorical clarity [17], [18].

IV. Results and Discussion

To evaluate the effectiveness and robustness of the proposed multimodal interaction system, a preliminary functional test was conducted involving six users. Each participant was involved in an individual session lasting approximately 30 minutes, during which they were instructed to interact with the WAI using both voice and facial control modes. During each session, each participant performed approximately 200 activation attempts per modality. For each modality, all activation attempts were logged, and recognition accuracy was calculated as the percentage of successful activations over total attempts.

Table I reports the recognition accuracies obtained across participants for both voice and facial commands. The results confirm a generally high level of accuracy across users. In particular, the voice-based mode exhibited consistently high accuracy, with minimal inter-user variability. Conversely, the facial gesture-based mode demonstrated slightly more variability in performance across users, although still within an acceptable operational range. This variation can be attributed

TABLE I
Accuracy of Voice and Facial Commands per User

User	Sex	Voice Accuracy (%)	Facial Accuracy (%)
User 1	F	91.5	90.7
User 2	M	92.5	87.7
User 3	M	93.8	86.6
User 4	F	93.1	95.5
User 5	M	90.4	92.6
User 6	F	91.6	93.8

to individual differences in facial expressiveness and gesture execution, which are inherently more diverse and harder to normalize than spoken commands.

While these findings validate the overall reliability of the prototype, they also highlight areas for further optimization. In the case of voice commands, future versions of the system could benefit from an expanded command dictionary that accounts for a broader set of phonetic variants and near-homophones. This would further minimize recognition errors due to accent, pronunciation inconsistencies, or ambient noise.

Regarding facial interaction, a promising direction is to integrate a sensitivity control slider into the WAI. This feature would allow users to manually adjust the sensitivity threshold for gesture detection according to their own facial mobility and comfort, enabling personalized calibration.

V. Conclusions

This work presented the design and preliminary validation of a microcontroller-based embedded system that extends standard audio interfaces with multimodal interaction capabilities, specifically tailored for EMP environments. The proposed solution integrates audio streaming with a browser accessible WAI, enabling users to control core production functionalities through voice and facial input, besides using manual controls. A functional evaluation involving six users demonstrated the preliminary effectiveness of the system. The results highlighted a high recognition accuracy for both interaction modes, with voice commands showing minimal variability between users and facial gestures showing slightly more variability with respect to individual differences. Future directions will include tests in EMP environments, evaluating both system performance and pedagogical effectiveness.

Acknowledgment

This publication is part of the project PNRR-NGEU which has received funding from the MUR- DM 118/2023. The authors acknowledge the support of the “Musical Metaverse: an inclusive Extended Reality platform for networked musical interactions” project (grant n. 2022CZWWKP) – funded by European Union – Next Generation EU within the PRIN 2022 program (D.D. 104 - 02/02/2022 Ministero dell’Università e della Ricerca) and of the MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement n. 101184379). This manuscript reflects only the authors’ views and opinions and the Ministry, the European Union or the European Innovation Council cannot be considered responsible for them.

REFERENCES

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

[2] A. P. Bell, *The process of production — The production of process: The studio as instrument and popular music pedagogy*, R. Wright, Ed. Canadian Music Educators' Association, 2017.

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

[4] G. Palaigeorgiou and C. Pouloulis, "Orchestrating tangible music interfaces for in-classroom music learning through a fairy tale: The case of improvischool," *Education and Information Technologies*, vol. 23, no. 1, pp. 253–271, 2018.

[5] 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 *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 2021, pp. 1–6.

[6] M. Bremmer, C. Hermans, and V. Lamers, "The charmed dyad: Multimodal music lessons for pupils with severe or multiple disabilities," *Research Studies in Music Education*, vol. 43, no. 1, pp. 132–149, 2021.

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

[8] J. Yu, T. Zhang, S. Wu, and X. Wu, "Architone: A lego-inspired gamified system for visualized music education," arXiv preprint arXiv:2410.15273, 2024.

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

[10] 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 *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, 2024, pp. 1–4.

[11] F. Sun, "Analysis of virtual reality-based music education experience and its impact on learning outcomes," *Scalable Computing: Practice and Experience*, vol. 25, no. 6, pp. 4755–4762, 2024.

[12] M. J. Cook, "Augmented reality: Examining its value in a music technology classroom. practice and potential," *Waikato Journal of Education*, vol. 24, no. 2, pp. 23–38, 2019.

[13] 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 *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2018, pp. 1–7.

[14] T. Colafiglio, C. Ardito, P. Sorino, D. Lofù, F. Festa, T. D. Noia, and E. D. Sciascio, "Neuralpmg: A neural polyphonic music generation system based on machine learning algorithms," *Cognitive Computation*, vol. 16, pp. 2779–2802, 2024.

[15] C. Myers, A. Furqan, J. Nebolsky, K. Caro, and J. Zhu, "Patterns for how users overcome obstacles in voice user interfaces," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, pp. 1–12.

[16] 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 *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 2022, pp. 373–384.

[17] K. Masai, K. Kunze, D. Sakamoto, Y. Sugiura, and M. Sugimoto, "Face commands - user-defined facial gestures for smart glasses," in *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2020, pp. 456–467.

[18] H. Y. Leng, N. M. Norowi, and A. H. Jantan, "A user-defined gesture set for music interaction in immersive virtual environment," in *Proceedings of the 3rd International Conference on Human-Computer Interaction in Indonesia (CHIuXiD)*. ACM, 2017, pp. 29–36.