Extended Reality Audio-Visual Instruments: Design Framework and Case Study

Esther Gruy
esther.gruy@univ-lille.fr
CRIStAL, CNRS, Univ. Lille
Lille, France

Florent Berthaut
florent.berthaut@univ-lille.fr
CRIStAL, CNRS, Univ. Lille
Lille, France

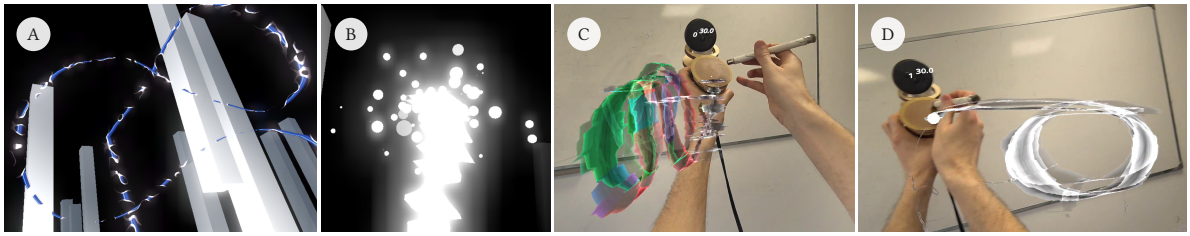

Figure 1: Diverging design of two XRAVI **from the same initial instrument. A) Visual-First instrument: creating strokes inside the environment; B) Visual-first instrument: particle effect when drawing; C) Audio-First instrument: generating complex visual strokes from virtual and physical sonic expression; D) Audio-first instrument: using visual traces as a guide to replay a sonic exploration.**

Abstract

Scientists, artists and performers have explored the relationship between visuals and sound on numerous occasions, to try and find/define ways to fuse both sensory experiences together. Immersive technologies further amplify the possibilities for multimodal expression in performance contexts. While previous frameworks exist for the design of immersive musical instruments and audio-visual instruments independently, we believe that their combination raises specific concerns, which might prevent the exploration of expressive opportunities. In this article, we define the concept of Extended Reality Audio-Visual Instruments (XRAVI) and propose a framework for their analysis and design, based on the literature on audio-visual and extended reality instruments. We evaluate it through a collaborative autoethnographic work, where an initial shared instrument has been developed and practised following two different approaches: audio-first and visual-first. From it, we derive insights and guidelines for the design of XRAVI.

Keywords

Audio-Visual, Extended Reality, Mixed Reality, Framework, Autoethnography, Instruments

1 Introduction

The relationship between visuals and sound has been explored for centuries [7]. The first instances of instruments mixing music with visual elements were called colour-organs, such as Castel's *Clavecin Oculaire*, Rimington's *Colour-Organ* (which has made the term generic), or Wilfred's *Clavilux* [33]. These instruments, however, were complex and expensive to conceive, which prevented their generalisation.

The twentieth century's technological advancements, with the widespread democratisation of computers, have made it easier for artists and performers to create their own audio-visual applications or instruments (*i.e., the simultaneous authoring of both dynamic image and sound* [21]). For example, the UPIC system, achieved in 1977, blends drawing on a 2D surface with sound production, creating audio-visual compositions that are very malleable. Similarly, Levin has worked on *painterly interfaces* [21], using drawings, shapes and animations as a metaphor for sound production, where both auditory and visual aspects are considered equally important in the design of the instruments and the composition process.

More recent advancements, such as the rise of Extended Reality (XR) technologies in the last decade, has led to the expansion of Extended Reality Musical Instruments (XRMI) [38]. As XR leads to the addition of 3D content inside a virtual space (whether it is mixed with the physical space or fully virtual), it allows for new interaction methods and opportunities for musical creation [3]. XR also benefits from its capacity to display complex visuals (for example with headsets or spatial projection), which makes it adapted to the creation of audio-visual instruments and performances. However, the combination of visual and sonic expression in Extended Reality, through what we can call Extended Reality Audio-Visual Instruments (XRAVI), has not been explicitly and thoroughly investigated.

As musical instruments are typically correlated with gestures [6], XRAVI create a tight relationship between the sound, the visuals and the performer's movements. Depending on the design process and the chosen mappings, these modalities can influence one another in various ways, and create multisensory experiences that are vastly different from one another. However, the absence of obvious mappings between sound properties and visual effects raises questions on the process by which such an instrument is designed. Moreover, the differences between 2D and 3D interaction induce a spatialised relationship with the instrument's content (audio and visual). If, on one hand, XR instruments do not typically raise the visual content at the same level as the auditory content, on the other hand, audio-visual interfaces are less reliant

NIME '26, June 23–26, 2026, London, United-Kingdom

on these spatialised interactions. This creates a specific set of constraints that lies at the intersection of both types of instruments, and which we believe requires thorough investigation.

In this article, we propose a framework for the analysis and design of XRAVI, based on previous literature, and establish several dimensions to provide a better understanding of the conceptual choices behind these instruments. We then describe the design process of two new XRAVI, derived from a common prototype and through a collaborative autoethnographic approach. The objective is to analyse how the design of these instruments can diverge over time when the focus is put either on the audio modality first, or on the visual modality first. We ground these instruments into our framework, which allows us to provide a discussion on the influence of each modalities on the design process, over a complete timeline. We conclude by giving guidelines on the design of XRAVI.

2 Related Work

Our research lies at the intersection of Extended Reality Musical Instruments and Audio-Visual Instruments.

2.1 Extended Reality Musical Instruments

XRMIs have been studied from multiple perspectives. Serafin *et al.* [35] present the history, opportunities and challenges of Virtual Reality Musical Instruments, followed by Turchet *et al.* [38] which expand the analysis to include Mixed Reality instruments. Berthaut [3] analyses immersive musical instruments through the lens of 3D interaction techniques. Among the 3D manipulation techniques described, some relate to changes in material and shape, which means that there are interactions with both the visual and sonic aspects of virtual content.

In these analyses and reviews, however, the primary focus remains on the control of sound. Often, the instruments afford very limited visual changes, such as transition/rotation/ scaling of preexisting virtual objects, and sometimes only the selection of these objects. The visual component of XRMIs often consists in replicating physical instruments (acoustic or electronic) at different scales and with different organisations. This paper instead investigates instruments that allow for a rich visual manipulation of virtual content.

2.2 Audio-Visual Instruments

Audio-visual instruments, or performances, can take many different forms and approaches. As briefly mentioned in the introduction, one possible interaction metaphor for sound production can be drawing, either in 2D [1, 17, 21, 34, 41], or in 3D [5, 9, 28, 30, 39]. Although some of these projects do not necessarily intend to provide a complete audio-visual experience and focus on the sound production first, the naturally visual aspect of drawing still puts them in the audio-visual instruments category.

Aside from drawing, audio-visual instruments can be driven from different artistic gestures, like pottery [11], or sculpting [25, 26], which all imply 3D interaction techniques, even when visualised on a 2D screen. Other projects work using gestures, whether they are driven by 2D [22] or 3D hand movements [29], by external controllers [13, 18, 37], or by full body movements [8, 16], like for example dancing [14, 36].

There exists interaction metaphors that are not necessarily based of off gestures, for example with biological signals [23, 27] or particle systems [31]. It is also possible to find audio-visual performances controlled by writing [20] or live coding [19].

We can note that some audio-visual instruments/performances use specific equipment in order to display the content or interact with the system, such as transparent reflective panels [5, 13], tabletop configurations [12, 15], or lasers [32].

In this work, although our instruments generate an audio-visual output using 3D drawing, we acknowledge the plurality of possible interaction metaphors, hence the idea of creating a framework for XRAVI.

3 Extended Reality Audio-Visual Instruments

As stated in the introduction, Extended Reality Audio-Visual Instruments (XRAVI) merge audio production with the generation of visual content inside a 3D space. The next subsection gives a more detailed definition of XRAVI, based on previous literature about either audio-visual instruments or XRMI.

3.1 Definitions

According to Levin [21], there are several elements that should compose an expressive audio-visual instrument, namely: 1) the simultaneous, real-time creation of both images and sound, 2) a produced content that is inexhaustible, variable and deeply plastic, 3) malleable sonic and visual dimensions, 4) the definition, by a performer, of their own visual languages, and 5) a system that is easily understandable while allowing for the development of a sophisticated practice. What is of interest for our definition of XRAVI are the notions of real-time content authoring, equal attention put on the audio and visual outputs, the variability of the interaction, and the possibility to develop a mastery of the instrument over time.

Looking at the definition of Musical XR given by Turchet *et al.* [38], their prerequisites encompass: 1) the existence of virtual elements in one or more sensory modalities, 2) the spatial persistence of these virtual elements, 3) the interactivity of the system, and 4) a sonic organisation to convey meaning, intent and focus that is an essential component of the user experience. In their framework, Zellerbach and Roberts [42] define Mixed Reality Musical Instruments as *an embodied system for expressive musical performance, characterized by the relationships between the performer, the virtual, and the physical environment.* Here again, what we can gather from these elements are the possibility to interact with virtual elements, no matter the sensory modality, inside a virtual space, and the relationship between the user, the environment and its content.

By merging these definitions, we propose the following criteria for what constitutes an XRAVI:

(1) The system makes possible the creation and performance of virtual, spatially persistent images and sounds in real-time.
(2) Sonic and visual dimensions are equally malleable and manipulated through 3D interaction techniques and physical interaction devices.
(3) The interaction with both visuals and sounds follows a "low entry fee / no ceiling on virtuosity" approach.
(4) There is an intricate relationship between gestures, visuals and sound in the virtual or physical space.

XRAVIs therefore combine both constraints and opportunities of XRMIs and audio-visual instruments, with respect to the "spatial" aspect of interactions, the immersion of performers and audiences, and the intricate relation between sounds and visuals.

3.2 A Framework for Analysing XRAVI

Several frameworks, related to XRMI or audio-visual instruments, have been proposed in recent years, to analyse design choices, propose conception guidelines, and uncover under-explored areas of a specific research field. For instance, Zellerbach and Roberts [42] have created a framework for Mixed Reality Musical Instruments, Gruy and Berthaut [10] have reviewed 2D and 3D musical drawing instruments, and Serafin *et al.* [35] have proposed an overview of Virtual Reality Musical Instruments. Although close in the themes they explore, these frameworks do not dive into the specifics of both XR and audio-visual instruments mixed together, since they either focus on the musical aspect exclusively, or only look at one interaction metaphor within audio-visual instruments (namely drawing). Thus, we believe that this framework could benefit the design of future XRAVI.

In order to find XRAVI and analyse their structure, we searched on several digital libraries (NIME, ACM, IEEE), exclusively choosing those that went through a peer-review process. We then proceeded to read the abstracts and kept the articles that seemed to fit the previously established criteria. We read the articles to refine the list, and gathered each instrument's characteristics to find the following common dimensions.

3.2.1 Performer Visual Transportation. The Performer Visual Transportation (PVT) follows the transportation dimension established by Benford *et al.* [2]: consequently, it here represents the extent to which the performer feels that they have left their physical visual space to enter a remote visual space. For example, the performer could see the visual environment on a simple screen positioned in front of them (Low), or be immersed inside a virtual environment where they cannot see the physical space (High).

3.2.2 Performer Sonic Transportation. The Performer Sonic Transportation (PST), similarly to the previous dimension, represents this time the extent to which the performer feels that they have entered a remote sonic space, in opposition to a local one. That would imply, for example, having sound come out from a single source where outside noises are distinguishable (Low), or oppositely, have spatialised sound that fills up the whole space (High).

3.2.3 Audience Visual Transportation. The Audience Visual Transportation (AVT) is similar to its performer counterpart, with the difference that audience members (or users that are the spectators of another person's interactions) are the main focus.

3.2.4 Audience Sonic Transportation. Similarly to the previous dimension, the Audience Sonic Transportation (AST) is the audience counterpart of the Performer Sonic Transportation.

3.2.5 Interaction Metaphor. The Interaction Metaphor (IMe) dimension corresponds to the type of gestures, performed in the virtual environment, that drives the interaction. That can be, for example, gestures that mirror what already exists in the physical space (drawing, sculpting, moving or rotating an object...), or that take advantage of XR environments and the possibilities they offer (revealing virtual content...).

3.2.6 Interaction Modality. The Interaction Modality (IMo) dimension looks at the physical means by which the interaction happens (regardless of the equipment or the hardware that is employed). A few examples could be clicking on buttons, using one's voice, doing various hand poses (*i.e.,* gesture recognition), or using a sound input from an acoustic instrument.

3.2.7 Inter-Modalities Mappings. The Inter-Modalities Mappings (IMM) dimension represents how each modalities (sound, visuals and gestures) interact with one another. For example, the gestures can drive the sound that then drives the visuals (G->S->V), the sound alone can have an effect on the visuals (S->V), or on the contrary, the visuals can influence the sound (V->S), and so on.

3.2.8 Collaborative Possibilities. The Collaborative Possibilities (CP) dimension looks at whether the XRAVI allows multi-user interactions, or is limited to a single person. It does not take into account the potential addition of collaborative features if they are not implemented, but are envisioned in future work.

3.2.9 Use of Space. The Use of Space (UoS) dimension dives into the way an XRAVI takes advantage of the surrounding space as part of the instrument's interaction modality. It can be, for example, divided into different zones, which change the mappings, the sounds, or the visuals. It can also be continuous, with the possibility to move around without changing the mappings. Finally, it can be static and only used as a way to display visual and sonic information (screens, projections, audio spatialisation...) to an audience.

3.3 Analysis of Existing Instruments

A total of 13 instruments were retained during the selection process: 10 of these were presented as conference papers [5, 8, 9, 11, 13, 18, 25, 26, 28, 37], 2 as posters [30, 39], and 1 as a demo [29].

The first instrument, *Reflets* [5], creates performances with 3D musical virtual interfaces. The audio-visual effects are activated by slicing though 3D objects, projected on audience members and reflected on a semi-transparent panel, so that they appear floating around the musician.

The *AirSticks* [13] also display the visual content using a semi-transparent panel and a projector. They use gestures to influence the sound and the visual parameters.

Cymbalism [37] makes use of the *AirSticks*, this time placing the accent on the visuals to inspire the gestural and sonic aspects of the instrument, and projecting the content on a surface.

GeKiPe [8] is an audio-visual gestural interface designed for performances. It works with cameras and sensors, and displays the visuals using a projector.

Entangled [18] is a smartphone-based, multi-users instrument that uses gestures and particles, influenced by gravitation, to drive the audio-visual features.

Musical Brush [39] is also smartphone-based. Drawings in the 3D space drive the sonic output.

Similarly, MagneTip [9] uses 3D drawing to produce sound, this time with a headset. The physical interaction between a copper coil and a magnet, that acts as a controller, generates the audio-visual output.

Drawing Sound in MR Space (DSMR) [28] works with a Mixed Reality headset, and 3D drawings produce audio and visual effects in a multi-users environment.

Drawing Lines to Connect (DLC) [30] expands on DSMR. In addition to the 3D drawing in mid-air, there are 360 AI generated

landscapes, and emotion estimation using physiological signals to change the strokes' colours.

Sound Sculpting [26] is a virtual instrument working with sensors to track hand gestures. The sonic output is controlled by interacting with 3D virtual objects, that are displayed on a screen.

Ashitaka [25] follows this 3D sculpting metaphor. The gestures, performed on a device with several sensors, drive the audio-visual output, with the image displayed on a screen.

Virtual Pottery [11] uses hand gestures to create 3D pottery objects and sound at the same time, with the visuals being projected on a surface.

Finally, Air Maestros [29] is a multi-users Mixed Reality music sequencer. 3D note objects are placed in the space, and audio-visual effects are activated by shooting them using gestures.

Table 1 details how these instruments fit into the previously described framework. Some equipment specifics were not given, as they might be context-dependent (*e.g.,* auditory displays impact both PST and AST dimensions), which explains the "unspecified" values. To fill the PST and AVT boxes, we used the information that was available regarding the possibilities of the system, or the setups in which performances were held (therefore, this analysis is potentially subject to interpretation). In general, if the sound parameters and mappings are explained, their displays can be easily changed or adapted to a venue.

A few conclusions can already be drawn. First of all, there are no high PVT, since a lot of the instruments were used in performances or multi-users settings, which induces the creation of a shared space between the performers and the audience. Completely leaving the physical space behind raises a lot of constraints, mainly equipment-wise (as a fully immersive VR setup requires headsets). Therefore, if the AVT is an important part of the instrument, the PVT can hardly go above medium, unless the content of the headset is streamed to the audience (which will most likely lead, in return, to a low AVT).

If we look at the IMo dimension, most of the instruments use some types of gestures, instead of physical objects, to drive the interaction. This is likely because the IMe are gesture-based, and therefore it creates a continuity between the two dimensions. As for the IMM, they follow the same idea, since all audio-visual effects are gesture driven, even though the visuals and sound can then influence one another (depending on the mappings).

For the UoS dimension, it is in part linked to the IMe, as the type of gestures implied by the metaphor at least partly defines how a performer will move inside the space. For example, all instruments that are based off of drawing use the space either continuously or in zones (as building a drawing takes up space), and those based off of sculpting are mostly static (the only exception being because there are several sculptures in the scene [11]), as a sculpture usually stays in one place.

Finally, for the CP dimension, every project could, on a technical standpoint, support collaborative features. Whether it is a single-user or multi-users instrument mostly depends on either the choices made by the designers, or by some technical or material limitations (building several instruments, adapting the setup to a venue...).

4 Studying the Divergent Design of two XRAVIs

In order to gain insights on and provide guidelines for the design of XRAVIs, we investigate the design and practice of two diverging instruments, analysed through our proposed framework.

4.1 Methodology

We followed a *collaborative autoethnographic approach* [24] to study the evolution and appropriation of two divergent versions of an instrument. Two of the authors therefore each designed and practiced a version that focused on a different modality (visual-first and audio-first), in order to ensure contrasting choices and to nourish the discussions and analyses. Our approach was to rely on the corresponding existing artistic practice of the authors (mostly visual/graphical and mostly musical) in order to inform the analysis.

The author who worked on the audio-first XRAVI has been playing music for more than 20 years (+ 10 years of acoustic drums), with electronic drum pads and live-looping, with augmented acoustic drums and with Extended Reality Musical Instruments, mostly in structured improvised performances.

The author who worked on the visual-first XRAVI has been drawing for over 15 years using various techniques (pen, pencils, alcohol markers, various inks, pastels, watercolour, acrylic paint) and has followed academic courses on art and its history.

The study lasted 5 months. Both authors did at least two individual design and practice sessions per week, between September 15th 2025 and January 20th 2026. Once a week, the two authors filmed a video of the current version and met to discuss and log all changes made to the instruments, newly discovered playing techniques and gestures, difficulties/issues, planned or desired changes and improvements.

4.2 Initial Shared Prototype

The initial shared prototype relies on the MagneTip instrument [9]. MagneTip relies on a "disassembled speaker" (Figure 2), with a flat coil layered on a membrane in which the amplified audio signal is sent and creates an oscillating magnetic field, and a magnet that the musician moves closer to the membrane in order to generate mechanical oscillations and therefore sonic output. The membrane is glued on top of a wooden resonant cylinder.

It is implemented with a Meta Quest headset (versions 2 and 3 were used). One of the controllers is attached to a physical device held in the user's non-dominant hand, while the other hand holds the stylus with the magnet. The sonic output and other interactions with the box are captured via a piezo transducer placed below and in contact with the membrane.

The software was developed using the IVMI-builder framework that relies on the Godot game engine and Pure Data [4].

The initial instrument therefore provides as inputs: controller and head positions and orientations, interaction with the buttons and joystick on the controller, audio input from physical interaction with the membrane and box. Outputs are: 3D visual rendering, sound output to the box membrane, sound output to an external speaker.

4.3 Final Version of the Audio-First Instrument

For the audio-first version of the instrument, the goal was to obtain an expressive tool for improvised musical performances,

Table 1: How different audio-visual instruments fit into the framework

Instruments	PVT	PST	AVT	AST	IMe	IMo	IMM	CP	UoS
Air Maestros [29]	Medium	Unspecified	Medium	Unspecified	Shooting	Gesture recognition	G->VS	Multi-users	Continuous
AirSticks [13]	Medium	Medium	Medium	Medium	Moving	Gestures	G->S->V	Single-user	Static
Ashitaka [25]	Low	High	Low	High	Sculpting	Gestures	G->VS	Single-user	Static
Cymbalism [37]	Medium	High	Medium	Medium	Percussion	Gestures	G->VS	Single-user	Continuous
DLC [30]	Medium	Unspecified	Medium	Unspecified	Drawing	Gesture recognition	G->V->S	Multi-users	Continuous
DSMR [28]	Medium	Unspecified	Medium	Unspecified	Drawing	Gesture recognition	G->V->S	Multi-users	Zones
Entangled [18]	Medium	High	Medium	High	Moving	Buttons	G->VS	Multi-users	Continuous
GeKiPe [8]	Low	High	Medium	High	Moving	Gesture recognition	G->VS	Multi-users	Zones
MagneTip [9]	Medium	High	Low	Medium	Drawing	Sound input	G->S->V	Single-user	Continuous
Musical Brush [39]	Low	Unspecified	Low	Unspecified	Drawing	Buttons	G->VS	Single-user	Continuous
Reflets [5]	Medium	Unspecified	Medium	Unspecified	Revealing	Spatial position	G->VS	Multi-users	Continuous
Sound Sculpting [26]	Low	Unspecified	Low	Unspecified	Sculpting	Footswitch	G->V->S	Single-user	Static
Virtual Pottery [11]	Low	Medium	Low	Medium	Sculpting	Spatial position	G->VS	Single-user	Zones

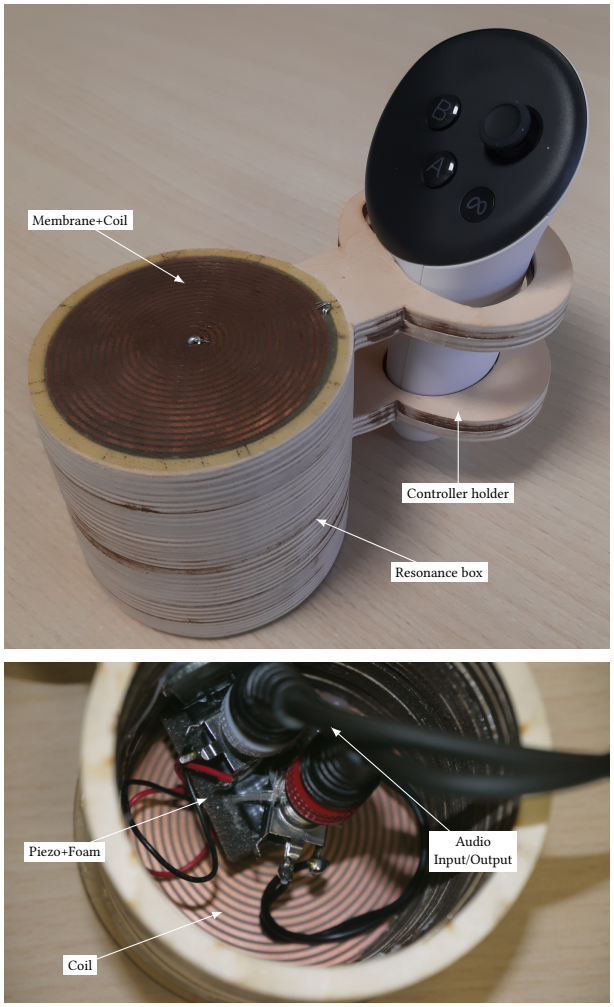

Figure 2: Initial instrument. Top: laser-cut resonance box and engraved membrane with coil. Bottom: connections to the coil as an audio output, and from the piezo pressed on the membrane as an audio input

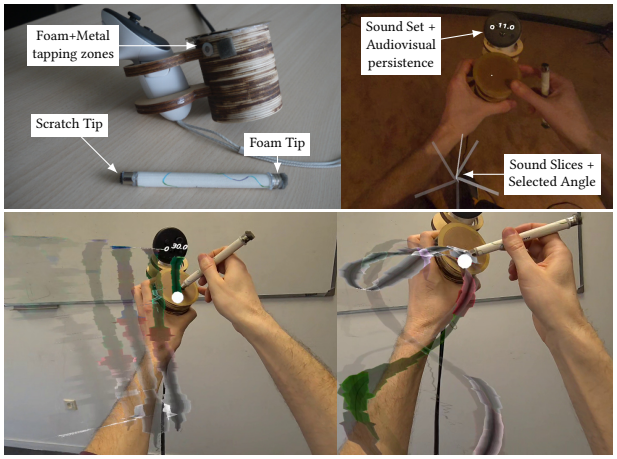

Figure 3: Physical and visual modifications in the audio-first instrument. Top left: added materials for extended interaction with the audio modality. Top right: visual feedback on the UoS with sound slices and buttons as IMo. Bottom left: sonic and visual persistence is controlled with the joystick or audio onsets, it allows for building audio-visual landscapes. Bottom right: hovering and scratching the membrane generates a variety of audio spectrum and visual strokes.

based on the exploration of a sonic space which would result in visual (*i.e.,* 3D brush strokes) and musical traces.

In order to increase the sonic capabilities of the instrument, the box and stylus were both modified, as shown in Figure 3. Foam and metal pieces were glued to the box, so that they can be used to play rhythms by tapping with the index and middle fingers of the non-dominant hand. The stylus was modified so that both ends can be used, *i.e.,* with a magnet but two different materials: scratch and foam. This allows for the mix of percussive playing techniques with the exploration of the electromagnetic field created by the coil. Consequently the membrane was reversed so that the thin layer of copper is not damaged when tapping and scratching the membrane. For such interactions to be possible,

the instrument is displayed in Mixed Reality, so that fine grained interactions on the membrane are clearly perceivable.

The explored sonic space is a cylinder ($height = 2m, radius = 1m$) centred on the musician. Sounds played using granular synthesis are assigned to slices of the cylinder (*i.e.,* each within an angle interval), meaning that moving the controller around the musician's position allows for crossfading between sounds. Within each slice, the vertical position of the controller moves the offset of the granular synthesis within the sound. A button press on the controller changes between cylinders, *i.e.,* between sets of sounds. A variable audio delay effect is applied to the sound captured on the physical box. Delay duration is either defined according to inter-onset duration, or at a fixed value changed with the controller's joystick. This allows the musician to build more complex phrases and rhythms, with either variable or fixed tempos.

As shown in Figure 3, the visual output of the instrument is driven by both the gestures and the sound: the spectrum and perceptual features extracted from the captured sound are mapped to the appearance of the strokes, while the stroke's position is mapped to controller's position. As seen in Figure 1, the pitch, brightness and noisiness features are respectively mapped with the hue, luminance and saturation of the stroke's colour. The overall shape, in layers from the centre of the stroke, depends on the spectrum of the sound, divided in 12 frequency bands. Drawing is activated when the sound loudness is above a defined threshold. The visual persistence of the 3D strokes delay feedback is mapped to the delay duration, so that the "sonic" strokes remain audible roughly as long as the visual strokes are visible.

Additional visual feedback includes delay control mode/duration and cylinder/sound set number shown on the controller, and the limits of cylinder slices/sounds and hand angle/distance shown on the floor, as seen in Figure 3.

A large variety of gestures and playing techniques were developed during the design process, some of which are shown in Figure 1. This includes: 1) hovering at various positions above the coil to explore the electromagnetic field and modulate the rendered sound; 2) tapping/scrubbing the surface with the two tips of the stylus for changes in noisiness and brightness; 3) repeating a musical phrase by following previous visual strokes, while introducing variations on positions or on physical interactions; 4) alternating between fast changes in sounds and visuals with the controller close to oneself and large strokes in the same sound with the controller further away; 5) visually and sonically filling the space with slow movements and loud sounds; 6) moving the head forward to displace the cylinder and be able to switch between sounds with circular gestures.

4.4 Final Version of the Visual-First Instrument

For the visual-first instrument, the goal was to create an environment that could be explored using the device, with strokes leading to various effects on the space and on the sound.

In order to immerse the user inside the environment, it is displayed in VR. The background is fully black, with white pillars randomly positioned in a circular pattern around the starting position. There are two lights inside the environment (under and slightly above the pillars to create depth), and they are dimmed according to the current visualisation.

For the exploration, four shaders were designed (with examples available on Figure 4):

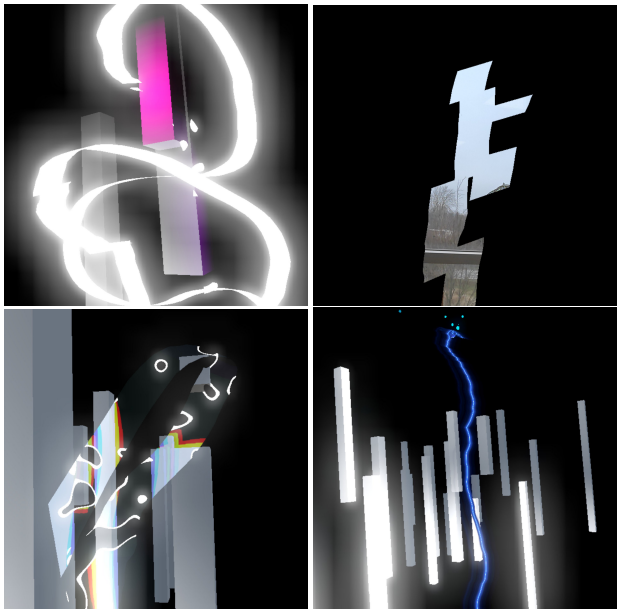

Figure 4: Visual-first instrument: images of the different brush modes. From top left to bottom right: *Glow*, *Passthrough*, *Chromatic* and *Thunder*.

(1) *Glow*: The strokes are white and made of an emissive material. White, sparkle-like particles are emitted from the brush when drawing. In this mode, all the environment lights are turned off. The brush and the strokes are the only elements producing light (with varying tints based on spatial position for the strokes).
(2) *Passthrough*: All the environment lights are turned off, leaving the background to be fully black. When drawing, the shader relies on the *shadow to opacity* Godot blend mode, which is designed to merge the physical and virtual spaces together. This reveals the physical space onto the strokes as if tearing the background open.
(3) *Chromatic*: This shader uses a stencil buffer to reveal elements (cubes) hidden inside the space (fully lit). The strokes appear transparent, but they also distort the environment behind them with a chromatic effect (red, green and blue hues). Only the hidden elements are not affected by it, setting them apart from the pillars.
(4) *Thunder*: The strokes resemble deep blue lightning bolts. The shader is dynamic, meaning that the thunder effect moves over time. Same as for *Glow*, particles are emitted from the brush. The environment is fully lit.

For all the visualisations, three temporalities (short, medium and long) can be chosen by clicking on a controller's button. This implies a fading effect on the strokes, as they are not persistent. For all the visualisations except *Passthrough*, which shrinks by getting thinner, a dissolving effect is applied (see Figure 4 with the *Chromatic* brush).

As for the sound, made using granular synthesis, it draws a direct inspiration from the visuals, in order for the two components to match. The gestures also have an effect on the resulting sound. It is described as follows:

(1) *Glow*: The sound is high pitched, with delay and reverberation to make it sound smoother. The height (*i.e.,* the Y-axis) of the brush creates slight variations of the pitch,

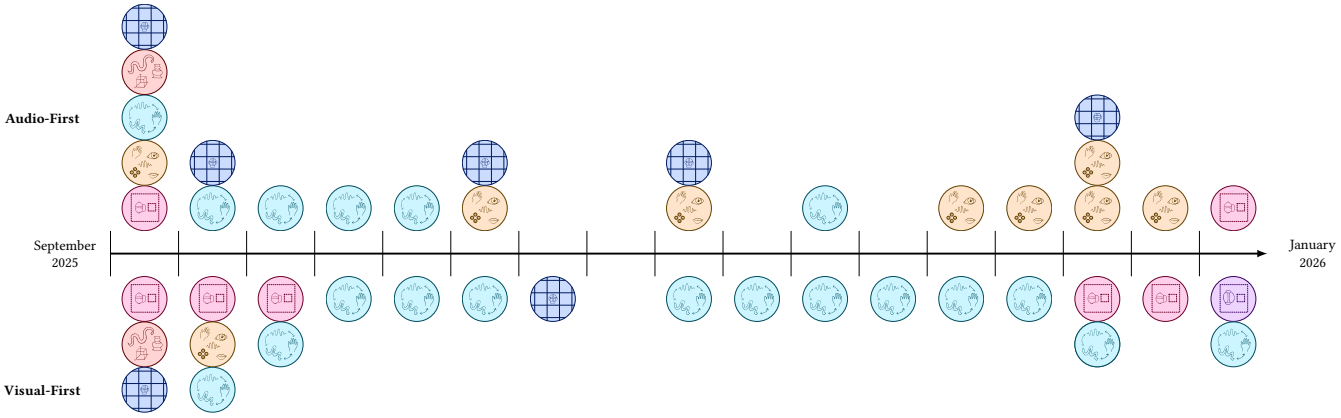

Figure 5: Timeline of per-week changes for each instrument organised by dimensions of the XRAVI framework

and the speed of the brush has an effect on the distortion (the faster it gets, the more distorted it becomes).

(2) *Passthrough*: The physical environment's sounds are amplified with added reverberation and delay.
(3) *Chromatic*: The sound resembles a low pitched white noise, with the addition of delay to make it seem like the space has echo. There is a pulsing effect that indicates the proximity with some of the hidden objects. The height of the brush changes the pitch.
(4) *Thunder*: The sound has a low pitch, somewhat granular, with reverberation and distortion for it to mimic actual thunder. The speed of the brush has an effect on the distortion (the faster it gets, the less there is), and the position of the brush on the Y-axis changes the pitch.

4.5 Analysis of the Design Process

Figure 5 shows which dimensions were modified during the design process for each versions of the instrument, and Table 2 details values for the final dimensions. The analysis of the design process is detailed in the next sections.

4.5.1 Performer Visual Transportation. For the audio-first instrument, the PVT was defined from the start as very low in order to preserve visibility over the physical instrument and fine-grained gestures which can be performed on and above the membrane. For the visual-first instrument, the PVT was the first consideration, since putting the environment in complete VR defined its main aesthetic. It was updated a few times during the design process, as the idea on how the environment would look like evolved over time.

4.5.2 Interaction Metaphor. The audio-first instrument remained focused on the drawing metaphor, therefore the IMe was not modified. For the visual-first instrument, an additional IMe was designed at the beginning. It allowed users to replay the sounds by moving the device around the strokes. However, it ended up being scraped, as it did not fit within the project as a whole.

4.5.3 Interaction Modalities. In the audio-first instrument, the IMo underwent the following modifications: onsets from tapping the box were used to control the audio-visual persistence of the strokes; new scrubbing and tapping techniques with the modified stylus and box; buttons and joystick to control sound synthesis parameters. Multiple weeks were dedicated to refining practice with new gestures, and controls at the end were added very incrementally, once it was clear that they were really necessary. For the visual-first instrument, the IMo was changed at the beginning to add an interface, made visible using a button. It allowed to disable the fading of the strokes and to erase the canvas. A few additional options were planned, but the idea was abandoned after a few weeks, since it would have added too much complexity to the overall design.

4.5.4 Inter-Modalities Mappings. In the audio-first instrument, although the interaction between gestures, visuals and sound was clear from the start, it was refined and amplified over time. The extracted audio features and the sound to visual mappings were often modified, so that the visual output correctly reflects changes in the perceptual features of sounds captured by the piezo. For the visual-first instrument, the IMM was the most explored dimension, as the interaction relied on the relationship between the gestures and the visual outcome.

4.5.5 Use of Space. In the audio-first instrument, the UoS became an important aspect as it allows for the selection of sounds and for granular synthesis parameter control. Over time the position on the Z-axis was removed from the mappings, but sets of sounds associated with corresponding subdivisions of space were made changeable. For the visual-first instrument, the UoS was tweaked a few times, as there are zones that change the active audio-visual effects. Once the code was correctly defined, there was no need to touch it anymore.

4.5.6 Performer Sonic Transportation. In the audio-first instrument, the PST was defined from the beginning and did not change, because the system already allowed for both local (with the physical box) and global sound diffusion. For the visual-first instrument, the PST was defined at the very end of the design process, because the sounds were added only when the visual side of the instrument was considered final (since the sonic aspect serves the purpose of grounding the visuals into the virtual environment in a cohesive way).

4.5.7 Unused dimensions. For both prototypes, some dimensions were not explored in the design process. Audience considerations were not situated at the core of the instrument, since no actual performance could be planned until we had a functioning prototype. Since both instruments work with an XR headset (and therefore will rely on projecting the performer's view to the audience), the AVT will most likely be low. The AST will depend

Table 2: How the audio-first and visual-first instruments fit into the framework

Instruments	PVT	PST	AVT	AST	IMe	IMo	IMM	CP	UoS
Audio-first	Medium	High	Low	Medium	Drawing	Sound	G->S->V	Single-user	Zones
Visual-first	High	High	Low	Medium	Drawing	Sound	G->VS	Single-user	Zones

on a venue's logistics, but a medium transportation should be expected (hence the values reported in Table 2). Collaborative possibilities were not explored, as both versions of the instrument focused on a single-user use case.

5 Discussion

From our review of existing instruments and the analysis of the design process of our two instruments, we derive insights and guidelines for the design of XRAVI.

5.1 Design through Dialogue and Interactions between Modalities

In both instruments, one can observe a *dialogic* rather than *synchronous* process between audio and visual modalities during design [40], *i.e.,* one modality was designed as a response to the other.

In the audio-first instrument, visual strokes served first as a guide for the sonic exploration of the physical space. Only when this aspect was sufficiently refined, in that case with the correct transparency settings to see overlapping strokes, that the focus changed to making sure that the visuals reflected expressively what was happening in the sound. Strokes were eventually made larger and longer lasting in order to make the visuals more salient and improve the balance between audio and visual output. In the visual-first instrument, the sound was used to add depth to the visual effects and ground them into the environment. As a result, it is only when a clear idea of the visual direction emerged that the sonic aspect of the instrument was explored.

In both cases, one can therefore observe different interactions between modalities, which evolve over time. We believe that this relation, through the IMM dimension of our framework, should be regularly examined during the design process.

5.2 Tensions between the Interaction Dimensions during the Design Process

As they are built from a shared prototype, both the audio-first and the visual-first instruments had pre-defined IMe (drawing) and IMo (magnet-coil interaction), in order to create a clear direction that they could each explore. The IMe and/or IMo then had an influence on the IMM, but also the UoS (as the IMe impacts how a performer moves around the space, see Section 3.3).

It is therefore both the IMM and the UoS that have made the variations between the prototypes more apparent. It can be seen on Table 2, and especially on Figure 5, where the IMM and UoS were frequently changed in each instrument's timelines. It is only in the audio-first case that the IMo also evolved over time, as it was used to add playing techniques, which was not necessary in the visual-first instrument.

We therefore believe that the dimensions of the framework should be considered together, through their tensions and conflicts, when designing an XRAVI.

5.3 Consider both Visual and Audio Transportation

Contrary to non-immersive audio-visual instruments, XRAVIs enable immersing both performers and audience members in the produced visual and sonic content. Instruments from the literature tend to restrict these dimensions to lower values, which could be due to technical constraints, aesthetic choices or a limited exploration of immersive content manipulation.

In our case study, the visual-first instrument had a stronger (*i.e.,* fully virtual) PVT than the audio-first instrument. This can be explained by stronger possibilities of manipulations of the visual environment in a virtual reality setting on one side, and by the need to preserve the perception of physical gestures on the other side.

Transportation also constitutes an issue for the audience which will require further research, either through dedicated, or shared devices [5]. This is especially the case with XRAVI that rely on headsets. Using our framework, one could imagine exploring instruments where the UoS corresponds to different scales to view a virtual content: *e.g.,* the performers could interact with a miniature audio-visual scene (with low PVT and PST), which appears at a much larger scale for the audience (with large AVT and AST), surrounding them.

5.4 Envisioned Long-term Application of the Framework

The audio-first instrument has now reached a playable state, from the author's point of view. The framework was useful in making sure both visual and sonic aspects were equally rich. The main issue for performing with the instrument will be to ensure a sufficient diversity of audio-visual content to enable long structured improvisations. While sounds, and eventually audio effects, can easily be added, in the current form of the instrument, the produced visual content will not be versatile. This will require more investigation of the IMM and UoS dimensions, for example with visual mapping presets or spatial zones.

The visual-first instrument has built a reasonable amount of visual content to experience, although there can always be more possibilities (visual and sonic) to be added. The framework has helped to raise concerns on the audience experience, and how to include them in a setting that is, equipment-wise, more fitted to a single-user. This could take the form, for the AST, of a spatialised sound setup. As for the AVT, the virtual content could affect the physical space without explicitly showing the virtual elements, thus creating a diverging experience for the performer and the audience, and bypassing the issue linked to the headset's viewpoint.

Finally, although the framework was validated through the analysis of existing instruments and of our own design process, we expect that in the long run, the dimensions might need to be revised, or even added/removed, depending on evolving practices and on changes in technologies.

6 Conclusion

In this paper, we defined the concept of Extended Reality Audio-Visual Instruments (XRAVI), and proposed a framework for their analysis and design, based on previous literature.

Following this, we presented the design of two XRAVI, diverging from the same prototype, to each focus on a different modality. This way, we were able to see how the focus on a specific modality above the other can influence the design process and the resulting instrument. We analysed the evolution of both prototypes according to our framework, provided a detailed discussion on the design choices, and proposed guidelines for the creation of XRAVI. Future work will focus on refining the instruments and the framework through long-term practice.

7 Ethical Standards

This research was conducted using public University funding for the equipment. One ethical issue comes from the use of proprietary headsets (Meta Quest), which cannot easily be maintained and for which the conditions of production are unknown. We attempted to avoid depending specifically on this hardware, through the use of open standards and protocols (OpenXR, OpenSoundControl). The rest of our implementation is done using Free and Open Source Software and will be released under a free license, including the hardware, with instructions on how to rebuild. In future versions, these instruments will also be implemented using more maintainable technologies such as projectors and cameras.

References

[1] Jerônimo Barbosa, Filipe Calegario, Veronica Teichrieb, Geber Ramalho, and Giordano Cabral. 2013. A Drawing-Based Digital Music Instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Graduate School of Culture Technology, KAIST, Daejeon, Republic of Korea, 499–502. https://doi.org/10.5281/zenodo.1178566

[2] Steve Benford, Chris Brown, Gail Reynard, and Chris Greenhalgh. 1996. Shared spaces: transportation, artificiality, and spatiality. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work* (Boston, Massachusetts, USA) *(CSCW '96)*. Association for Computing Machinery, New York, NY, USA, 77–86. https://doi.org/10.1145/240080.240196

[3] Florent Berthaut. 2020. 3D interaction techniques for musical expression. *Journal of New Music Research* 49, 1 (2020), 60–72. https://doi.org/10.1080/09298215.2019.1706584

[4] Florent Berthaut. 2025. GDPD / IVMI-BUILDER: A Libre Software Framework for Extended Reality Musical Instruments and Sonic Installations. In *Proceedings of the 19th Linux Audio Conference*. Lyon, France. https://hal.science/hal-05095979

[5] Florent Berthaut, Diego Martinez Plasencia, Martin Hachet, and Sriram Subramanian. 2015. Reflets: Combining and Revealing Spaces for Musical Performances. In *New Interfaces for Musical Expression (NIME)*. Baton Rouge, United States. https://inria.hal.science/hal-01136857

[6] Claude Cadoz and Marcelo Mortensen Wanderley. 2000. Gesture-music. *Trends in gestural control of music* (2000).

[7] Ernest Edmonds, Andrew Martin, and Sandra Pauletto. 2004. Audio-visual interfaces in digital art. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology* (Singapore) *(ACE '04)*. Association for Computing Machinery, New York, NY, USA, 331–336. https://doi.org/10.1145/1067343.1067392

[8] José Miguel Fernandez, Thomas Köppel, Nina Verstraete, Grégoire Lorieux, Alexander Vert, and Philippe Spiesser. 2017. GeKiPe, a gesture-based interface for audiovisual performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Aalborg University Copenhagen, Copenhagen, Denmark, 450–455. https://doi.org/10.5281/zenodo.1176312

[9] Esther Gruy and Florent Berthaut. 2024. MagneTip: Reintroducing a Physical Interaction Loop for 3D Musical Drawing in Extended Reality. In *New Interfaces for Musical Expression*. Utrecht (Netherland), Netherlands. https://hal.science/hal-04649511

[10] Esther Gruy and Florent Berthaut. 2024. Musical Drawing in 2D and 3D: Dimensions and Perspectives. In *Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures* (Milan, Italy) *(AM '24)*. Association for Computing Machinery, New York, NY, USA, 181–188. https://doi.org/10.1145/3678299.3678317

[11] Yoon Chung Han and Byeong jun Han. 2012. Virtual Pottery: An Interactive Audio-Visual Installation. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. University of Michigan, Ann Arbor, Michigan. https://doi.org/10.5281/zenodo.1178273

[12] Yuma Ikawa and Akihiro Matsuura. 2020. Playful Audio-Visual Interaction with Spheroids. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Romain Michon and Franziska Schroeder (Eds.). Birmingham City University, Birmingham, UK, 188–189. https://doi.org/10.5281/zenodo.4813311

[13] Alon A Ilsar, Matthew Hughes, and Andrew Johnston. 2020. NIME or Mime: A Sound-First Approach to Developing an Audio-Visual Gestural Instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Romain Michon and Franziska Schroeder (Eds.). Birmingham City University, Birmingham, UK, 315–320. https://doi.org/10.5281/zenodo.4813383

[14] Andrew Johnston. 2013. Fluid Simulation as Full Body Audio-Visual Instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Graduate School of Culture Technology, KAIST, Daejeon, Republic of Korea, 132–135. https://doi.org/10.5281/zenodo.1178572

[15] Sergi Jordà, Günter Geiger, Marcos Alonso, and Martin Kaltenbrunner. 2007. The reacTable: exploring the synergy between live music performance and tabletop tangible interfaces. In *Proceedings of the 1st International Conference on Tangible and Embedded Interaction* (Baton Rouge, Louisiana) *(TEI '07)*. Association for Computing Machinery, New York, NY, USA, 139–146. https://doi.org/10.1145/1226969.1226998

[16] Pelin Kiliboz and Cumhur Erkut. 2024. Multimodal Looper: A Live-Looping System for Gestural and Audio-visual Improvisation. In *Proceedings of the 9th International Conference on Movement and Computing* (Utrecht, Netherlands) *(MOCO '24)*. Association for Computing Machinery, New York, NY, USA, Article 8, 8 pages. https://doi.org/10.1145/3658852.3659068

[17] André Knörig, Boris Müller, and Reto Wettach. 2007. Articulated Paint : Musical Expression for Non-Musicians. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. New York City, NY, United States, 384–385. https://doi.org/10.5281/zenodo.1177155

[18] Myungin Lee. 2021. Entangled: A Multi-Modal, Multi-User Interactive Instrument in Virtual 3D Space Using the Smartphone for Gesture Control. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Shanghai, China, Article 37. https://doi.org/10.21428/92fbeb44.eae7c23f

[19] Sang Won Lee, Jungho Bang, and Georg Essl. 2017. Live Coding YouTube: Organizing Streaming Media for an Audiovisual Performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Aalborg University Copenhagen, Copenhagen, Denmark, 261–266. https://doi.org/10.5281/zenodo.1176242

[20] Sang Won Lee, Georg Essl, and Mari Martinez. 2016. Live Writing : Writing as a Real-time Audiovisual Performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Queensland Conservatorium Griffith University, Brisbane, Australia, 212–217. https://doi.org/10.5281/zenodo.1176060

[21] Golan Levin. 2000. *Painterly interfaces for audiovisual performance*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[22] Golan Levin and Zachary Lieberman. 2005. Sounds from Shapes: Audiovisual Performance with Hand Silhouette Contours in The Manual Input Sessions. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Vancouver, BC, Canada, 115–120. https://doi.org/10.5281/zenodo.1176772

[23] Ryan Mcgee, Yuan-Yi Fan, and Reza Ali. 2011. BioRhythm : a Biologically-inspired Audio-Visual Installation. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Oslo, Norway, 80–83. https://doi.org/10.5281/zenodo.1178105

[24] Masuko Miyahara and Akiko Fukao. 2022. Exploring the use of collaborative autoethnography as a tool for facilitating the development of researcher reflexivity. *System* 105 (2022), 102751.

[25] Niall Moody, Nick Fells, and Nicholas Bailey. 2007. Ashitaka : An Audiovisual Instrument. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. New York City, NY, United States, 148–153. https://doi.org/10.5281/zenodo.1177199

[26] Axel GE Mulder, Sidney S Fels, and Kenji Mase. 1999. Design of virtual 3D instruments for musical interaction. In *Graphics Interface*, Vol. 99. 76–83.

[27] Ryu Nakagawa and Shotaro Hirata. 2017. AEVE: An Audiovisual Experience Using VRHMD and EEG. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Aalborg University Copenhagen, Copenhagen, Denmark, 497–498. https://doi.org/10.5281/zenodo.1176336

[28] Ryu Nakagawa, Ryo Komatsubara, Taku Ota, and Takahisa Mitsumori. 2018. Drawing Sound in MR Space: A Multi-User Audiovisual Experience in Mixed Reality Space. In *Proceedings of the Virtual Reality International Conference-Laval Virtual*. 1–4.

[29] Ryu Nakagawa, Ryo Komatsubara, Taku Ota, and Hidefumi Ohmura. 2018. Air Maestros: A Multi-User Audiovisual Experience Using MR. In *Proceedings of the 2018 ACM Symposium on Spatial User Interaction* (Berlin, Germany) *(SUI '18)*. Association for Computing Machinery, New York, NY, USA, 168. https://doi.org/10.1145/3267782.3274685

[30] Ryu Nakagawa, Hidefumi Ohmura, Kenta Hidaka, Sho Kato, and Ai Kawaguchi. 2025. *Drawing Lines to Connect: A Multi-Participant XR Audiovisual Art Experience Using Internet-Connected Video See-Through MR with EEG/PPG*. Association for Computing Machinery, New York, NY, USA.

[31] Ireti Olowe, Giulio Moro, and Mathieu Barthet. 2016. residUUm: user mapping and performance strategies for multilayered live audiovisual generation. In

Proceedings of the International Conference on New Interfaces for Musical Expression. Queensland Conservatorium Griffith University, Brisbane, Australia, 271–276. https://doi.org/10.5281/zenodo.1176098

[32] Gorkem Ozdemir, Anil Camci, and Angus Forbes. 2016. PORTAL: An Audiovisual Laser Performance System. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Queensland Conservatorium Griffith University, Brisbane, Australia, 338–343. https://doi.org/10.5281/zenodo.1176102

[33] Kenneth Peacock. 1988. Instruments to perform color-music: Two centuries of technological experimentation. *Leonardo* 21, 4 (1988), 397–406.

[34] Sourya Sen, Koray Tahiroğlu, and Julia Lohmann. 2020. Sounding Brush: A Tablet based Musical Instrument for Drawing and Mark Making. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Romain Michon and Franziska Schroeder (Eds.). Birmingham City University, Birmingham, UK, 331–336. https://doi.org/10.5281/zenodo.4813398

[35] Stefania Serafin, Cumhur Erkut, Juraj Kojs, Niels C Nilsson, and Rolf Nordahl. 2016. Virtual reality musical instruments: State of the art, design principles, and future directions. *Computer Music Journal* 40, 3 (2016), 22–40.

[36] Fabián Sguiglia, Pauli Coton, and Fernando Toth. 2019. El mapa no es el territorio: Sensor mapping for audiovisual performances. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Marcelo Queiroz and Anna Xambó Sedó (Eds.). UFRGS, Porto Alegre, Brazil, 146–149. https://doi.org/10.5281/zenodo.3672902

[37] Sam Trolland, Alon Ilsar, and Jon McCormack. 2025. Visually-Led Design for Gestural Audiovisual Instruments. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Doga Cavdir and Florent Berthaut (Eds.). Canberra, Australia, Article 45, 9 pages. https://doi.org/10.5281/zenodo.15699633

[38] Luca Turchet, Rob Hamilton, and Anil Çamci. 2021. Music in extended realities. *IEEE Access* 9 (2021), 15810–15832.

[39] Rafael Valer, Rodrigo Schramm, and Luciana Nedel. 2020. Musical brush: Exploring creativity in an ar-based tool combining music and drawing generation. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 634–635.

[40] Anna Weisling, Anna Xambó, ireti olowe, and Mathieu Barthet. 2018. Surveying the Compositional and Performance Practices of Audiovisual Practitioners. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Thomas Martin Luke Dahl, Douglas Bowman (Ed.). Virginia Tech, Blacksburg, Virginia, USA, 344–345. https://doi.org/10.5281/zenodo.1302609

[41] Mark Zadel and Gary Scavone. 2006. Different Strokes: a Prototype Software System for Laptop Performance and Improvisation. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Paris, France, 168–171. https://doi.org/10.5281/zenodo.1177025

[42] Karitta Christina Zellerbach and Charlie Roberts. 2022. A Framework for the Design and Analysis of Mixed Reality Musical Instruments. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Andrew McPherson and Emma Frid (Eds.). Auckland, New Zealand, Article 29. https://doi.org/10.21428/92fbeb44.b2a44bc9