Layered Motion and Gesture Sonification in an Interactive Installation

Grigore Burloiu, Valentin Mihai, Stefan Damian

▶ **To cite this version:**

Grigore Burloiu, Valentin Mihai, Stefan Damian. Layered Motion and Gesture Sonification in an Interactive Installation. Journal of the Audio Engineering Society, 2018, 66 (10), pp.770-778. 10.17743/jaes.2018.0047 . hal-03015480

HAL Id: hal-03015480
https://hal.science/hal-03015480v1

Submitted on 19 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

G. Burloiu, V. Mihai and Ş. Damian, "Layered Motion and Gesture Sonification in an Interactive Installation"
J. Audio Eng. Soc., vol. 66, no. 10, pp. 1–10, (2018 October.).
DOI: https://doi.org/10.17743/jaes.2018.0047

Layered Motion and Gesture Sonification in an Interactive Installation

GRIGORE BURLOIU[1], VALENTIN MIHAI[2], AND ŞTEFAN DAMIAN[1]
(grigore.burloiu@unatc.ro) (valentin.mihai.syl@gmail.com) (stefandamian.sd@gmail.com)

[1] *CINETic, UNATC "I.L. Caragiale" Bucharest, Romania*
[2] *University "Politehnica" Bucharest, Romania*

SoundThimble is an interactive sound installation based on the relationship between human motion and virtual objects in 3D space. A Vicon infrared motion capture system and custom software are used to track, interpret, and sonify the movement and gestures of a performer relative to a virtual object. We define three possible interaction dynamics, centered around object search, manipulation, and arrangement. We explore the resulting possibilities for layered sonification dynamics and extended perception and expression in internal tests as well as a public demonstration. Experimental evaluation reveals an average object search time of around 60 s, as well as thresholding ranges for effective gesture spotting. The underlying software platform is open source and portable to similar hardware systems, leaving room for extension and variation.

0 INTRODUCTION

High resolution three-dimensional motion tracking is traditionally used for animation in film and games as well as in life sciences research and engineering applications [1]. This technology has long been utilized by the audio computing community [2, 3], but its applications generally remain limited to standard paradigms of isolated body motion audification [3, 4] or sound control interfaces [2, 5, 6].

The *SoundThimble* project harnesses current motion capture technology and gesture detection algorithms to enable new modes of real-time sound exploration and transformation, coupled with layered interaction scenarios. The result is a framework for interactive sound installations, dynamic composition, and augmented dance performance.

This paper presents the pilot application of the *SoundThimble* framework as an installation of the same name. Audience members entering the tracking area shift between the roles of game player, sonic performer, and composer/arranger, following an iterative interaction schema. The central vehicle in all three layers is the "sound-thimble" itself—a virtual object with particular spatial, sonic, and interaction attributes.

Our implementation uses a state-of-the-art Vicon motion capture system[1] containing eight Vantage 5-megapixel infrared cameras and two Bonita video cameras. Since the open-source software developed in this project[2] is built around Vicon's Datastream SDK,[3] the framework can be ported to both older and future Vicon-based systems.

In the remainder of the paper we review relevant literature and technology (Sec. 1), we describe the *SoundThimble* installation (Secs. 2, 3), we evaluate several aspects of the system (Sec. 4), and finish with a survey of future challenges and perspectives (Sec. 5).

1 STATE OF THE ART

Infrared motion capture (mocap) systems have been revealed as a precise and robust means for expressive interaction in a controlled environment, with many features being translatable to more portable technologies [7, 8]. This makes mocap a rich tool for the research and instantiation of new sonic interaction paradigms. In particular, Vicon motion capture systems have been used for over a decade for music applications [2, 3, 5, 8, 9]. A software bridge for streaming OSC data from a Vicon source was developed [5] as part of a concluded project;[4] it unfortunately proved to be incompatible with our current system.

Gesture is a main focal point in music technology research [10]. *EGGS* [4] is an early real-time gesture sonification system, which has been employed in various interactive installations and performances including enhanced

[1] See https://www.vicon.com.

[2] Available at https://github.com/RVirmoors/viconOSC.

[3] See https://www.vicon.com/products/software/datastream-sdk.

[4] See http://sonenvir.at/downloads/qvicon2osc/.

circus [11]. Fohl et al. developed a control system for virtual sound source spatialization [6]. As with *EGGS*, it relies on a set of elementary gestures mapped to specific sonic actions. The *MLWorkbench* software [12] provides interactive access to machine learning parameters for generative use, moving the focus from the algorithms' output (classification, position, etc.) to its internal parameters.[5] The *MaD* system [14] (later integrated into *MuBu* [15]) introduced on-the-fly motion-sound mapping, enabling user-defined gestures to flexibly navigate through specific sounds via concatenative synthesis. Finally, the *3DinMotion* system [9] is an example of a modern (albeit closed-source) integrated motion sonification and visualization system, supporting several mocap platforms, including Vicon.

These and other recent qualitative advances in the interaction between human gesture and sound behavior have been made possible by real-time gesture recognition and the following tools [14, 12, 16–18]. *Gesture spotting* is the automatic detection of gestures from a continuous stream of motion, also separating anticipated gestures from unrelated movement. This on-the-fly classification can be achieved through methods based on probabilistic/HMM-based models [19, 18] or artificial neural networks [20], with particular interest given to the cost of triggering the classification early [21].

Most existing gesture-based interactive sound installations center around a particular activity. A typical example is *Grainstick* [22], where two controllers manipulate a spatialized virtual "sound tube." However, more complex scenarios are technically possible, where gesture is used both for direct sonification and for multi-level control of system behavior—features that our project channels into a coherent, open-source framework.

2 CONCEPT

The "sound-thimble," as the basic building block of our framework, extends the concept of *sound object* in the Schaefferian sense, as a clearly delimited sounding unit, open to manipulation, arrangement, and composition [23].

Such an entity, once instantiated, can retain an ambiguous nature (spatially and acousmatically) or can switch to a more material state (positioned in space and tied to a causal source) [24]. The duality between the latent positioning of the object (which can be inferred from phenomena other than spatial sound reproduction), and the active sound spatialization and transformation, enables a variety of sonic art practices, such as sound sketching, auditory games, and other real-time interactions.

Motion capture technology is integral to this concept, due to the ability to define and employ absolute coordinates in 3D space. For alternatives such as wearable sensors or augmented instruments, only relative measures are directly available and a stable anchoring to an independent point or trajectory in space is hard to impossible to accomplish.

[5] An early iteration of this idea can be found in [13].

2.1 Interaction Scenario

Our installation instantiates the idea above in three phases: search, manipulation, arrangement—related to schemas of play, performance, and composition, respectively. The reasoning for this design is manifold. A layered experience is apt to appeal to a wider audience (from the general expo/museum visitor, to professionals in sound or dance) on several levels: exploring basic mechanics, achieving game goals, expressing sonic and kinetic creativity. The more levels the visitor actively engages with, the more listening modes [25] they traverse, and the richer and more lasting their involvement is likely to be.

The experience begins as an immersive game, with a human player attempting to find a sound-thimble (stationary and randomly positioned in 3D space), by analyzing cues that are constantly shifting in the sonic fabric based on the hand's movement relative to the object. Analogously to the traditional game of *Hunt the Thimble* (a.k.a *Hot or Cold*), the space between the human and the virtual object is correlated to dynamic sound generation parameters, guiding the player's hand towards its target.

Once the object is found it attaches to the hand, and its sonic manifestation gains a richer causal relationship: the player becomes a performer and is now able to explore the object's sonic palette and define a number of gestures that can be re-performed later, re-called, and used to trigger or manipulate sonic shifts and events.

Finally, the user can position the virtual object on the floor or discard it by "pushing" it outside of the installation boundaries. This triggers a new object to be randomly generated, while the player retains a degree of control over the initial object by recalling recorded gestures. Both objects are now in a latent state, with the new one guiding the player's search and the previous one responding to the learned set of gestures.

This recursive scenario is outlined in Fig. 1: objects are randomly generated, the performer finds them, defines gestures, and interacts sonically before arranging them in a desired configuration. The different nature of each phase is designed to provide a sense of structure, resulting in a layered and varied user experience.

2.2 Performance Aesthetic

The human-object dynamic central to the *SoundThimble* framework results in certain interaction features that circumscribe the aesthetics of the installation.

Our approach is informed by Worrall's study [25], which reveals a necessity for the mapping of minute gestural inflections to alter sonic material with a view to certain modes of listening. Our installation aims to tackle these modes: reflexive, kinaesthetic, connotative, empathetic, semantic, reduced. Their sound design correspondences are laid out in Sec. 3.4.

The cross-modality between different kinds of sense perception guides the performer's attention to the various sonic responses to physical actions. The multi-modal information is processed in real time, continuously redefining the affordances enabled by the system.

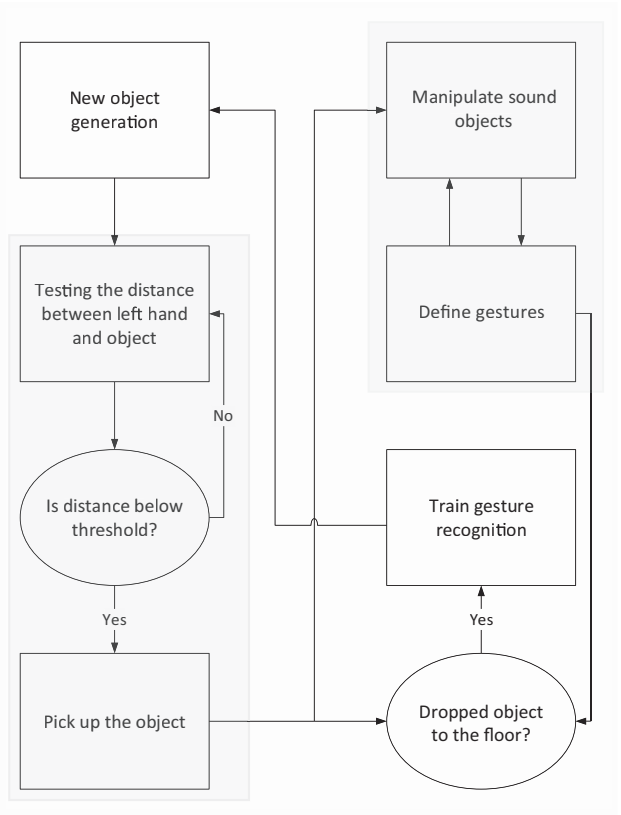


Fig. 1. *SoundThimble* interaction flow. Left highlight: search phase. Right highlight: manipulation phase. Overall: arrangement phase.

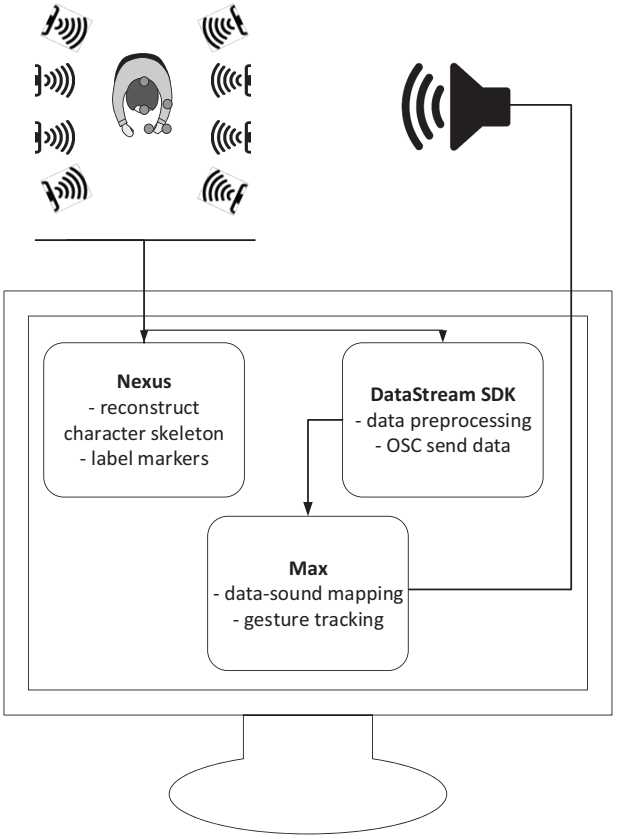


Fig. 2. *SoundThimble* framework architecture.

The sonic interaction occurs via two complementary paradigms: in the search phase, the sound acts as the encompassing guide to a performer's movements; in the manipulation phase, the performer has agency over the sound and its parameters. Finally, the arrangement phase integrates the two. We further delineate the space of kinetic and sonic affordances opened by our installation in Sec. 4.

3 IMPLEMENTATION

The framework architecture diagram is laid out in Fig. 2. Three-dimensional sensor data is streamed into the Vicon Nexus software, which is able to reconstruct and label the underlying character skeleton. The gesture recognition and sonification algorithms are programmed in Max[6], which receives control data via the OSC[7] protocol. Since Vicon systems do not support OSC out of the box, we used the *oscpack*[8] library to extend the DataStream C++ SDK and send OSC bundles to Max. A demonstration video of this implementation pipeline is available online[9].

3.1 Character Design

We pursued a minimal amount of markers for ease of setup and prototyping. The resulting configuration—sufficient for tracking hand gestures, while ensuring redundancy in case a marker is obscured from the cameras—consists of five markers: two positioned on the head, one on the forearm, and two on the hand (thumb and index finger). These are processed and sent via OSC bundles as two 3D coordinates: the head and the hand.[10]

To minimize data loss we set the system frame rate f by taking into account the maximum movement speed v_{max} and the minimum spacing between markers d_{min} [26].

$$f > \frac{1}{K}\frac{v_{max}}{d_{min}} \tag{1}$$

For the constant value $K = 0.41$ as determined in [26], and setting d_{min} to 40 mm (minimum hand to head distance[11]) and v_{max} to 1350 mm/s (the maximum speed of human hand movement [27]), we obtain a minimum value of $f = 82.3$. This means that a frame rate of 100 fps is more than sufficient for our purposes.[12] This configuration produces highly stable and responsive inputs into the Max system

[6]Max is a state-of-the-art programming environment for real-time multimedia: http://cycling74.com/.

[7]OpenSoundControl is a multimedia communication protocol: http://opensoundcontrol.org/.

[8]See http://www.rossbencina.com/code/oscpack.

[9]See https://youtu.be/K2Xni2lWswg.

[10]The forearm marker only serves for skeleton reconstruction.

[11]The two hand markers contribute to a single coordinate, so we do not require permanent labeling accuracy between them.

[12]The major variable here is d_{min}: in a different configuration (e.g. if two hands are clapping), smaller inter-marker distances would result in a larger f values.

with a spatial resolution of 1 mm and an end-to-end latency of 13 ms.

3.2 Object Generation and Interaction Mechanics

The following object-related mechanics are implemented as basic algorithms: generation, detection, release.

Objects are generated at random positions within the boundaries of the motion capture field. Detection of the sound-thimble occurs when the distance between hand and object falls below a set threshold. By default this threshold is set at 120 mm radial distance; lowering it can make the game considerably more difficult. Once the object is detected, it becomes mobile, its coordinates tracking those of the hand's.

Finally, a simple thresholding of the z-axis (height) value of the hand position serves to release the object. At this point, a new object is generated. The system keeps track of all object coordinates, as they appear over time.

3.3 Gesture Recognition

We use the thumb-index finger distance value to enable gesture recording while the two fingers are kept close together. In the initial test runs it quickly became evident that raw Cartesian coordinates perform poorly when the user changes position and orientation. Thus, the input features captured into *MuBu* multi-buffer containers [15] consist of cylindrical triplets:

- $\Delta\theta/\Delta t$, with $\theta = \tan^{-1}(\frac{y}{x})$;
- $r = \sqrt{x^2 + y^2}$;
- z,

where x, y, z are the respective differences between head and hand Cartesian coordinates, as received via OSC. This feature preprocessing serves two purposes:

First, gestures are recorded based on the position of the hand relative to the head, thus becoming invariable to the performer's absolute *position* within the space. Second, by considering the variation of angle θ over time (as opposed to its absolute value), gestures become invariable to the performer's *orientation* on the horizontal plane. Thus, gestures can be recorded and recalled anywhere within the space, irrespective of the direction the performer is facing.

The motion data is fed into *MuBu*'s Hiearchical Hiden Markov Models (HHMM) continuous classifier module [18], which encodes each recording into a sequence of 10 states with corresponding transition probabilities. We leverage the continuous class and state likelihood data to perform gesture spotting [19], by setting two control parameters:

- **Classification threshold**: when one class is at maximum likelihood (positive value), the second-most likely class should be *below* this threshold;
- **Hold time**: a minimal number of consecutive states, in monotonous progression.

These two conditions ensure a degree of *certainty* of one class over the others, and *consistency* of the choice over several consecutive states, for the corresponding gesture to be marked as active. Each parameter can be adjusted, depending on the error tolerance and responsiveness desired from the system, starting from a default setting we reached through the evaluation in Sec. 4.2.

When an object is released to the floor, the classifier is (re)trained with the newly recorded gesture data, and consequently gestures can act as activators for a particular sonic behavior (if they are performed/spotted one at a time), or as continuous controllers (if they are repeated by the performer). When several objects exist in the space, a specific gesture might act on one or more objects depending on their spatial relationship (angle, proximity) to the head-hand segment.

3.4 Sound Design

In their conceptual study [7], Skogstad et al. distinguish four strategies for IR mocap-based gesture sonification: modeling sound-producing actions (excitation and modification), touchless actions ("in the air"), feature mapping, and spatialization. While the first comes closest to simulating actual instruments, we mostly employ the remaining three to provide an immersive experience, making direct use of the active space.

Each sound-thimble has a corresponding sound design Max patch, differing in (a) the source sound material used, (b) the synthesis techniques applied, and/or (c) the control mapping schema to the object search and manipulation variables. The various combinations of (a), (b), and (c) give rise to a growing library of objects, each with its own character. By varying the interaction rules for each object, we link them to spatially aware parameters adding up to a continuously evolving, organic soundscape. An octophonic ring of speakers is used for monaural diffusion in the search phase and for conveying directionality and spatialization in the manipulation and arrangement phases using vector based amplitude panning (VBAP).

As with [13, 5, 3] et al., the main sound generation engine relies on granular synthesis, a technique that reorganizes short grains of acoustic material towards shaping new sounds. The source audio files consist of environmental recordings and electronic soundscapes that were pre-edited in order to gradually shift their intensity (amplitude, spectral content, and micro-events within the sound) from beginning to end. This way, the granular engine is able to focus on specific audio characteristics, depending on which area of the source material is being explored. Other techniques including additive and wavetable synthesis were considered, however granular synthesis proved to be more flexible in terms of mapping options and resulting soundscapes.

For the search phase, divergent mapping [28] is used to sonify the hand-thimble distance through the parameters of grain position, size, density, and of envelope shape. This translates in a sustained low-amplitude, darker sound with a larger grain duration and density when the distance is large, transitioning into more of an attack-decay type sound [29],

brighter in nature, with shorter grain duration and sparser density as the distance is decreased. The resulting sound is decoded monaurally on all speakers: only timbre and dynamics are to guide the search process, not any sound spatialization.

After the thimble is found, during the manipulation phase, new rules of interaction apply, suggesting an increased affordance range [30]. First, the main input to the granular engine, previously represented by the hand-thimble distance is linked to a primary dimension such as the hand position on the height axis, with the original mappings scaled accordingly. Second, gesture recognition is used to toggle different types of processing states for the active thimble, such as reverb, echo, various modulations, and spatial manipulations. Each sound-thimble has a predetermined collection of such states that transform the sonic character and every newly defined gesture is attributed randomly to one set of such events. These sonic actions are designed in correspondence to the affordances enabled by each thimble source sound. Finally, the amount of head-hand distance acceleration allows the user to augment the active sound through enveloping, creating rhythmic patterns and structures. These patterns persist when the object is placed on the floor.

Transient events (finding the object or dropping it) also trigger a short audio signal consisting in an enveloped burst of random grains from the active thimble's material.

In the arrangement phase, the position of a dropped thimble originates a cylindrical "hotspot," in the form of a virtual pillar reaching from the floor to the top of the installation, which alters the sonic content via hand position triggering. While the first two phases referred only to the active thimble, the arrangement phase is perpetual. Even when none of the hotspots are active, remnants of the original sounds are always audible, fading in intensity and density as new sound-thimbles are created.

The emerging soundscape consists of a sonic background of existing, aging sound objects and a foreground determined by the active thimble. From a spatial perspective, background components are distributed over the eight speakers in accordance to their positioning on the floor. Decorrelation-based techniques [31] such as chorusing and filtering are used to increase the spatial spread when old thimbles are more centrally stationed and diffused almost equally on all speakers. As for the foreground, in the manipulation phase the sound is panned following the user's hand position, with a subtle spatial echo to amplify the effect.

Referring back to Sec. 2.2, our design aims to elicit the following modes of listening: (a) **reflexive** as a new sound appears, into **connotative** as the user learns to approach the thimble; (b) **semantic** for the auditory icon of grabbing or dropping the thimble; (c) **connotative** into **reduced**, as the user explores manipulation affordances and focuses in on sonic qualities; and (d) **reduced** into **empathetic** as the user arranges the sounds into a pleasing or otherwise emotionally significant setup. Naturally, all the preceding should coexist with the **kinaesthetic** sense of motor-movement sustained by the constant link between motion and sound.

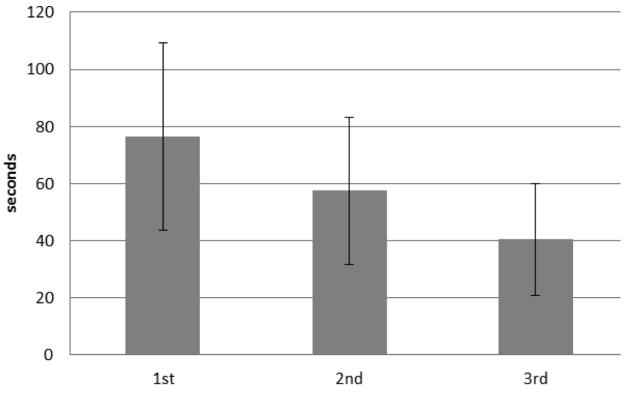


Fig. 3. Average search time per trial. As users gain experience, finding the object is consistently quicker.

4 EVALUATION

To assess the implementation of our design, we conducted a set of three tests, each focusing around a key aspect of the *SoundThimble* installation. Fourteen participants (5 female) aged between 26 and 34, without prior exposure to the system, 8 with a music, and 3 with a performing arts background, were asked to test the sound object search, the gesture definition and recall, and the ensuing sonic manipulation.

4.1 Search Time

In order to validate the interaction and sound design choices for the search phase, we started by studying how fast first-time users manage to find the sound-thimble and how the search time changes over subsequent trials.

We conducted 42 trials overall, giving each participant 3 chances to find the thimble. Fig. 3 shows the average times for each attempt, decreasing from the first to the third, around a global average of 59.14 s, with a sample standard deviation of 29.72 s.

Considering this test, one must bear in mind the many possible sources of variability. The first is the subjects' period of accommodation to the environment, which is revealed by the decreasing variance from one attempt to the next. The pairing of source sound (which we drew randomly among four non-repeating options) and listener influences the ease of search. Finally, there were outliers where the thimble was hit upon before the user started actively searching for it. Such a case was deemed irrelevant to measuring the effectiveness of search action, so we discarded the corresponding data point and generated a fresh trial.

Certainly there is room for a more detailed evaluation that takes into account these control variables, as well as the actual installation scenario, wherein the second search attempt happens with the first object still manifesting itself in the space, interfering with the search signal and thus compensating (by design) for the user's adaptation to the system mechanics. Still, the current experiment offers a quantified baseline that validates the search phase design.

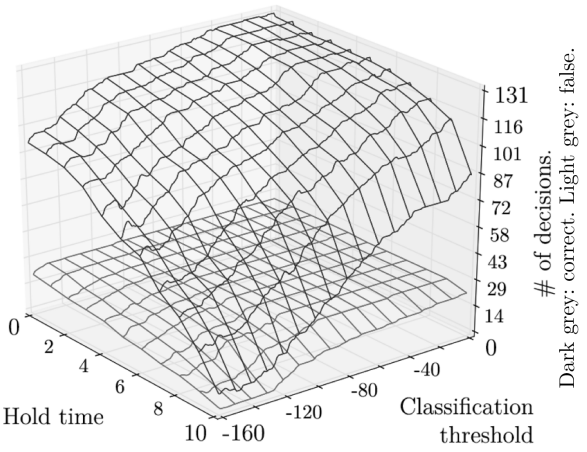


Fig. 4. Impact of classification threshold and hold time on the number of triggered classifications. The lower the hold time and higher the threshold, the more triggered classifications and more false positives. Dark grey: Correct decisions (max: 131, from 158 trials). Light grey: False positives (max: 25).

4.2 Gesture Spotting

The second test aimed to quantify the performance of the architecture is described in Sec. 3.3. Subjects were asked to record three separate gestures, each representing a class to be later recognized. After training the *MuBu* HHMM-based classifier, the participants were to record themselves[13] performing the gestures in an arbitrary order, at varying orientation and spatial or temporal scale.

For each such trial, we recorded a stream of classification likelihoods and state progressions. As a benchmark, we labeled each trial by an optimistic empirical analysis of whether a best-case algorithm would be able to correctly classify the gesture from the recorded likelihood data. This empirical approximation resulted in 131 positive tags out of a total of 158 trials, or a 83% maximal correct classification rate. We observed two general causes of mislabeling, or inconclusive likelihood outputs. One is when the trained recordings were too short or too similar between each other. The second is for inexact or reversed gesture execution, e.g., someone recording a circle clockwise, and then performing it counter-clockwise.

Fig. 4 shows the result of passing the likelihood and state progression data through the algorithm described in Sec. 3.3 for a range of classification threshold and hold time values. The hold time is sampled from zero (instantaneous) to 10 (the whole gesture) states. The classification threshold goes down to –160, corresponding to very high certainty. Since these ranges are shown to produce the entire gamut of activation numbers, we make them available in the installation's settings interface.

As a default, we set the hold time to 3 states and the classification threshold to –40, which amounts to 113 total decisions, of which 99 are correct (75% of the maximal 131) and 14 are false positives (12% of the decisions triggered).

[13] by pinching the thumb and index finger, similarly to the initial gesture recording, as described in Section 3.3.

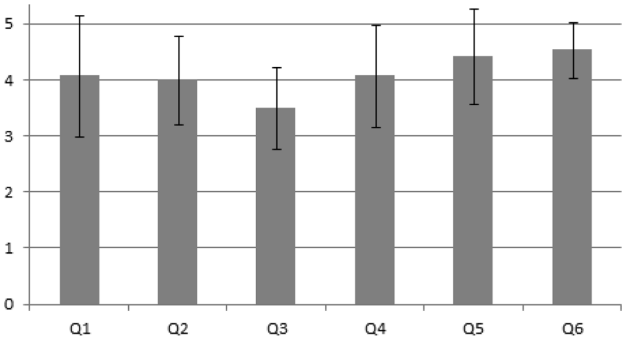


Fig. 5. Answers to the questionnaire in Sec. 4.3 (1 = poor to 5 = very good).

At this setting, subsequent experience reveals a general satisfaction with gestural responsiveness, with the occasional misfire not severely impacting the experience. Note that the heuristic maximal number of 131 correct classifications is never reached, even when setting both parameters to zero, which results in 125 correct triggers (95%) and 29 false positives (19%).

4.3 Sonic Interaction

The last step consisted in a subjective evaluation after the users had freely explored the installation for 10–15 minutes. Participants were asked to rate six aspects of their experience on a five-point Likert scale ranging from "poor" to "very good." The following questions were asked:

Q1: How easy was it to find the thimble?
Q2: How engaging are the interaction mechanics?
Q3: How expressive is the thimbles' sonic character?
Q4: How clear is the range of possible sounds?
Q5: How evident is the game's structure?
Q6: What is your general evaluation of the platform?

As the results in Fig. 5 show, reactions were generally positive. We attribute the slightly lower marks for Q3 to the incipient state of the sound design, as presented in our previous paper [32]; in fact, several features from Sec. 3.4 were added following the feedback we received over these sessions.

The questionnaire concluded with a section for free text comments and suggestions. We received several comments praising the search phase for easing them into the installation mechanics and lending the sound a physical character. Subjects enjoyed the alternation between interaction paradigms, and several users reported slipping from one listening mode to another, as we indicated in Sec. 2.2. The great majority found the interaction intuitive and engaging, mostly treating it as sound exploration rather than music. We counted several ideas for increasing sonic variation (melodic and textural) and interaction complexity. There were also a number of requests for expanding the character skeleton and the installation space; one person suggested interconnecting several physical rooms.

Fig. 6. Interacting with *SoundThimble* at the "Theatre meetings in Sulina" conference.

5 DISCUSSION AND FUTURE WORK

This paper introduced *SoundThimble*, a three-phase interactive installation based on a new framework for real-time motion-music interaction. All the developed software (including the C++ code for data preprocessing and transmission, and the Max patches for gesture tracking and sound design) is open source and publicly available.

Our main contribution consists in the aesthetic and immersive context provided by the installation, allowing for the makeup of state-of-the-art motion and gesture control technology to be experienced from different perspectives: (absolute—thimble search) motion guided by sound, (relative—cylindrical coordinates) motion defining gesture, gesture spotting triggering sonic actions, and sound guided by (absolute and relative) motion.

In most applications of gesture recognition, the training phase occurs apart from the performance, often using many examples of each gesture in a particular set. By incorporating both the training and execution phases into the installation, our paradigm empowers the user to build their own set of gestures, taking advantage of the *MuBu* HHMM model's ability to generalize from single examples of each class. Our experience has shown that the combination of classification likelihood (for certainty) and state progression (for responsiveness) thresholding provides a reliable system for controlling the flexibility and sensitivity of gestural interaction. Still, as the evaluation results in Sec. 4.2 confirmed, gesture spotting remains a challenging problem.

Aside from the internal testing and evaluation, the installation has been presented as a public demonstration in September 2017 at the international "Theatre meetings in Sulina" event[14] and is being prepared for public gallery presentations and other destinations[15]. An audience member in Sulina engaging with the installation is shown in Fig. 6. The verbal feedback given by participants, both for the evaluation and in Sulina, was generally encouraging.

Our granular synthesis engine has proved to be flexible enough as to be suggestive of location cues, while producing varied and interesting sounds that react well to processing, blending, and layering. In addition to the granulator's internal parameters, scale factors, transfer functions, and thresholding are equally important for designing usable sonic interaction schemas. Future plans for sound design development include experimenting with more synthesis techniques and evaluating a Wave Field Synthesis-based output system, which could prove more effective in a larger multi-user environment—being a holophonic approach, a more developed sense of direction could be achievable [22].

In terms of software development, we are working to incorporate the Vicon data acquisition and processing into a custom Max external, also adding support for different character skeletons. Eventually we aim to release a fully featured SDK and expanding to platforms such as Pure Data. We also plan to support several participants at once, implementing the shared control of a sound-thimble.

Finally, we have commenced the outreach to composers, artists, and creative programmers, to apply our framework to more innovative projects and to engage in collaborative practice-led research.

[14] See http://bit.ly/2m4anXL (in Romanian).

[15] At revision time (Feb 2018), the system is being adapted for (1) interaction with disabled children, as a means to improve attention and motor skills, and (2) the sonification of human posture changes, as feedback within physical therapy procedures.

6 ACKNOWLEDGMENT

The authors thank Bogdan Golumbeanu and Ştefan Pârlog for their contributions to developing and documenting the project.

7 REFERENCES

[1] G. Welch and E. Foxlin, "Motion Tracking: No Silver Bullet, but a Respectable Arsenal," *IEEE Computer Graphics and Applications*, vol. 22, no. 6, pp. 24–38 (2002).

[2] C. Dobrian and F. Bevilacqua, "Gestural Control of Music: Using the Vicon 8 Motion Capture system," *International Conference on New Interfaces for Musical Expression*, pp. 161–163 (2003).

[3] A. Kapur, G. Tzanetakis, N. Virji-Babul, G. Wang, and P. R. Cook, "A Framework for Sonification of Vicon Motion Capture Data," *Conference on Digital Audio Effects*, pp. 47–52 (2005).

[4] M. Goina and P. Polotti, "Elementary Gestalts for Gesture Sonification," *International Conference on New Interfaces for Musical Expression*, p. 150 (2008).

[5] G. Eckel, D. Pirro, and G. K. Sharma, "Motion-Enabled Live Electronics," presented at the *International Sound and Music Computing Conference* (2009).

[6] W. Fohl and M. Nogalski, "A Gesture Control Interface for a Wave Field Synthesis System," presented at the *International Conference on New Interfaces for Musical Expression* (2013).

[7] S. A. v. D. Skogstad, A. R. Jensenius, and K. Nymoen, "Using IR Optical Marker Based Motion Capture for Exploring Musical Interaction," presented at the *International Conference on New Interfaces for Musical Expression* (2010).

[8] G. Vigliensoni and M. M. Wanderley, "A Quantitative Comparison of Position Trackers for the Development of a Touch-less Musical Interface," presented at the

International Conference on New Interfaces for Musical Expression (2012).

[9] A. Renaud, C. Charbonnier, and S. Chagué, "3DinMotion: A Mocap Based Interface for Real Time Visualisation and Sonification of Multi-User Interactions," *International Conference on New Interfaces for Musical Expression*, pp. 495–496 (2014).

[10] A. R. Jensenius, "To Gesture or Not? An Analysis of Terminology in International Conference on New Interfaces for Musical Expression Proceedings 2001–2013," *International Conference on New Interfaces for Musical Expression*, pp. 217–220 (2014).

[11] L. Elblaus, M. Goina, M.-A. Robitaille, and R. Bresin, "Modes of Sonic Interaction in Circus: Three Proofs of Concept," presented at the *International Computer Music Conference* (2014).

[12] J. Schacher, C. Miyama, and D. Bisig, "Gestural Electronic Music Using Machine Learning as Generative Device," presented at the *International Conference on New Interfaces for Musical Expression* (2015).

[13] J. Williamson and R. Murray-Smith, "Audio Feedback for Gesture Recognition" (2002).

[14] J. Françoise, N. Schnell, and F. Bevilacqua, "MaD: Mapping by Demonstration for Continuous Sonification," presented at the *ACM SIGGRAPH 2014 Emerging Technologies*, SIGGRAPH '14 (2014).

[15] N. Schnell, A. Röbel, D. Schwarz, G. Peeters, R. Borghesi, et al., "MuBu and Friends–Assembling Tools for Content Based Real-Time Interactive Audio Processing in Max/MSP," presented at the *International Computer Music Conference* (2009).

[16] B. Caramiaux, J. Françoise, N. Schnell, and F. Bevilacqua, "Mapping through Listening," *Computer Music J.*, vol. 38, no. 3, pp. 34–48 (2014).

[17] B. Caramiaux, N. Montecchio, A. Tanaka, and F. Bevilacqua, "Adaptive Gesture Recognition with Variation Estimation for Interactive Systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 4, no. 4, p. 18 (2015).

[18] J. Françoise, N. Schnell, R. Borghese, and F. Bevilacqua, "Probabilistic Models for Designing Motion and Sound Relationships," *International Conference on New Interfaces for Musical Expression*, pp. 287–292 (2014).

[19] Y. Yin and R. Davis, "Gesture Spotting and Recognition Using Salience Detection and Concatenated Hidden Markov Models," *International Conference on Multimodal Interaction*, pp. 489–494 (2013).

[20] P. Neto, D. Pereira, J. N. Pires, and A. P. Moreira, "Real-Time and Continuous Hand Gesture Spotting: An Approach Based on Artificial Neural Networks," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 178–183 (2013).

[21] R. Tavenard and S. Malinowski, *Cost-Aware Early Classification of Time Series*, pp. 632–647 (Springer International Publishing, Cham, 2016).

[22] G. Leslie, B. Zamborlin, P. Jodlowski, and N. Schnell, "Grainstick: A Collaborative, Interactive Sound Installation," *Proceedings of the International Computer Music Conference*, p. 4 (2010).

[23] P. Schaeffer, G. Reibel, B. Ferreyra, H. Chiarucci, F. Bayle, A. Tanguy, J.-L. Ducarme, J.-F. Pontefract, and J. Schwarz, *Solfège de l'objet sonore* (INA, 1998).

[24] B. Kane, *Sound Unseen—Acousmatic Sound in Theory and Practice* (Oxford University Press, New York, 2014).

[25] D. Worrall, "Understanding the Need for Micro-Gestural Inflections in Parameter-Mapping Sonification," presented at the *International Conference on Auditory Display* (2013).

[26] M.-H. Song, and R. I. Godøy, "How Fast Is Your Body Motion? Determining a Sufficient Frame Rate for an Optical Motion Tracking System Using Passive Markers," *PloS one*, vol. 11, no. 3, p. e0150993 (2016).

[27] K. M. DeGoede, J. A. Ashton-Miller, J. M. Liao, and N. B. Alexander, "How Quickly Can Healthy Adults Move Their Hands to Intercept an Approaching Object? Age and Gender Effects," *The Journals of Gerontology: Series A*, vol. 56, no. 9, pp. M584–M588 (2001).

[28] G. Kramer, *Auditory Display: Sonification, Audification, and Auditory Interfaces* (Perseus Publishing, 1993).

[29] D. Smalley, "Spectromorphology: Explaining Sound-Shapes," *Organized Sound*, vol. 2, no. 2, pp. 107–126 (1997).

[30] A. Altavilla, B. Caramiaux, and A. Tanaka, "Towards Gestural Sonic Affordances," *International Conference on New Interfaces for Musical Expression*, pp. 61–64 (2013 May).

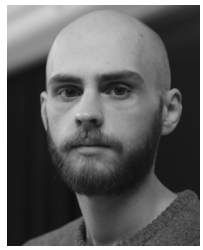
[31] G. S. Kendall, "The Decorrelation of Audio Signals and its Impact on Spatial Imagery," *Computer Music J.*, vol. 19, no. 4, pp. 71–87 (1995).

[32] G. Burloiu, S. Damian, B. Golumbeanu, and V. Mihai, "Structured Interaction in the *SoundThimble* Real-Time Gesture Sonification Framework," presented at the *Audio Mostly International Conference* (2017).

THE AUTHORS

Grigore Burloiu

Valentin Mihai



Ştefan Damian

Grigore Burloiu defended his Ph.D. thesis, *DynamicMusic Representations for Real-Time Performance*, in 2016. He is a researcher at the CINETic center in Bucharest, heading the Sound-Light Digital Interaction Lab and coordinating the newly established Interactive Technology for Performing and Media Arts MA program. In 2014–15 he worked as a visiting student researcher on the Antescofo project at IRCAM, Paris. His research interests include sonification and interactive music systems.

•

Valentin Mihai is currently a Ph.D. student at University "Politehnica" Bucharest, where he recently obtained an M.Sc. in audio production with the dissertation project *Sound Control by Gestures with the Help of a Vicon System*. His research activity covers radar and IR communications and human-computer interaction.

•

Ştefan Damian is pursuing a Ph.D. in sound design for interactive media at the UNATC "I.L. Caragiale" Bucharest. He has an MA in Sonic Arts from SARC Queen's University Belfast. His area of activity includes synthesis, spatialization, and electroacoustic music.