An Environment for Gestural Interaction with 3D Virtual Musical Instruments as an Educational Tool

C. Garoufis [1,3], A. Zlatintsi [1,3], K. Kritsis [2,4], P.P. Filntisis [1,3], V. Katsouros [2] and P. Maragos [1,3]
[1]School of Electr. & Comp. Enginr., National Technical University of Athens, 15773 Athens, Greece
[2]Institute for Language and Speech Processing, Athena Research Center, 15125 Maroussi, Greece
[3]Robot Perception and Interaction Unit, Athena Research Center, 15125 Maroussi, Greece
[4]Department of Informatics, University of Piraeus, 18534 Piraeus, Greece
Email: cgaroufis@mail.ntua.gr; [nzlat, maragos]@cs.ntua.gr; filby@central.ntua.gr; [kosmas.kritsis, vsk]@ilsp.gr

***Abstract*—This paper presents a finalized version of an environment intended for performance and gestural interaction with three-dimensional virtual musical instruments, developed as a part of a larger educational platform, the iMuSciCA workbench. The environment can employ either a Leap Motion or a Kinect sensor, and enables interaction with a variety of virtual musical instruments, namely virtual interpretations of a bichord, a xylophone, a drumming set, a guitar and an upright bass, by means of performing and recognizing hand gestures similar to the ones needed to play their physical counterparts. In order to showcase the usability of the platform in an educational context and measure its effectiveness, we designed a scenario, where the user tries to keep a steady rhythm while drumming. A usability study of the above scenario, involving 22 users, demonstrates that the audiovisual feedback can actually provide assistance to the user.**

***Index Terms*—virtual musical instruments, gesture recognition, gestural interaction, educational tool, HCI**

I. Introduction

The development of virtual musical instruments (VMIs) and virtual reality musical instruments (VRMIs) has been an emerging field in Human-Computer Interaction (HCI). A virtual musical instrument [17] is defined as a non-physical object that constitutes a simulation of an existing musical instrument, focusing on sonic emulation, while a virtual reality musical instrument extends on that, by providing visual as well as auditory feedback. Their engaging nature, as well as the recent development of non-intrusive sensors [6], such as Microsoft Kinect[1] and Leap Motion[2], make VRMIs suitable for musical education. They can, for instance, provide an appealing tool for conveying concepts from music theory, since, as argued in [1], a number of musical concepts is easier to explain by linking them to specific movements.

Indeed, motion and gestural interactions with digital, virtual and virtual reality musical instruments have been receiving increased attention from the scientific community in recent years. An early attempt is documented in [12], where both data gloves and vision sensors are deployed for interacting with VRMIs, along with 3D stereoscopic data and a physical modeling sound engine.

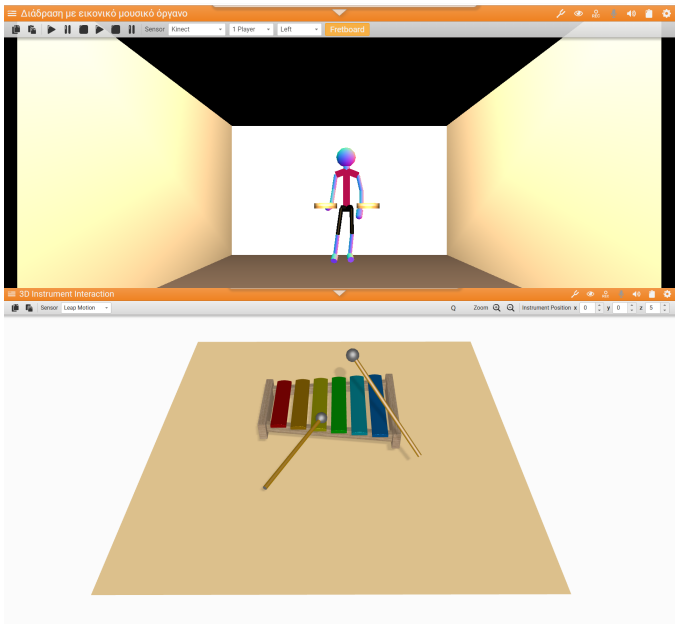

Fig. 1: Snapshots of the iMuSciCA performance environment, using the Kinect (top) and Leap Motion (bottom) sensors.

Since then, a large amount of DMIs and VRMIs have employed commercial motion sensors, such as Kinect and Leap Motion. Aside from their non-intrusiveness, these sensors also offer a skeletonized and easily trackable model of the human body and the hand palm, respectively. Kinect has been used in a variety of virtual music instruments, such as implementations of guitar and drums [8], or even an augmented piano that utilizes tracking of the performer's hand above the piano to manipulate a variety of synthesis parameters [4]. Moreover, it has been employed in more abstract explorations of musical spaces, including a conductor-like system that enables control of various sonic options [14], and Crossole [16], a visual platform that represents chords using blocks. The performer can manipulate them by moving his/her hand in front of a Kinect sensor, constructing in this way chord progressions. On the other hand, [18] reports on the development of an augmented handpan, using a Leap Motion sensor, potentially encompassing various educational needs.

This work is supported by the *iMuSciCA* project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 731861.

[1]https://developer.microsoft.com/en-us/windows/kinect

[2]https://www.leapmotion.com/

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

In recent years, the concept of integrating artistic activities in traditional STEM (Science, Technology, Engineering and Mathematics) curricula, also referred to as STEAM [19] (where A stands for Arts), has gained traction due to argued societal deficiencies in traditional STEM pedagogy [20]. The advantages of this approach are twofold: Not only does the integration of arts in STEM curricula foster problem-solving, critical thinking, creativity and communication skills, but it also makes STEM subjects more attractive, by using cross-disciplinary activities as a means of showcasing a practical use case for them. Examples of such activities include the design of digital musical instruments [7] and 3D printing using Computer-Aided Design software [2], as they can combine engagement in core STEM subjects with artistic expressiveness.

In fact, it can be claimed that the development of online technologies has opened new avenues for musical education [5], aided by the implementation of Javascript libraries intending to facilitate interactive music creation in browser environments [13]. With regards to using these technologies towards classroom education, a number of works can be found in literature. For instance, [15] describes a number of educational scenarios and applications geared towards collaborative creation and intended for use in high school education, while [3] presents an interactive environment where sounds can actually be drawn on a whiteboard, thus enabling intuitive exploration of various sonic parameters.

iMuSciCA (interactive Music Science Collaborative Activities[3]) is a research project grounded in the need for STEAM pedagogy, that aims to build a web platform that will utilize music-related activities and innovative educational technologies in order to promote cross-disciplinary learning. Specifically, the aim is to develop: (i) original and innovative enabling technologies to facilitate open co-creation tools incorporated in music activities to support STEM learning, (ii) a set of practical activities to give learners the opportunity to discover about different phenomena/laws of physics, geometry, mathematics and technology through creative music activities, and (iii) to encourage students to engage in innovative interactive music activities with advanced multimodal interfaces, raising this way their interest in science and technology with the support of creative and artistic interventions.

In this work, which is developed as a part of iMuSciCA, we propose a complete virtual environment that facilitates gestural interaction with three-dimensional VRMIs in an educational context. This work expands on the work reported in [21] and [10], providing a larger variety of instruments and interactions, in an environment compatible with both Kinect and Leap Motion sensors, incorporating gestural recognition during the performance. Furthermore, a rhythmic analysis test using the virtual drumkit was designed in order to determine the usability of our environment in educational applications, by measuring in what extent the audiovisual feedback helps the user to follow a steady beat, inspired by a similar study conducted in [11].

[3]http://www.imuscica.eu/

The remainder of the paper is structured as follows: Section II describes the overall architecture of the environment, as well as the gestural interactions with the virtual musical instruments. In Sec. III, the usability study is described, and its results are analyzed and discussed, while the conclusions and future research directions are discussed in Sec. IV.

II. System Architecture

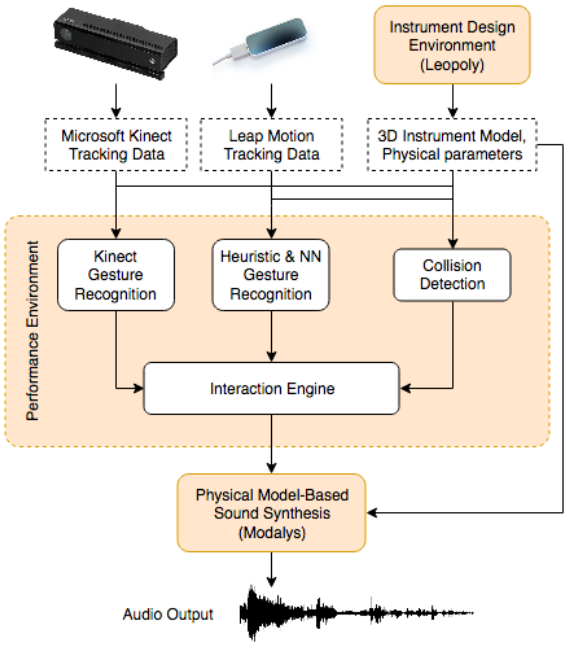

Fig. 2: General architecture of the proposed system.

We have developed a 3D virtual environment (see Fig. 2), where one or two players can interact with virtual instruments using hand gestures. Depending on the sensor, we visualize either the performer's hands (Leap Motion) or a skeletonized full-body avatar of the performer (Kinect). We have also designed and developed the gestural interactions between the performers and the VIs, described in Sec. II-A and Sec. II-B.

The developed performance environment uses external sources for both the design of the virtual instruments and the respective sonic output. Regarding the first part, we use a set of realistic 3D visualization technologies implemented by Leopoly[4], who have developed an environment where 3D virtual musical instruments with user-defined physical parameters can be designed. The sonic output that corresponds to each instrument originates from Modalys[5], a physical modeling sound synthesis engine developed by IRCAM.

A. Interactions with Leap Motion

The goals and technical details of the Leap Motion enabled performance environment have been previously presented in [10]. However the project is under development and additional instruments have been designed and implemented with

[4]https://leopoly.com/
[5]http://forumnet.ircam.fr/product/modalys-en/

their corresponding interactions, such as plucked and bowed string instruments (i.e. monochord, guitar, upright bass) as well as mallet-based percussion (i.e. xylophone). Furthermore, previously developed heuristic-based interaction methods have been updated [10], while experimenting with state-of-the-art deep Neural Network architectures for improving our gesture recognition component [9]. Other supported modules that enrich the educational and creational aspects of the environment include a gesture recorder and a musical/rhythmical quantizer, enabling the user to edit his/her recordings while giving a deeper insight of his/her performances by reproducing the same visual and auditory feedback. The different interaction approaches are described next.

String Interaction: In this approach, the interaction engine takes into account the raw 3D positions of the fingers' joints, as provided from the Leap Motion sensor, in order to reconstruct the hand skeleton. Next, two factors are calculated; first, the Gesture Recognition module evaluates the temporal movement of the joints and decides whether the user performs a plucking gesture. When the gesture is recognized, the Collision Engine calculates the distances between the 3D positions of the strings and the distal phalanges bone of the finger that performed the gesture. Finally, the string with the shortest distance is triggered, while translating the 3D point of the collision to the physical (local) plucking point on the modeled string.

Membrane Interaction: For percussion instruments consisting of membranes, the interaction engine computes the transformation matrix (i.e. position, rotation and translation) of the visible palms, in order to control a set of virtual 3D drumsticks. Then the Collision Engine calculates the distances between the 3D positions of the tips of the drumsticks and the 3D surface that models the membrane's position and shape (e.g. square, circular). If the distance is less than a predefined threshold, then a collision is detected and the surface is triggered, while considering the 3D collision point as the physical (local) impact point.

Surface Interaction: This interaction method is implemented similar to the membrane interaction, with the difference that the considered instrument consists of multiple rectangle surfaces, in our case the bars of the xylophone. Furthermore, the exported transformation matrices of the palms are applied on a set of virtual mallets. Consequently, the collision engine calculates the distances between the xylophone bars and the tips of the mallets and triggers a collision event when the head of a mallet lies within the margins of the surface of the modeled bar.

B. Interactions with Microsoft Kinect

Currently, the Kinect environment supports a variety of interactions, including plucked and bowed string instruments (i.e. guitar, upright bass), membranes (i.e. drums), as well as mallet-based percussion (i.e. xylophone). In all cases, the interaction takes place by means of recognizing instrument-specific gestures, after processing the skeletal data from the Kinect sensor; see Fig. 3 for the Kinect-based interactions.

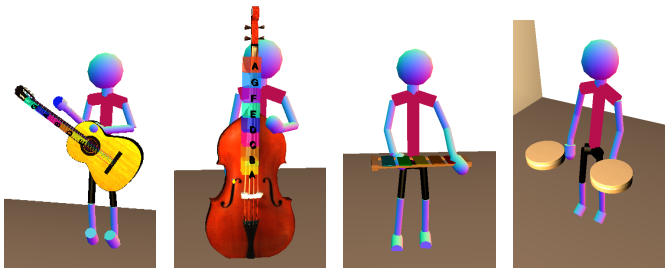

Fig. 3: Interactions with the Kinect-based virtual musical instruments: Air Guitar, Upright Bass, Xylophone and Drumkit (from left to right).

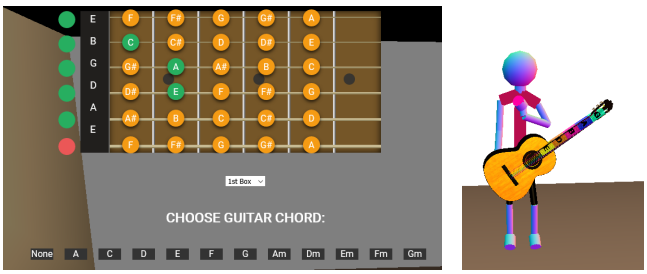

Fig. 4: Visualizations of the fretboard, where the chords of the five guitar positions can be defined (left) and the air guitar for a left-handed player (right).

Other supported options include multiplayer performance, see also [21], and the functionality for both left- and right-handed players in the case of the guitar and the upright bass, as shown in Fig. 4 (right) for the guitar.

Air Guitar Interaction: The dominant hand of the performer is placed at the height of their waist, and performs an up-and-down plucking gesture in order to trigger sonic events. The chords played depend on the positioning of the performer's non-dominant hand, which is placed along the virtual guitar's fretboard that is divided in five distinct areas, each corresponding to a different chord. The chord that corresponds to each area is either user-defined (by determining the fingerings for each string), or chosen between some predefined chords, in the fretboard pop-up menu, see Fig. 4 (left).

Upright Bass Interaction: Similar to the air guitar interaction, the dominant hand of the performer is placed at the height of their waist, and performs a left-to-right bowing gesture. A bowing sound is produced as long as the above gesture is recognized. The height of the performer's non-dominant hand determines the pitch of the produced sound; the lower the hand is placed, the sharper the sound is.

Xylophone Interaction: In this case, the player places both hands in front of him/her, at the height of their waist, performing hitting gestures as if the hands are used as mallets. Once the gestures are recognized, a sound, corresponding to the pitch of the specific xylophone bar, is produced. The pitch of the produced sound is determined by the x-axis positioning of the hand that collides with the virtual bar.

Drumkit Interaction: The interaction occurs similar to the virtual xylophone. A pair of membranes are generated in front of the user, corresponding to the left and right hand respectively. Each membrane is generated as a circular object, of predetermined center (placed at 40cm diagonally from

the performer, at approximately the height of their waist) with a user-defined radius, see Fig. 3. The users perform downwards, hitting gestures with their hands, and whenever a collision between their hands and any membrane is detected, a percussive sound is produced.

III. Usability Study

As previously mentioned, the developed environment for interacting with 3D VRMIs is a part of the larger educational platform of the iMuSciCA project. Thus, apart from the development of interactive musical activities with advanced multimodal interfaces, we would like to explore some educational scenarios as well. In this context, we designed a usability study, and specifically a scenario involving the virtual drum membranes, intending to investigate to what extent the audio-visual feedback of the interactive environment can operate as an assisting tool towards maintaining specific rhythmic patterns while drumming.

Experimental Protocol: In order to accomplish this, the following four setups were considered for evaluation:

- Lack of feedback (LF).
- Visual-only feedback (VF): The user along with the virtual drumkit appears on the browser, but no sound is produced.
- Audio-only feedback (AF): No avatar corresponding to the user is shown, but sound is produced whenever the user's hands collide with the virtual drumkit.
- Audio-visual feedback (AVF).

In all setups, a metronome keeping a steady tone in 40 BPM (beats per minute) was used, in order to both enable the baseline comparisons with the users' actual hitting and to offer a minimal aiding tool. In total, 22 users tested the various feedback combinations, 14 having prior musical background. The users were instructed to stand in front of the Kinect sensor, and try to match the following rhythmic patterns:

- Experiment 1: Steady 80 BPM rhythm with the dominant hand, for a duration of 15 seconds.
- Experiment 2: Steady 80 BPM rhythm with the dominant hand, while the subdominant hand plays at 40 BPM, for a duration of 15 seconds.
- Experiment 3: The dominant hand plays successively at 40 BPM, 80 BPM, 120 BPM and 160 BPM, for a duration equal to 5 metronome hits (7.5 seconds) each. This is equivalent to hitting once, twice, thrice and four times per each metronome hit, respectively.

To ensure that the produced results are not skewed in favor of any of the 4 setups, all experiments were executed four times, with a randomized setup order.

A. Evaluation Results and Discussion

For all combinations of the above scenarios and setups, we used the following evaluation metrics (calculated in ms), which should ideally be as low as possible:

- The standard deviation (std) of the recorded time intervals between successive drum hits, as a measure of keeping a steady rhythm.

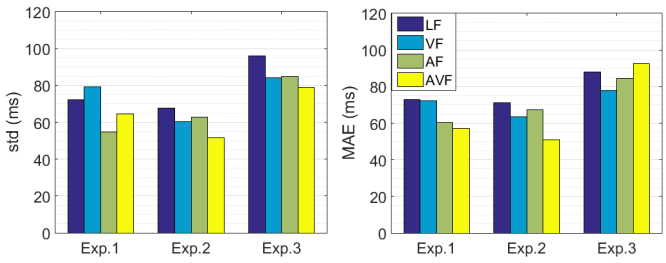

Fig. 5: Average standard deviations (std) of the intervals between successive user hits and average mean absolute errors (MAE) between recorded and dictated successive user hit intervals, in ms, for each of the 4 tested setups, for all 3 experiments.

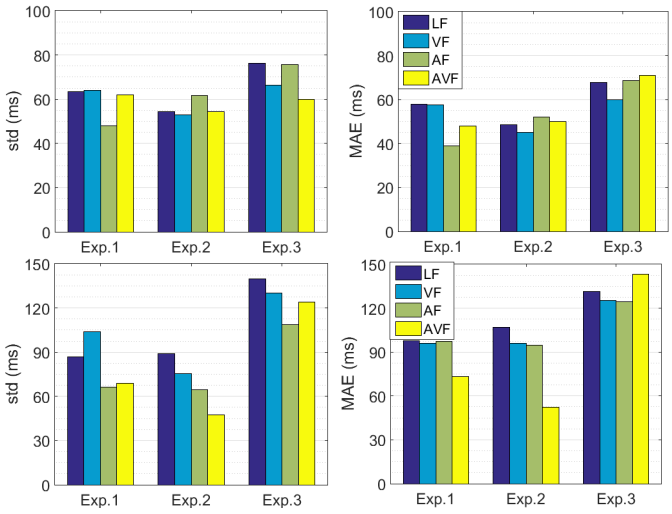

Fig. 6: Average standard deviations (std) of the intervals between successive user hits and average mean absolute errors (MAE) between recorded and dictated successive user hit intervals, in ms, for each of the 4 tested setups, for all 3 experiments, for the users with (top row) and without (bottom row) a musical background.

- The Mean Absolute Error (MAE) of the recorded time intervals between two successive drum hits, compared to the ones dictated by the scenario, as a measure of keeping the predetermined rhythm.

For example, a set of successive drum hits in almost identical time intervals, that differ from the ones dictated by the scenario, would achieve a low std score, but a relatively high MAE score.

The results of the evaluation can be found in Fig. 5. From these results, we can deduce that the presence of audio-visual feedback does help in maintaining specific rhythmic patterns. We may also note that, when comparing auditory and visual feedback, the auditory feedback is improving the metrics in all experiments, while the visual feedback only does so in the second and the third experiment. A possible explanation could lie in the fact that the second experiment involves movements of both hands instead of only one, being thus harder with regards to body part coordination, while the third requires relatively quick hand movements.

In order to get further insights regarding the above results, we calculated the values of both the std and the MAE, for all experiments and possible feedback combinations, while taking into consideration the musical background of the users. The results are presented in Fig. 6.

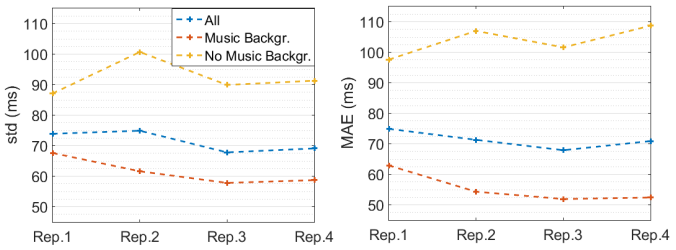

Fig. 7: Average standard deviations (std) of the intervals between successive user hits and average mean absolute errors (MAE) between recorded and dictated successive user hit intervals, in ms, averaged over all three experiments, ordered by repetition.

We observe that while the behavior of both user groups is similar in the third experiment, the users with a musical background did not achieve a remarkable improvement in the metrics in the first two experiments. On the other hand, those without a musical background were significantly assisted by the audio-visual feedback. This could be explained, since the patterns in the first two experiments are easy enough for someone with musical background to perform instinctively, something that does not generally apply to people without a musical background.

Finally, in Fig. 7, we analyze the effect of the repetitions towards learning the predetermined rhythmic patterns. As we can see, there is a decreasing trend regarding both std and MAE scores, especially for the users with musical background. For the users without musical background, we may conclude that the presence of audiovisual feedback is the primary factor towards improving the ability to successfully follow the predetermined rhythmic patterns.

IV. Conclusions and Future Work

This paper describes our work on the development of an environment, where users interact with virtual musical instruments by means of hand gestures, which are recognized after processing skeletal data from motion sensors. A usability study involving the use of the virtual drumkit as an educational tool is also presented, showing encouraging results with regards to the effect of the audio-visual feedback, as well as the pattern repetitions. Potential future research avenues are twofold. On the one hand, the gestural interactions can be further enhanced, so as to give the players greater creative control over their performances, as well as alleviate potential issues due to the inherent latency of the various technical components of the system. On the other hand, regarding the educational aspect of our environment, further experiments are designed, regarding either theoretical musical knowledge or more practical playing skills. For instance, in the case of the guitar, where the interaction is more "metaphorical" than physical, experiments could be designed to explain concepts from music theory, such as consonant and dissonant chords and melodic chord progressions.

Acknowledgments

The authors would like to thank all the students and colleagues at NTUA and ATHENA RC for participating in the experiments, as well as the project partners from Leopoly, for developing the 3D instrument design environment, IRCAM for providing the JavaScript port of Modalys, and EA for the discussions about the experiment design.

References

[1] A. Antle, M. Droumeva, and G. Corness, "Playing with the sound maker: Do embodied metaphors help children learn?" in *Proc. 7th Intl. Conf. on Interaction Design and Children*, Chicago, IL, USA, 2008.

[2] A. Bicer, S. B. Nite, R. M. Capraro, L. R. Barroso, M. M. Capraro, and Y. Lee, "Moving from stem to steam: The effects of informal stem learning on students' creativity and problem solving skills with 3d printing," in *2017 IEEE Frontiers in Education Conf.*, Indianapolis,IN,USA, 2017.

[3] M. Bouillon, F. Simistira, R. Ingold, and M. Liwicki, "Drawme: Drawing canvas for music creation - a new tool for inquiry learning," in *Proc. of the 4th Intl. Conf. on Learning and Teaching*, Singapore, Singapore, 2018.

[4] W. Brent, "The gesturally extended piano." in *Proc. NIME 2012*, Ann Arbor, MI, USA, 2012.

[5] A. M. Burns, S. Bel, and C. Traube, "Learning to play the guitar at the age of interactive and collaborative technologies." in *Proc. SMC 2017*, Espoo, Finland, 2017.

[6] M. Gleicher and N. Ferrier, "Evaluating video-based motion capture," in *Proc. of the 2002 Computer Animation Conf.*, Switzerland, 2002.

[7] J. Harriman, "Start em young: Digital music instruments for education," in *Proc. NIME 2015*, Baton Rouge, LA, USA, 2015.

[8] M.-H. Hsu, W. Kumara, T. Shih, and Z. Cheng, "Spider king: Virtual musical instruments based on microsoft kinect," in *Proc. 2013 Intl. Joint Conference on Awareness Science and Technology and Ubi-Media Computing*, Aizuwakamatsu, Japan, 2013.

[9] K. Kritsis, A. Gkiokas, M. Kaliakatsos-Papakostas, V. Katsouros, and A. Pikrakis, "Deployment of lstms for real-time hand gesture interaction of 3d virtual music instruments with a leap motion sensor," in *Proc. SMC 2018*, Limassol, Cyprus, 2018.

[10] K. Kritsis, A. Gkiokas, Q. Lamerand, R. Piechaud, C. Acosta, M. Kaliakatsos-Papakostas, and V. Katsouros, "Design and interaction of 3d virtual music instruments for steam education using web technologies," in *Proc. SMC 2018*, Limassol, Cyprus, 2018.

[11] T. Maki-Patola, "User interface comparison for virtual drums," in *Proc. NIME 2005*, Vancouver, BC, Canada, 2005.

[12] T. Maki-Patola, J. Laitinen, A. Kanerva, and T. Takala, "Experiments with virtual reality instruments," in *Proc. NIME 2005*, Vancouver, BC, Canada, 2005.

[13] Y. Mann, "Interactive music with tone. js," in *Proc. 1st annual Web Audio Conf.*, Paris,France, 2015.

[14] A. Rosa-Pujazon, I. Barbancho-Perez, L. Tardon, and A. Barbancho-Perez, "Conducting a virtual ensemble with a kinect device," in *Proc. SMC 2013*, Stockholm, Sweden, 2013.

[15] J. Sastre, J. Cerda, W. Garcia, C. Hernandez, N. Lloret, A. Murillo, D. Pico, J. Serrano, S. Scarani, and R. Dannenberg, "New technologies for music education," in *Proc. 2nd Intl. Conf. on E-Learning and e-Technologies in Education*, Lodz, Poland, 2013.

[16] S. Senturk, S. W. Lee, A. Sastry, A. Daruwalla, and G. Weinberg, "Crossole: A gestural interface for composition, improvisation and performance using kinect." in *Proc. NIME 2012*, Ann Arbor, MI, USA, 2012.

[17] S. Serafin, C. Erkut, J. Kojs, N. C. Nilsson, and R. Nordahl, "Virtual reality musical instruments: State of the art, design principles, and future directions," *Computer Music Journal*, vol. 40, pp. 22–40, Fall 2016.

[18] C. Steinmeyer and D. Becking, "Towards creating an augmented handpan using leap motion," in *Proc. SMC 2018*, Limassol, Cyprus, 2018.

[19] G. Yakman, "Steam education: An overview of creating a model of integrative education," in *Proc. of the Pupils' Attitudes Towards Technology Conf.: Research on Technology, Innovation, Design & Engineering Teaching*, Salt Lake City, UT, USA, 2008.

[20] D. Zeidler, "Stem education: A deficit framework for the twenty first century? a sociocultural socioscientific response," *Cultural Studies of Science Education*, vol. 11(1), pp. 11–26, 2016.

[21] A. Zlatintsi, P.-P. Filntisis, C. Garoufis, A. Tsiami, K. Kritsis, M. Kaliakatsos-Papakostas, A. Gkiokas, V. Katsouros, and P. Maragos, "A web-based real-time kinect application for gestural interaction with virtual musical instruments," in *Proc. AM 2018*, Wrexham, Wales, 2018.