Gestural User Interface for Audio Multitrack Real-time Stereo Mixing

Konstantinos Drossos
Digital Audio Processing & Applications Group
Audiovisual Signal Processing Laboratory
Dept. of Audiovisual Arts, Ionian University
Corfu, Greece
kdrosos@ionio.gr

Andreas Floros
Digital Audio Processing & Applications Group
Audiovisual Signal Processing Laboratory
Dept. of Audiovisual Arts, Ionian University
Corfu, Greece
floros@ionio.gr

Konstantinos Koukoudis
Digital Audio Processing & Applications Group
Audiovisual Signal Processing Laboratory
Dept. of Audiovisual Arts, Ionian University
Corfu, Greece
t11kouk@ionio.gr

ABSTRACT

Sound mixing is a well-established task applied (directly or indirectly) in many fields of music and sound production. For example, in the case of classical music orchestras, their conductors perform sound mixing by specifying the reproduction gain of specific groups of musical instruments or of the entire orchestra. Moreover, modern sound artists and performers also employ sound mixing when they compose music or improvise in real-time. In this work a system is presented that incorporates a gestural interface for real-time multitrack sound mixing. The proposed gestural sound mixing control scheme is implemented on an open hardware micro-controller board, using common sensor modules. The gestures employed are as close as possible to the ones particularly used by the orchestra conductors. The system overall performance is also evaluated in terms of the achieved user experience through subjective tests.

Categories and Subject Descriptors

J.5 [**Arts and Humanities**]: Performing Arts

General Terms

Algorithms

Keywords

Gestural real-time sound mixing, gestural interaction, real-time sound mixing, real-time interaction

1. INTRODUCTION

Music is an important aspect of every day life [3]. In order to allow for a variety of different usage scenarios of listening to music (i.e. mobile, desktop/high-fidelity or surround), a wide range of audio formats has been defined that supports different application types. However, the majority of the available musical content is available in stereo [4, 5].

The process of creating such audio material involves several stages, such as recording, mixing and mastering [4]. During the mixing stage, the combination of the waveforms corresponding to the desired sound sources in two audio channels is performed. In particular, the sound engineer processes the sources both in terms of their frequency components and their dynamics and integrates them in two channels by defining their relative gain. This controlled gain variation is performed under specific panning laws (such as [10, 11]) and results into a simple means for spatial positioning of the sound sources within the virtual stereo scene.

The aforementioned process usually imposes the employment of particular hardware, like a mixing console. However, this is not the only application case where sound mixing is required. For example, in the case of a musical orchestra, all musical instrument groups are located in a fixed spatial position, indicated by the type and the arrangement of the orchestra itself. In addition, the conductor can alter the dynamics and, thus, the reproduction gain of each musical instruments group. Consequently, it can be stated that an orchestra conductor seems to realize a kind of audio mixing, following a specific panning law. Obviously, the interface for the aforementioned process is the conductor's gestures. Moreover, considering that the sound sources (i.e. the musical instruments groups) are in fixed positions, the alteration of their reproduction gain is also a way to alter the balance of the overall sound produced by the orchestra.

Furthermore, new music genres are based on real-time mixing. In live concerts of for example, some artists utilize mixing consoles and multitrack sequencers/recorders in order to compose their music on-the-fly. Thus, the available interface to exploit their artistic result is restrained by the hardware itself and therefore it is likely to also suppress their artistic expression by not employing intuitive interaction. Recently developed technologies use gestural interaction as a means for easy-to-use and easy-to-learn interfaces [5, 13]. Focusing on the employment of gestural interfaces in artistic applications, such technologies are expected to allow the development of novel means for human-machine interaction and thus enhance the artistic potential [8].

In this work we present a prototype gestural interface for

AM '13, September 18 - 20 2013, Piteå, Sweden
Copyright 2013 ACM 978-1-4503-2659-9/13/09 ...$15.00
http://dx.doi.org10.1145/2544114.2544123.

multitrack audio mixing in real-time. This interface incorporates gestures very similar to the ones used by an orchestra conductor. The developed prototype utilizes an open hardware micro-controller board as the physical interface platform along with infrared proximity sensors, accelerometers and buttons. It can be combined with any digital sequencer that is capable to receive Musical Instrument Digital Interface (M.I.D.I.) messages and controls the reproduction gain, the spatial positioning and the application of one audio effect for each of the audio tracks. The overall mixing efficiency and usability of the proposed sound-mixing gestural scheme is assessed using subjective evaluation tests. The results obtained show a clear preference for the developed system over classic means for multitrack mixing.

The rest of the paper is organized as follows: Section 2 contains a brief overview of related works. In Section 3, the developed prototype is introduced and analytically described. Next, Section 4 outlines the evaluation procedure followed for assessing the subjective performance of the proposed system. Finally, Section 5 summarizes the results obtained along with their discussion, while Section 6 concludes the work.

2. RELATED WORK

Although gestural interaction and control is widely used in consumer devices, e.g. smartphones, to the best of authors' knowledge there is a scarcity of projects and publications related to stereo audio mixing. Nevertheless, there are published works focusing on gestural interaction using mobile devices. In [9] for example, an analytic study is performed for the gestural and audio metaphors in controlling such devices. This particular work showed that the creation of an interface without visual interaction is indeed feasible. Moreover, a recent research work focused on auditory cues for gestural audio control interaction and showed that there is a genre-dependent bias on the cues used [7]. This work employed a system based on the widely-known WiiMote controller [15].

The WiiMote controller tends to be used in many other experimental implementation of prototypes aiming to investigate effective means for gestural control of audio synthesis and mixing [6, 12, 14]. Nonetheless, the utilization of an additional or alternative controller seems to introduce some extra effort in terms of the pre-defined handling operations, while some issues on the correct placement of the source are also reported [12].

One of the few published works that performed gestural audio mixing without using an off-the-shelf remote controller has incorporated binaural mixing [5]. Although this work represents an attempt towards efficient audio mixing, it incorporates binaural and not stereo audio. In the present work we are concerned with a stereo mixing system without any remote controller: we use only gestures which are chosen properly in order to enhance the intuitive interaction of both listeners and musicians/composers. Finally, we limit our implementation to stereo audio due to its wide spread and acceptance.

3. THE GESTURAL STEREO MIXER IMPLEMENTATION

The proposed gestural sound mixing system accepts as input the user's gestures or selections and controls the mixing of audio tracks in real time. Generally, it consists of three sensors, two button groups, one single extra button and one micro-controller board. The sensors employed are two infrared proximity and one accelerometer. All buttons were push buttons and in total were nine (9). Eight (8) of them were in two groups of four (4). The micro-controller board selected was the Arduino Mega, based on the ATmega1280 micro-controller [2]. The micro-controller was programmed using the Arduino Integrated Development Environment (IDE) [1], utilizing both C and C++ programming languages.

User's gestures are captured through the sensors mentioned above, while user selections are performed using the push buttons. Specifically, the identified gestures control a) the applied panning, b) the reproduction gain as well as c) the parameters of one pre-assigned audio effect. All user selections are made using the two groups of buttons. These selections are aiming to define which tracks are currently reproduced and which tracks are affected by the gestures. The extra button functionality corresponds to a selection confirmation that enables the gesture application to the selected track.

All the controlling information derived by the above modules is transmitted using the M.I.D.I. protocol to a multitrack sequencer, which is responsible for handling the audio tracks. Aim of the overall system is to offer an intuitive gestural control for real time multitrack mixing. An image of the system's parts, sensors and buttons is illustrated in Figure 1. The sensors are appeared at the top of the image. From left to right one can see: a) the two infrared proximity sensors and b) the accelerometer employed. At the bottom of the Figure, the two groups of buttons along with the extra selection button are also appeared.

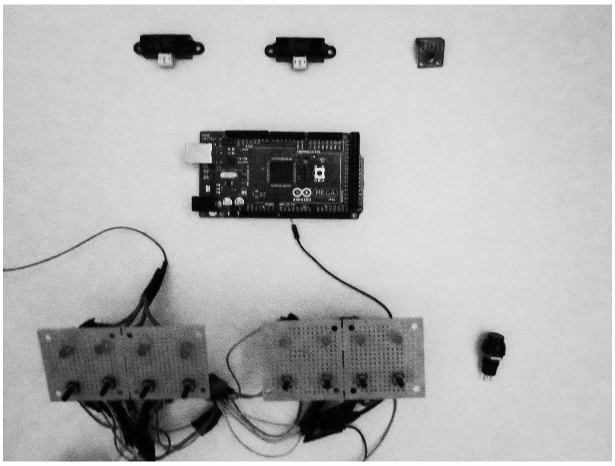

Figure 1: The hardware modules of the implemented system

3.1 Sound Mixing Gestural Control Details

Gestural control was realized based on mapping distance and rotation values to integers ranging from 0 to 255. More specifically, gestural panning and gain control employed the distance between the user's hands in two dimensions, as derived by the proximity sensors. Additionally, the one hand rotation value obtained through the accelerometer was used to control the application of the pre-assigned audio effect.

The mapping (or quantization) of the measured distance values to integers was performed with the built-in function of Arduino-IDE *map()* [1].

Focusing on the stereo gain parameter, a value of 0 mutes the sound output for the corresponding stereo channel, while a value equal to 255 corresponds to 0 $dBFS$. For the panning law control, a derived value of a) 0 was mapped to 100% left panning (equal to M.I.D.I. value for pan of 0) b) 128 was equal to center panning (equal to M.I.D.I. value for pan 64) and c) 255 was equal to 100% right panning (equal to M.I.D.I. value for pan 127). For the application of the pre-assigned audio effect, a value of 0 was mapped to dry only sound, whereas a value of 255 to wet only sound (applied for the selected track only).

All values were mapped to 0 - 127 range prior sending them in forms of M.I.D.I. messages. Moreover, due to hardware implications, the useful range of distances for the proximity sensors was 15cm to 60cm, measured from the center of each sensor. Any other range was truncated to the closest of the boundaries of useful ranges, i.e. a range of 10cm was set equal to 15cm, whereas a range of 70cm was limited to 60cm. Thus, for the infrared proximity sensors the distance of 15cm was set equal to 0 and the distance of 60cm was equal to 255.

The proximity sensors were placed appropriately in order to measure the distance between the sensor and the user's hands in both the horizontal and vertical axes. The measured distance in the horizontal axis was assigned the control of the spatial positioning of the track (i.e. the panning law implementation), while the corresponding distance in the vertical axis was mapped to the overall reproduction gain. Moreover, in order to achieve a more flexible and integrated user interaction scheme, the user had to press with ihis foot the available extra button and release it when he wanted to end the control of the selected track. This allowed consequent processes for each track, without altering the previously achieved state of the specific track.

As simple usage scenarios, if the user wants to increase the reproduction gain of the selected track, he will have to press the extra button and raise his hand above the proximity sensor assigned to the vertical axis. If the wishes to control the spatial position of the selected track, then he will have to move his hand to or away from the proximity sensor assigned to the horizontal axis, while pressing the extra button. Figure 2 includes an illustration of the exact layout for the proximity sensors.

The previously described gestural scheme for controlling the reproduction gain and spatial positioning was selected as the closest to a widely accepted method, i.e. the one used by the orchestras conductors, where the vertical axis is assigned to the overall gain that will be produced from the orchestra. The horizontal positioning of the conductor's hands is used as a selection for a group of musical instruments and, therefore, it can be considered to be equivalent to a kind of spatial positioning since the location of the musical instruments is fixed in space.

The application of the audio effect to a specific track was performed through the rotation of the hand that also controlled the spatial positioning, following the functionality of turning a knob. The graphical representation of controlling the audio effect is illustrated in Figure 3. It should be noted here that the audio effect control can be performed concurrently with the control of spatial positioning and the definition of the overall reproduction gain.

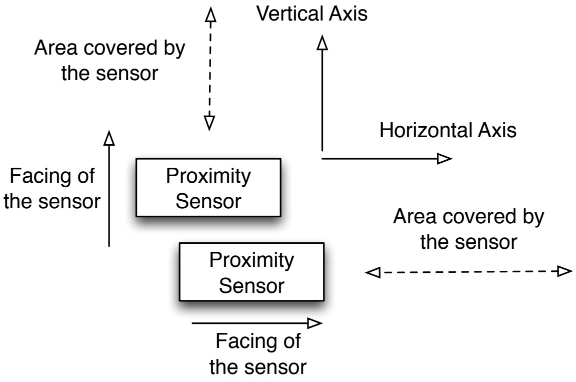

Figure 2: Schematic diagram of the volume and spatial positioning control (figure not in scale)

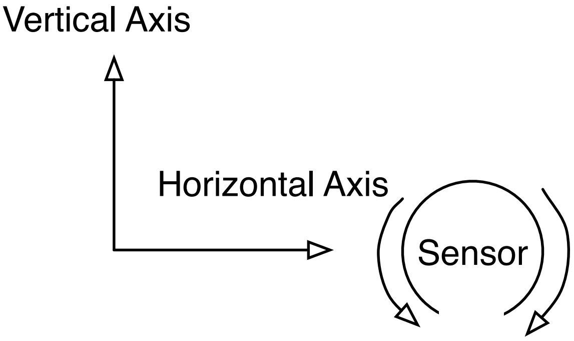

Figure 3: The gesture recognised by the accelerometer sensor (figure not in scale)

An example of the exact gestural notations for controlling the reproduction gain and the spatial position is included in Figure 4. The same information that corresponds to controlling the reproduction gain, the the spatial positioning and audio effect application is summarized in Figure 5. Finally, Figure 6 illustrates the block diagram of the complete gesture-control algorithm.

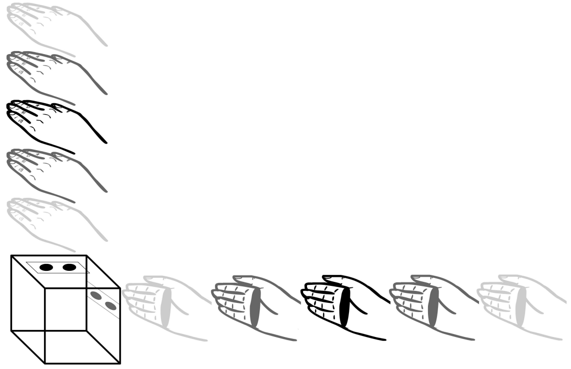

Figure 4: Example of reproduction gain and spatial positioning control usage (figure not in scale)

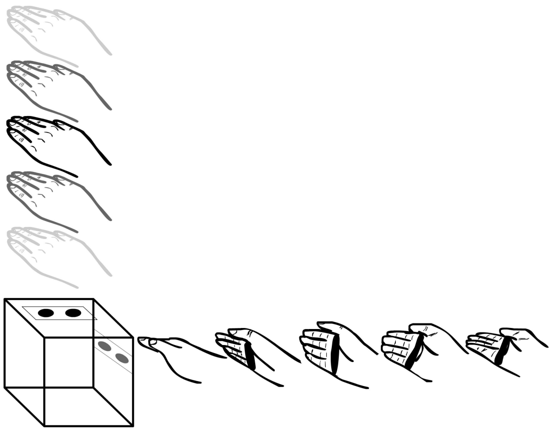

Figure 5: Example of reproduction gain, spatial positioning and audio effect application control usage (figure not in scale)

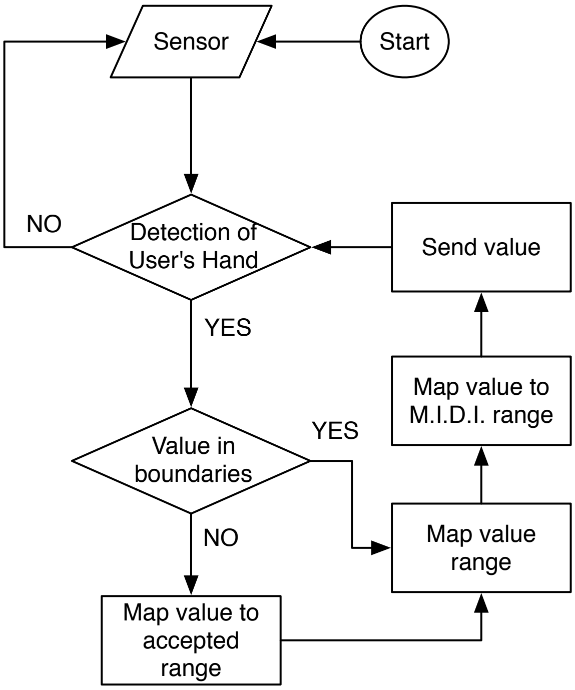

Figure 6: Illustration of algorithm for the gestural interaction

Following Figure 6, when the user selects a track and activates it for processing, the system maps the range of the distance values between the user's hand and the maximum accepted distance to the range of the current value and the maximum accepted one. This extra mapping functionality ensures that no accidental changes will be made due to different distances measured by the sensors.

3.2 Track Selection

Track selection and track control was performed through the available groups of buttons mentioned previously. The first group controls which track (or tracks) will be reproduced. The second, enables the full control of the selected track. Figure 7 clarifies the above selection and control scheme, with the upper part illustrating the track selection process for reproduction and the lower one demonstrating the selection of the track to be controlled. In this Figure, four tracks can be controlled. The first and the third track are selected for reproduction (top) and the third one is selected for control (bottom). Finally, a light indication provides visual feedback showing which tracks were selected per group of buttons.

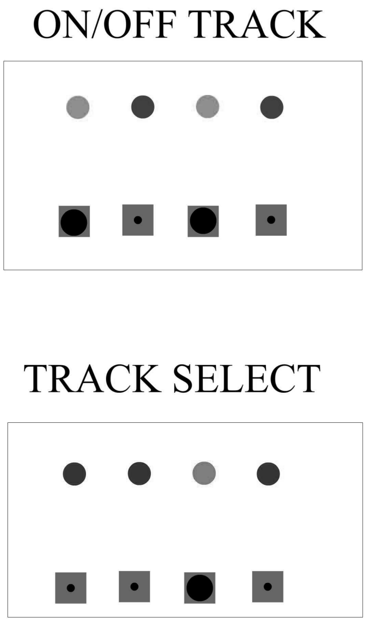

Figure 7: Example of the button groups usage for track reproduction (top) and controlling (bottom) selection.

The extra button was used to enable the overall control functionality of the system. Thus, each time the user wanted to control the selected track he had to press and hold the extra button. In order to avoid the employment of gestures for this critical task, pressing the extra button was designed to be made by foot.

4. SYSTEM EVALUATION

The gestural sound mixing scheme evaluation was performed through a sequence of subjective measurements. Aim of these measurements was to provide an estimation of the user experience quality, as well as an indication related to the artistic expression potential that can be achieved by the proposed gestural set. Twenty-two students from the department of Audiovisual Arts, Ionian University participated in the subjective evaluation. The measurements took place on a small mixing studio, using the presented system, a personal computer, a digital console and a pair of monitor loudspeakers. Four tracks (see Table 1) were selected for the real time mixing procedure, which were parts of a musical piece composed for the needs of the particular experiment. In all track-cases, a simple reverberation effect was defined as the pre-assigned audio effect.

Table 1: The tracks used in the system's evaluation

Track No.	Track Content	Track Mode
1	Guitar	Monophonic
2	Synthesizer	Stereophonic
3	Drums	Stereophonic
4	Electric Bass	Monophonic

Prior to the subjective measurements, the reproduction system gain (i.e. the digital mixing console and the loudspeakers overall gain) was calibrated. This offered an equal sound pressure level from both loudspeakers at the specific point where the participant would be standing during the experiment. Each participant was given a short introductory brief, explaining the gestures used and the track selection process. Consequently, the participants were instructed to perform a mixing session of the audio material according to their own preferences. The maximum duration of each real-time mixing session was limited to 10 minutes.

At the end of each mixing session, a questionnaire with twelve questions was handed out to each participant. All questions were organized in three sections: a) personal information regarding gender, age, relation with music (listener, composer or performer), b) overall satisfaction evaluation, and c) evaluation of the artistic expression potential. Questions in sections (b) and (c) were scored using a five grade evaluation scale, i.e. 1) minimum, 2) little, 3) average, 4) much, and 5) maximum.The complete list of questions is shown in Table 2.

70% of the participants were listeners, that is they are not capable to compose music or play a musical instrument. The remaining 30% were musicians (either composers or musical instrument performers or both). Moreover, 55% were males and 45% females. Regarding the age group of the participants, 49% were $18-20$ years old, 18% were $20-22$, 14% belonged to the $22-24$ group and the remaining 19% were above 24 years old. Finally, only 9% of the participants did not had any prior experience with non contemporary musical interfaces.

5. RESULTS & DISCUSSION

The graphs in Figure 8 and 9 summarize the results obtained for the system usage satisfaction evaluation and the artistic expression potential respectively. From these results it can be clearly observed that musicians tend to rate higher the gestural interface than the ordinary listeners. More specifically, although all users rated their experience with the proposed system to be over average (score equal to 3), musicians provided a score of ”much convenience” (question 8). Additionally, focusing on the results of Figure 8, there is an indication that the gesture for controlling the spatial positioning of the sources is not as satisfactory as the gesture for controlling the reproduction gain and the audio effect module. Regarding the artistic expression potential of the proposed gestural mixing scheme, it can be observed that musicians rated the system above ”much” (question 12). Also, for every aspect of the presented system, all participants groups provided a score above average.

Focusing on the gesture employed for the reproduction gain control and on the listeners group only, it can be seen that although the user experience obtained a rate close to ”much” score, the artistic expression through the specific gesture was rated close to average. But, if this fact is combined with the obtained scores from the musicians group of participants, then it can be inferred that musicians seem to be more familiar with the used gesture than listeners. Therefore, musicians seem to be better artistically expressed when using the specified gesture for controlling the reproduction gain.

The utilized gesture for spatial positioning control has been rated with the lowest scores among all three gestures considered by both groups of participants. One possible rea-

Table 2: The questions that the participants answered at the system's evaluation

No.	Question
Personal Information	
1	Please state your age group
2	Please state your gender
3	How are you connected to music
4	Prior experience with non contemporary musical interfaces
System's Usage Evaluation	
5	Grade the usage convenience of the process for reproduction gain control
6	Grade the usage convenience of the process for spatial positioning control
7	Grade the usage convenience of the process for audio effect control
8	Grade the usage convenience of the device in general
Artistic Expression Evaluation	
9	Grade the degree of your artistic expression enhancement using the gesture for reproduction gain control
10	Grade the degree of your artistic expression enhancement using the gesture for spatial positioning control
11	Grade the degree of your artistic expression enhancement using the gesture for audio effect control
12	Grade the degree of your artistic expression enhancement by using the device in general

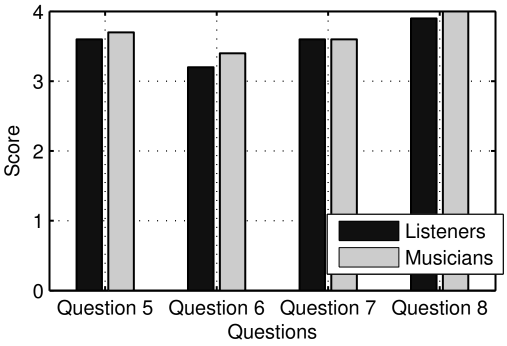

Figure 8: Results for the evaluation of the system's usage satisfaction

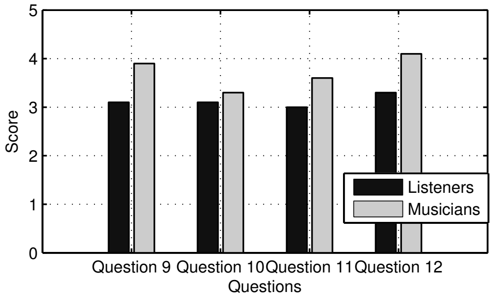

Figure 9: Results for the evaluation of the artistic expression offered by system's usage

son for these low rating results can be the form of the gesture itself, in conjunction with the other gestures: It seems that the majority of users can not easily perform at the same time both the gestures for controlling the reproduction gain and the spatial positioning.

The gesture for audio effect application achieved a"much" average score, regarding the user experience from both participant groups. On the other hand, focusing on the enhancement of the artistic expression potential, the listeners group rated the specific gesture lower than the musicians. This outcome confirms the fact that the musical background seems to have a significant impact on the usage efficiency of the proposed system's interface. Finally, the proposed sound mixing scheme seems in general to achieve an overall score of "much" for both subjective evaluation criteria (i.e. user experience and enhancement of artistic expression potential) and for both participant groups.

6. CONCLUSIONS & FUTURE WORK

In this work a novel gestural user interface for multitrack audio real-time mixing is introduced. The utilized gestures are chosen in order to be as closest as possible to legacy gestures employed by an orchestra conductor. The system implementation is performed using easy to find and program hardware equipment, including proximity sensors, accelerometer, push buttons and a physical computing interface.

A number of subjective evaluation measurements were performed in order to assess the efficiency of the proposed system in terms of the achieved user experience and the enhancement of the artistic expression potential. The results obtained demonstrate that the majority of the users found that the proposed gestural scheme do enhance both the artistic expression and the user experience. From the partial results it can be also concluded that the employed gestures do help the user to express freely and intuitively.

Future investigations towards the proposed gestural scheme evolution may include additional aspects of conducting that are not considered here, such as setting the tempo for music performances, as well as an analytic investigation of slight gestural variations, optimized for sound mixing purposes and applications.

7. REFERENCES

[1] Arduino. Software. http://arduino.cc/en/main/software, 2005. [Online; accessed 20-April-2013].

[2] Arduino. ArduinoBoardMega. http://arduino.cc/en/Main/arduinoBoardMega, 2009. [Online; accessed 20-April-2013].

[3] K. Drossos, R. Kotsakis, G. Kalliris, and A. Floros. Sound events and emotions: Investigating the relation of rhythmic characteristics and arousal. In *Fourth International Conference on Information, Intelligence, Systems and Applications*, IISA 2013. IEEE, July 2013.

[4] K. Drossos, S. Mimilakis, A. Floros, and N. Kanellopoulos. Stereo goes mobile: Spatial enhancement for short-distance loudspeaker setups. In *Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, IIH-MSP, pages 432–435. IEEE, July 2012.

[5] N. Grigoriou, A. Floros, and K. Drossos. Binaural mixing using gestural control interaction. In *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound*, AM '10. ACM, September 2010.

[6] C. Kiefer, N. Collins, and G. Fitzpatrick. Evaluating the wiimote as a musical controller. In *International Computer Music Conference*, ICMC, August 2008.

[7] J. M. Morrell, D. J. Reiss, and T. Stockman. Auditory cues for gestural control of multi-track audio. In *The 17th International Conference on Auditory Display*, ICAD-2011, June 2011.

[8] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human computer interaction: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):667–695, July 1997.

[9] A. Pirhonen, S. Brewster, and C. Holguin. Gestural and audio metaphors as a means of control for mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 291–298, 2002.

[10] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of Audio Engineering Society (JAES)*, 45(6):456–466, June 1997.

[11] V. Pulkki. Localization of amplitude-panned virtual sources ii: Two- and three-dimensional panning. *Journal of Audio Engineering Society (JAES)*, 49(9):753–767, September 2001.

[12] P. Quinn, C. Dodds, and D. Knox. Use of novel controllers in surround sound production. In *Audio Mostly 2009 - the 4th Conference on Interaction with Sound*, AM '09, September 2009.

[13] E. Sato, T. Yamaguchi, and F. Harashima. Natural interface using pointing behavior for human-robot gestural interaction. *IEEE Transactions on Industrial Electronics*, 54(2):1105–1112, April 2007.

[14] R. Selfridge and J. Reiss. Interactive mixing using wii controller. In *130th Audio Engineering Society Convention*, AES Convention. AES, May 2011.

[15] W. O. Site. What is Wii. http://www.nintendo.com/wii/what-is-wii/, 2013. [Online; accessed 13-May-2013].