OPEN ACCESS Freely available online

Testing A Novel Gesture-Based Mixing Interface

MICHAL LECH,[1] *AES Member,* **AND BOZENA KOSTEK,**[2] *AES Fellow*
(mlech@sound.eti.pg.gda.pl) (bokostek@audioacoustics.org)

[1]*Multimedia Systems Department*
[2]*Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications & Informatics, Gdansk University of Technology, Gdansk, Poland*

In this article, a sound-mixing system controlled by hand gestures recognized in a video stream is presented. This novel approach to DAW (Digital Audio Workstation) controlling, was motivated by the limited ergonomics of the computer mouse and keyboard interface, as well as by the influence of audio information visualization on sound mixing. The article reviews existing approaches to gesture controlled audio, and presents the engineered system architecture and approach to gesture sonification. The methodology involved examining the system with the help of professional audio engineers in tests conducted to assess, among others, what influence the visualization of audio parameter values may have on mixing results. The results of a questionnaire and the subjective assessment of the obtained mixes have been given. The system efficiency and gesture recognition reliability have been assessed.

0 INTRODUCTION

Nowadays, large, well-equipped music studio facilities are often substituted by smaller project studios. In such places the mixing software (mixing in the box) approach dominates. The reasons behind this solution are primarily economic. However, many respected sound engineers claim that mixing in the box provides worse results than a mixing desk [1]–[4]. The main cause for this is the difference in quality between the algorithms of mixing software and their corresponding physical equivalents in expensive analog mixing desks [1]–[3]. There are also audio engineers who believe that the quality of algorithms is not a significant factor [5]. According to their observations, the results are affected mainly by the ergonomics of a mixing interface [3] [5].

Graphic interfaces of many audio plugins imitate control panels of their hardware prototypes. Knowing the original device implies knowing the software emulation ad hoc. However, knob-based plugins operated with a mouse and keyboard may have some drawbacks such as that only one parameter can be handled at the same time, or that setting a parameter it not as handy as with a physical switch.

Editing a parameter is also not as handy as using a physical switch. Thus, various compact sound-mixing interfaces have been developed. The faders, knobs, and meters of mixing desks offer better ergonomics in such equipment. The small size is preserved by assigning a chosen potentiometer to the chosen function. However, it can sometimes be necessary to use a mouse to reassign a particular control's settings and its role [6]. Despite the improvement in ergonomics, the price of such interfaces contradicts the idea of the low-cost home recording studio.

Another issue is the visual aspects of Digital Audio Workstations (DAWs). Many sound mixing engineers claim that modifying audio signals with graphic information representing parameter changes leads to worse aesthetics effects [4]. The reason may be in the common physiology of sensory systems and multimodal perception mechanisms, in which sight plays the primary role [7]–[10]. As a result, mixing engineers may be distracted from their original task of audio signal processing by visual information [4]. The visual representation of the changes to an audio signal parameter may also affect the perception of sound at lower levels of the sensory systems. Visual objects may "attract" the person's attention, thus sound sources may seem to be localized closer to the screen. The shift of virtual sound source toward the visual stimulus is often described by the term ventriloquism effect or image proximity effect and occurs unconsciously and regardless of the will of people taking part in tests [8], [11], [12]. Mixing for many years with visual support may also cause a distortion of the cognitive scheme. According to Cohen [13] in ambiguous situations a person can be guided by a cognitive scheme which is formed on the basis of past experience and helps shape the expectations of the subject. However, these schemes are not fully adequate for the reality and distort the object under perception. The solution to the above issues can be seen in eliminating intermediary devices between the engineer and the sound system by employing hand gestures. This would create an opportunity for a greater immersion in the process of sound mixing. Thus, the impact of visual

stimuli on sound perception could be minimized. Such an approach could also improve ergonomics in comparison to the computer mouse and keyboard interface, because when using two hands, an engineer can more easily manage two audio parameters simultaneously.

An additional advantage is that mixing with hand gestures provides conditions in which sound between an engineer and the studio monitors may propagate in a semi-free acoustic field. In such a case, soundproof materials on the floor, ceiling, and walls would eliminate sound colorations [14], [15].

Given the above observations, we have engineered a mixing interface handled by dynamic (i.e., motion) and static (i.e., pose) hand gestures recognized in a video stream. The system has been developed in such a way that mixing operations can be performed both with or without visual support.

First, the article reviews solutions applied in known audio interfaces. Then, the main features of the developed system along with its architecture, Graphic User Interface (GUI) and gesture sonification are provided. Finally, the methodology of the experiments and their results are given.

1 STATE OF THE ART

A review of literature on sound-mixing systems leads us to conclude that none of the well-known solutions provides gestural control of all the key operations of sound mixing. In the work by Marshall et al. [16], the systems that support hand gesture controlling of sound panorama have been reviewed. The majority of the reviewed systems additionally enable the control of parameters associated with reverberation of a virtual space in which the panned audio sources are placed. However, the purpose of the presented systems is to support musicians, not mixing engineers. Gestures which naturally occur while playing a musical instrument can be recognized and used to trigger sound processing effects. Thus, musical performance can be enriched. Another solution in the immersive virtual instrument domain has been proposed by Valbom et al. [17]. The system, called WAVE (Virtual Audio Environment), enables the triggering of music loops or the playing of tones of chromatic scales using hand gestures. The hand motion is transformed into the movement of a virtual wand on the computer screen. To provide three-dimensional (3D) immersion the solution employs virtual reality technologies and 3D sound techniques based on a near-field stereo-sound system coupled with a 4.1 surround-sound system. Berthaut et al. [18] proposed a new solution in the immersive instrument area. They introduced a new hardware control, called Piivert to manipulate the graphic widgets. Piivert is composed of infrared targets placed on its extremity and of pressure sensors located below the thumbs, index fingers, middle fingers, and ring fingers of each hand. Thus, this solution requires a dedicated hardware, attached to the user's body.

The usefulness of gestures in operating the DAW software was noticed by Balin and Loviscach [19]. However, their system utilizes only gestures from the dictionary of movements performed with a computer mouse. Moreover, gestures are used typically to operate editing functions rather than mixing.

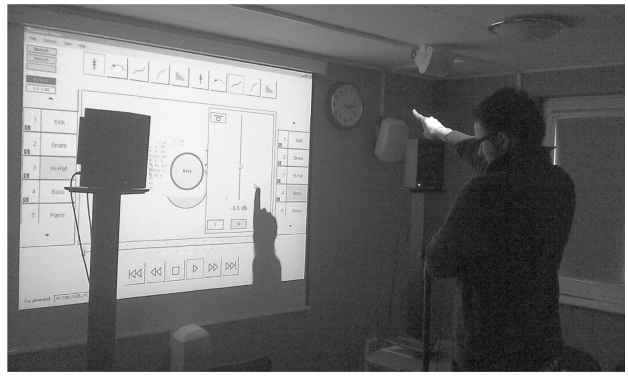


Fig. 1. Placement of system components and the location of the user.

The solution which enables the mixer to step away from a mixing desk or any other physical interface, and handle the process remotely via gestures has been presented by Selfridge and Reiss [20]. The motion of a Wii [21] controller is used to adjust the levels of parameters on a variety of digital audio effects. The authors of the solution examined the possibility of using infrared sensors contained in the Wii for the purpose of gesture-based audio mixing. However, when applied to mixing, infrared diodes introduced limitations to the range of the controller angular motion. Also, it lacked the user requirement to make a free choice of a sound monitoring position. Another serious problem with the Wii controller was the sensitivity of accelerometers. As stated by the authors, movements which were too gentle did not cause the accelerometers to register the motion, and thus no change in parameters took place.

Karjalainen et al. examined three different gesture controllers for the purpose of virtual air guitar playing [22]. They considered datagloves as an expensive solution but one which provides the highest control over hand gestures. In contrast, they said that specialized control sticks supply a variety of low-cost possibilities and they judged video tracking to provide the lowest quality.

It was concluded that controllers, infrared sensors, or accelerometers do not provide sufficient ergonomics to be adopted for sound mixing purposes. Therefore, we engineered a nonobtrusive sound mixing interface in which gestures are recognized purely on the basis of camera stream processing.

2 THE ENGINEERED SYSTEM

2.1 System Overview

The engineered system is composed of a PC, a webcam, a multimedia projector and a screen for the projected image. A camera lens is directed at the projection screen. The whole projected image and the shadows of the user's hands are visible in the captured video stream (Fig. 1).

A user is situated in a sweet spot located between the screen and the multimedia projector, from where he or she

can control the mixing processes via hand gestures. No infrared diodes, infrared cameras, gloves, or markers were needed. The system can be used with either a dual or a multichannel sound system.

The system is based on subtracting the video stream captured by the camera from the image projected by the multimedia projector and locating the hands in the processed output. Both dynamic gestures (motion trajectory) and static gestures (palm shape) are recognized by the system. For the dynamic gesture recognition a fuzzy-rule inference system is used. Static gestures are recognized by Support Vector Machines (SVMs) of a C-Support Vector Classification type. Dynamic gestures are strongly associated with static gestures. Thus, performing the same motion with a different palm shape has various meanings. Moreover, the order in which gestures are performed can represent a gesture category.

The detailed description of the algorithms used in the system can be found in other papers by the authors [23], [24].

2.2 System Architecture

The software of the system has been divided into two parts, that is, the application recognizing gestures and relevant actions, and a gesture dedicated graphic overlay for any DAW software. The communication with the DAW software is based on the MIDI protocol. The graphic overlay receives system actions generated by the gesture recognition application and sends relevant MIDI messages. Native functions, such as changing the track level, playing the session, or soloing the track are handled by the MIDI HUI protocol. The parameters of plug-ins other than the native ones are associated with particular gestures using the MIDI learn function which is provided in the majority of professional DAW systems.

After initializing the gesture-recognition application, a user can set the SVM classifiers separately for the left and right hand. This enables the assignment of audio parameters for each hand independently and modify two parameters simultaneously.

2.3 System GUI and Gesture Sonification

Considering the examination of the influence of audio information visualization on sound mixing, all sound mixing operations can be handled with a GUI that does not provide visual support for audio changes or with full graphic representation of sound modifications. The middle part of the application window in Fig. 2 contains circles representing audio sources. The size of the circle represents the level. The horizontal and vertical positions represent the panorama and equalizer gain, respectively. Directing a hand over the circle with an index finger extended selects the particular audio source. With the audio source selected, hand movements cause respective circle position changes and thus the panorama or equalizer gain can be smoothly adjusted. A similar approach to visualizing mixes has been adopted by Aaron Holladay in an application called Audio Dementia [25]. Every track in a song in this solution has an icon on

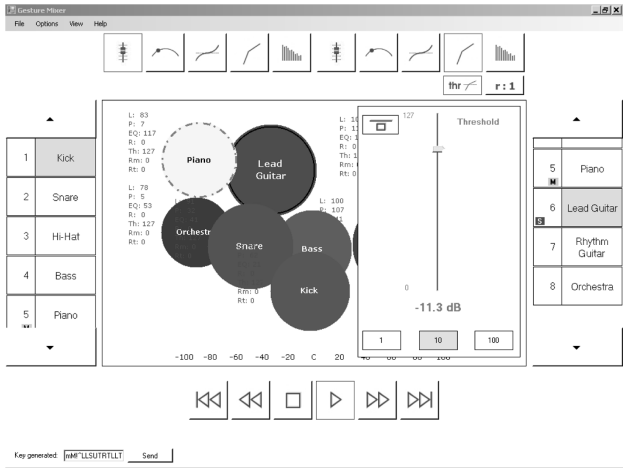


Fig. 2. GUI of the application.

Table 1. Default gesture set

G1	✋	G7	↓
G2	☝	G8	⇔
G3	✌	G9	⇒⇐
G4	→	G10	⇑⇓
G5	←	G11	⇓⇑
G6	↑		

a stage area that represents its volume and pan with respect to a central icon on the stage. Changing the panorama or level is performed by clicking and dragging the track icon. According to the author's words, such an interaction makes music mixing more natural and allows musicians to relax and enjoy the music being created.

The GUI contains menu strips with iconographic representation of all available sound mixing operations (Fig. 2). A user can choose parameters and operations by directing a hand over these icons. Some of these functions can be chosen directly by performing a dynamic gesture with a palm appropriately shaped. The interface can also be entirely managed with a mouse and keyboard.

For the purpose of efficient gesture controlling, the unified gesture dictionary has been created (Tables 1 and 2). Holding the hand flat has no action assigned. Thus, it is possible to comfortably choose mixing parameters or functions by directing a hand over the menu icons. To perform a meaningful gesture, the palm must take one of the shapes presented in Table 1. Dynamic gestures from Table 2 are represented by motion trajectories indicated by single line arrows for one-hand movements and double line arrows for both hands. While training classifiers, a user can define other static gestures. The dictionary has been created in such a way that dynamic gestures are semantically associated with functions. For example, choosing a compression threshold is drawing a capital "T" letter in the air. Every parameter can be modified by moving a hand up or down, for increasing or decreasing its value, respectively.

Table 2. Default gesture-action assignments

ID	Gesture	Default action
1		no action
2		Choosing source (audio signal)
3	⇔	Increasing level
4	⇒ ⇐	Decreasing level
5	→ / ←	Play
6	← / →	Stop
7	→ / ←	Forward (if playing)
8	← / →	Backward (if stopped)
9		Solo / unsolo
10		Mute on / unmute
11	$(\ )_2$	Unsoloing all tracks (performed with both hands)
12	$(\ )_2$	Unmuting all tracks (performed with both hands)
13	↑ / ↓	Increase / decrease chosen parameter value
14		Choosing reverb time for setting
15		Choosing dynamic compression ratio for setting
16		Choosing dynamic compression threshold for setting
17		Choosing shelving equalizer gain

During this motion, an index finger is extended. A flat hand finishes the parameter edition. Each parameter can be modified using one hand only. Thus, two arbitrary parameters can be modified simultaneously. As mentioned earlier, the level, panorama and gain of the shelving equalizer can be adjusted directly by manipulations on circles displayed on the screen.

3 EXPERIMENTS

The experiments were constructed in such a way that the influence of parameter visualization on sound mixing and the ergonomics of the interface in comparison with a mouse and keyboard could be verified.

The experiments were carried out using the engineered interface and the Steinberg Cubase Studio 5 music-production system.

There were two hypotheses formulated, referenced later in the article:

Hypothesis 1.

Visualization of audio signal parameters adversely affects the aesthetic value of the mixes.

Hypothesis 2.

Mixing by hand gestures leads to mixes of a higher aesthetic value than mixing with a mouse and keyboard.

It should be remembered that sound quality is a complex and multilayered phenomenon consisting of a variety of elements and features. Moreover, the relevance and salience of each of quality to be judged depend on the specific case. It is, therefore, very important to design subjective tests in an appropriate way [26].

3.1 Sound mixing methodology

Ten professional mixing engineers were involved in the experiments. The task of each engineer was to mix eight audio tracks with significantly different musical and signal features. Each track contained recordings of a single instrument or of a group of instruments. The instruments were: a bass drum, snare, hi-hat, bass guitar, grand piano, lead guitar, rhythm guitar, and symphonic orchestra. The composition was created by the authors. They played real instruments for the guitar and piano sounds and used plug-ins for sampled sounds. The genre of the music was either instrumental rock or film soundtrack.

None of the engineers was familiar with the provided audio material before the experiments. Each mixer was asked to develop the individual idea for the final qualities of a mix. The aim was to preserve this idea in all mixing and thus ideally obtain an identical mix every time. The engineers were also asked to adopt a fixed methodology for all mixing methods.

In order to examine the influence of parameter visualization on the mixing results and compare the ergonomics of gesture interaction with a mouse and keyboard, the following five methods of sound mixing were considered:

① mixing via gesture using the engineered system, without visual information reflecting audio parameter changes;

② mixing via gesture using the engineered system, with visual information reflecting audio parameter changes provided;

③ mixing using the engineered system, controlled by mouse and keyboard, without visual information reflecting audio parameter changes;

④ mixing using the engineered system, controlled by mouse and keyboard, with visual information reflecting audio parameter changes provided;

⑤ mixing directly using a music production system controlled by a mouse, keyboard, and MIDI controller for parameter editing.

In ⑤, the mixing operations which could be performed by the engineer were limited to the set of operations available during mixing with the engineered system. The motivation for carrying out the experiments based on the five methods presented above has been given in Table 3.

The order of the mixing methods was different for each engineer. Its aim was to eliminate the effect of learning the process which could lead to serial correlation. When finished, each engineer was asked to fill in a questionnaire examining various aspects of the system. The qualities under review were precision, convenience, and intuitiveness. The engineers were also asked to order their own mixes from the best sounding to the worst sounding.

Table 3. Information that can be obtained from various combinations of test pairs

Pair of mixes	Information provided by pair comparison
① ∧ ②	Checking the impact of visual stimuli reflecting audio parameter changes on sound perception
① ∧ ③	Checking ergonomics/precision of the system controlled by hand gestures
① ∧ ④	Control pair
① ∧ ⑤	Analyzed with a pair ① ∧ ⑤, when ① > ⑤ provides information whether the key relevance is given to the impact of visual stimuli on sound perception (① > ⑥) or the ergonomics of the system engineered (⑥ > ①) (MIDI controller provides ergonomics comparable with gesture handling regarding the possibility of simultaneous editing of two parameters)
② ∧ ③	Checking whether it is the way it is controlled (② > ③) or is it the presence of visual stimuli (③ > ②) that has greater influence on mixing results
② ∧ ④	Checking ergonomics/precision of the system controlled by hand gestures
② ∧ ⑤	Comparison of ergonomics of the engineered system controlled by gestures and music production system handled with MIDI controller
③ ∧ ④	Checking the impact of visual stimuli reflecting audio parameter changes on sound perception when controlling the system by mouse and keyboard
③ ∧ ⑤	Control pair (due to significant diversity of experiment conditions (systems) being compared in this pair, it cannot be a basis for inference considered separately)
④ ∧ ⑤	Control pair (due to significant diversity of experiment conditions (systems) being compared in this pair, it cannot be a basis for inference considered separately)
⑤ ∧ ⑤	Checking whether using MIDI controller indeed provides better mixing results than mouse and keyboard

3.2 Experiment conditions

Both the mixing of audio signals and the subjective assessment were conducted in identical conditions in an acoustically adopted conference room with absorptive panels under the ceiling and diffusive wooden panels on the walls. This particular room was chosen due to the proper configuration of the projector, screen, and audio monitors and a short reverb yet natural sound propagation. Yamaha MSP5 studio monitors placed on Ultimate Support MS-45B2 stands were used. The distance between the monitors equaled 1.85 m. The mixing engineer was situated in the sweet spot.

No special time restrictions were placed on mixing. However, the engineers were advised to consider such qualities of a mix that can be obtained within a 25-min mixing session, no matter which method they decided to use. After each mixing method, a 5-min break was taken. Mixing duration varied between one and a half and three and a half hours. Average mixing time per engineer equaled approximately two and a half hours. It has been observed that even if the developed system and the gesture interaction was used for the first time, the duration of the mixing was not longer than the time needed for mixing with DAW.

3.3 Subjective assessment methodology

Subjective evaluation was conducted using a rank order test [27]. The assessed samples were 15-s excerpts of mixes from all five mixing methods. The reason for the 15-s duration was the fact that longer samples would drastically increase tests time and might cause weariness and hearing fatigue. The ranking order of mixes from each engineer was analyzed via pair comparison, according to Table 3.

Information for the confirmation or contradiction of hypothesis 1 is provided by the result of comparing methods ① ∧ ② and ③ ∧ ④. The predominance of the chosen mixing method over the second one in the pair has been figuratively denoted by sign ">" between pair numbers. The confirmation of hypothesis 1 could be inferred based on the result of the comparison according to relation 1.

$$① > ② \wedge ③ > ④ \tag{1}$$

When using the engineered system with only a mouse and keyboard it is necessary to look at the screen in order to choose the system option. Such a constraint exists independently from the option to either activate or deactivate the visual stimuli reflecting parameter changes. When the system is controlled by gestures, it is possible to close one's eyes and perform the operations without involving eye sight also when visual stimuli are provided. This feature may have been used by the mixers and this could be confirmed by the smaller difference in the predominance of manner ① over the manner ② than of manner ③ over the manner ④. Extending the relation (1) with comparison of methods ③ ∧ ⑤, according to relation (2), would enable us to compare the system engineered with the chosen music production software in the context of multimodal perception. The predominance of the method ⑤ over the method ③, with other pairs consistent with the relation 1, could indicate that when using the chosen music production software there exist other factors than the ones researched and they could have a significant impact on the sound mixing results and do not exist when using the developed system.

$$① > ② \wedge ③ > ④ \wedge ③ > ⑤ \tag{2}$$

The insufficient ergonomics of the mouse and keyboard interface during sound mixing might be inferred from relation (3).

$$① > ③ \wedge ② > ④ \tag{3}$$

Adding the pair ② ∧ ⑤ and obtaining results according to relation (4) enables us to exclude the existence of other factors which could have a greater impact on the mixing results than the controlling interface used while mixing with the chosen music production system.

$$① > ③ \wedge ② > ④ \wedge ② > ⑤ \tag{4}$$

A sequence that would explicitly show that mixing without eye sight involvement is superior to mixing supported with graphic message takes form 5. It would also suggest

that a gesture handled interface is more ergonomic than a mouse and keyboard controlled system.

$$①>②\wedge①>③\wedge①>④\wedge①>⑤\wedge$$
$$②>④\wedge②>⑤\wedge③>④\wedge③>⑤\wedge⑤>④ \quad (5)$$

One can notice that in the above relation pairs ② ∧ ③, ③ ∧ ⑤ have not been included. Based on the result of comparing the mixing methods contained in these pairs one cannot confirm or deny the hypotheses. The relation ② > ③, depending on the results of other relations, could only indicate that interface ergonomics have a greater influence on mixing than eye sight involvement. In contrast, the relation ③ > ② could signify that involving eye sight through the influence on sound perception has a stronger connection with the mixing results than the usage of the less ergonomic interface of the mouse and keyboard. Additionally, for the relation ③ > ②, the results of the comparison of the pair ① ∧ ③ provides information whether the predominant role in the process is given to affecting sight (① > ③) or insufficient precision of reflecting hand movements in changes of parameter values (③ > ①). Based on the result of comparison of a pair ③ ∧ ⑤, depending on the outcome of other relations, one can assess whether the greater importance in the mixing process shall be associated with using a MIDI controller or ensuring interaction not affecting eye sight.

The inverse relation to the relation ⑤ > ④ in relation (5) could indicate that obtaining better results while mixing with the system, we have developed was associated with factors other than those under research.

3.4 Analysis of the influence of ergonomics and visualization on mix parameters

For each track of every mix, the audio parameter values have been collected. Panorama, gain of the shelving equalizer and level have been visualized in figures, according to the method of displaying information in full graphic mode, described earlier. In Fig. 3. sample visualizations for one engineer (engineer no. 3) and all five mixing methods have been presented. In Table 4, all parameter values are given.

For the six mixing engineers there were clear differences in the location of audio sources depending on the GUI mode. Mix visualizations of the five engineers among this group revealed that mixing with full visual information support resulted in a greater spread of sources both in the horizontal and vertical axis. This reflected the broader panorama and more intensive use of the shelving equalizer, respectively. This phenomenon occurred irrespective of whether the interaction was through gestures or the mouse and keyboard. One could regard such an outcome as a surprise, thinking that visual support of source displacement should result in easier and thus earlier perception of parameter change. Conversely, it turned out that when not supported by visualization and displayed parameter values, the engineers seemed to devote much more attention to the sound balance. In fact, what looked balanced in the visualizations turned out to be imbalanced in terms of audio assessment. However, changes among mixes in the remaining parameters made it impossible to associate a smaller spread of sources with greater aesthetic value regarding statistic significance.

3.5 The evaluation of the degree of visual involvement in the process of sound mixing

Nine of the 10 engineers confirmed in the questionnaire that in at least one of the mixing methods sight was involved to a smaller extent (in comparison with other means), that is, it was easier to focus on the sound. Eight of them considered the engineered system, handled by gestures in limited GUI mode, as enabling them to focus on the sound better. For six persons in this group handling the system with a mouse and keyboard did not prevent them from recognizing that the system involved sight to a smaller extent. Two of them also considered the DAW software to involve sight to a smaller extent. This may be associated with the intensive use of the MIDI controller and keeping operations performed by mouse and keyboard to an absolute minimum. Another reason may be due to considering the method of displaying information in the DAW software as involving sight to a smaller extent than the method adopted in the engineered system. One of the mixing engineers considered the methods employing gesture interaction as enabling him to focus better on the sound, regardless of the presence or absence of visual information. It can be associated with the fact that the system has been designed in such a way that it is possible to choose and modify most of the parameters with eyes closed. This person also considered the DAW software as enabling him to focus better on the sound. For one engineer, the engineered system involved sight to a smaller extent only when handled by a mouse and keyboard.

3.6 Evaluation of the repeatability of operation performance

Before the subjective assessment of their own mixes, the engineers had been asked in the questionnaire whether, in his or her opinion, it was possible to create an identical mix each time. The answer to this question was positive for half of the engineers (engineers no.: 1, 2, 4, 6, and 7). The analysis of the visualizations of the mixes with registered values of all parameters, and an auditory evaluation enabled us to conclude that none of the engineers obtained mixes similar to such an extent that distinguishing them would cause difficulties. The fact that half of the engineers were convinced about obtaining identical mixes, before listening to them, and tendencies observed within the visualizations may suggest that the means of interacting with the system, as well as the specificity of perception can indeed affect the choice of audio parameter values while mixing. Regardless of a positive or negative answer, none of the engineers had any difficulties in ordering the mixes with regard to aesthetic value.

After listening to the mixes, the engineers were asked to indicate the ones which they found the most different. In this assessment, it was possible to choose both better or worse

Table 4. Values of MIDI controllers for all parameters of a mix of engineer no. 3; (a) handling by gestures / limited GUI, (b) handling by gestures / full GUI, (c) handling by mouse and keyboard / limited GUI, (d) handling by mouse and keyboard / full GUI, (e) direct use of DAW software

(a)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	69	127	0	0	0
Snare	100	64	61	127	0	0	0
H-H	89	64	66	127	0	0	0
Bass	96	64	70	127	0	0	0
Piano	88	40	64	127	0	0	0
Lead	100	76	64	127	0	12	29
Rth.	85	82	64	127	0	0	0
Orch.	80	64	64	127	0	29	25

(b)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	60	51	127	0	0	0
Snare	90	41	48	127	0	0	0
H-H	94	37	120	127	0	0	0
Bass	86	90	82	127	0	0	0
Piano	86	34	42	127	0	0	0
Lead	95	60	45	127	0	25	47
Rth.	68	98	31	127	0	0	0
Orch.	68	63	75	127	0	27	52

(c)

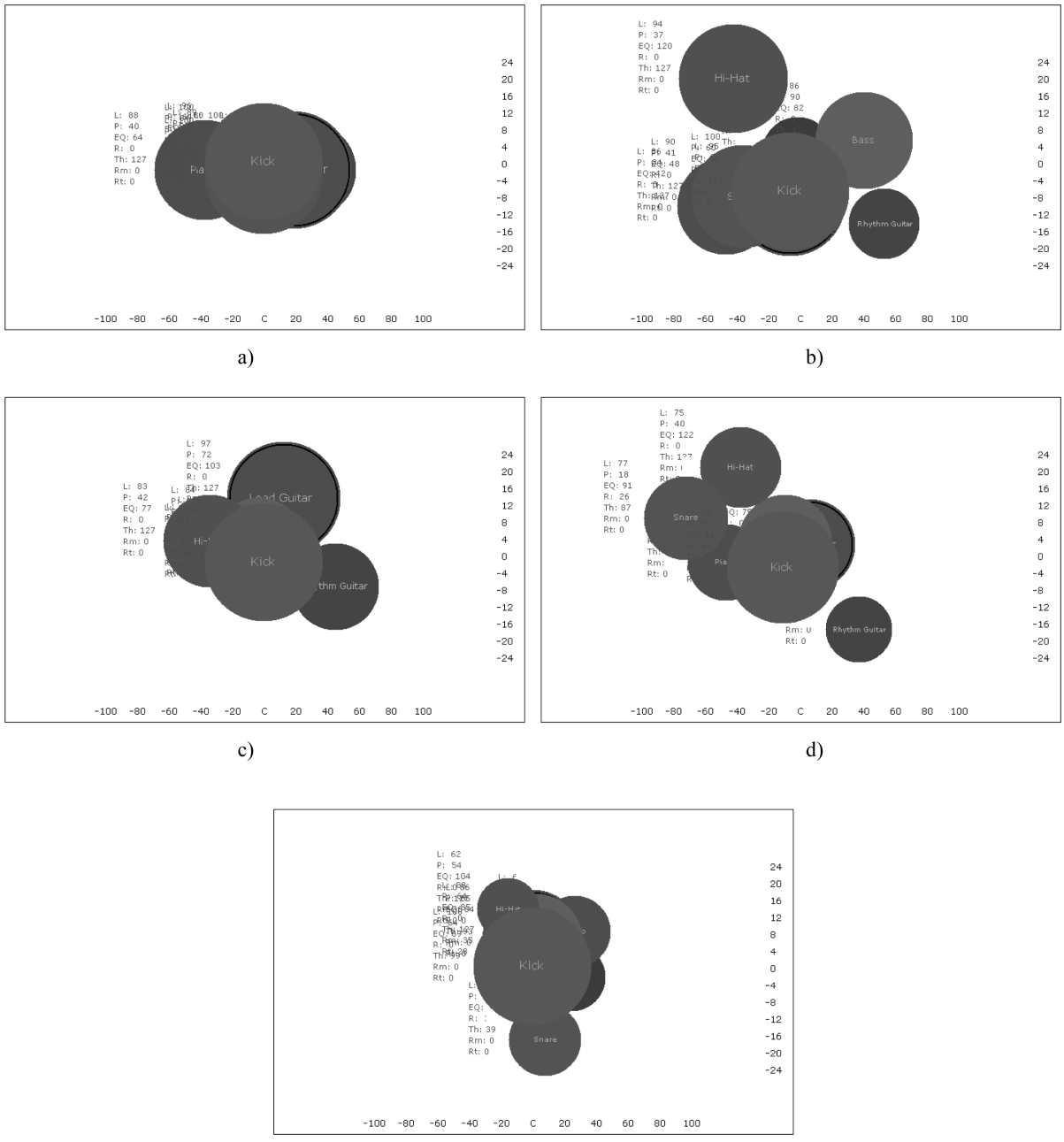
	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	64	74	22	0	0
Snare	79	63	69	127	0	0	0
H-H	83	42	77	127	0	0	0
Bass	84	62	75	118	12	0	0
Piano	84	60	65	127	0	0	0
Lead	97	72	103	127	0	13	40
Rth.	79	93	49	127	0	0	0
Orch.	77	64	64	127	0	32	28

(d)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	96	57	61	50	68	0	0
Snare	77	18	91	87	26	0	0
H-H	75	40	122	127	0	0	0
Bass	83	58	77	127	0	0	0
Piano	72	34	64	127	0	0	0
Lead	82	68	75	127	0	23	25
Rth.	65	88	23	127	0	0	0
Orch.	63	61	60	127	0	40	60

(e)

	Lev	Pan	EQ	Thr	Rat	Mix	Tim
Kick	100	64	69	127	0	0	0
Snare	69	69	24	39	39	0	0
H-H	62	54	104	127	0	0	0
Bass	86	65	84	127	0	0	0
Piano	69	81	90	127	0	0	0
Lead	88	64	85	127	0	35	28
Rth.	61	73	58	127	0	0	0
Orch.	66	80	62	52	6	33	57



a) b) c) d) e)

Fig. 3. Visualizations of mixes of engineer no. 3: (a) handling by gestures/limited GUI, (b) handling by gestures/full GUI, (c) handling by mouse and keyboard/limited GUI, (d) handling by mouse and keyboard/full GUI, and (e) direct use of DAW software.

mixes. The results have been presented in Fig. 4. In Fig. 5, the distribution of answers to the question regarding the reason for differences between the mixes has been given. The answers have been collated in Table 5.

Among the factors categorized in, the engineers mentioned weariness resulting from insufficient ergonomics (one person) and the running order of the mixing methods (two people). Four engineers were not able to identify the cause of the differences.

Whereas for the majority of engineers the selections of differing mixes and indications of the reasons for the differences reflected the lowered grades of aesthetic value, for one person (engineer no. 3), surprisingly, this relation was inversed. The engineer indicated a mix obtained employing gesture interaction in full GUI mode and a mix obtained using mouse and keyboard with the system in limited GUI mode as substantially different from other mixes. As a reason for such an outcome the engineer mentioned insufficient precision, lack in convenience, and presence of visual information when mixing by gestures. However, both mixes were assessed as sounding the best. Based on such a result one can conclude that better ergonomics do not necessarily mean better aesthetic results.

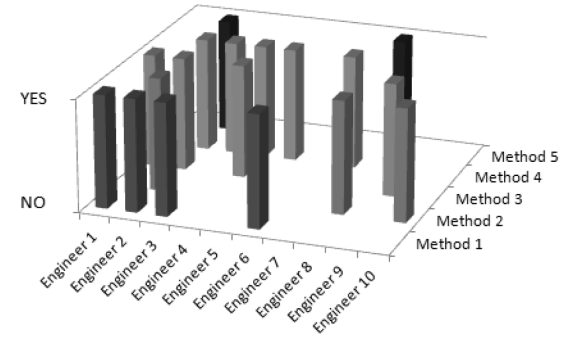


Fig. 4. Distribution of mixes considered different from others within each engineer's own mixes.

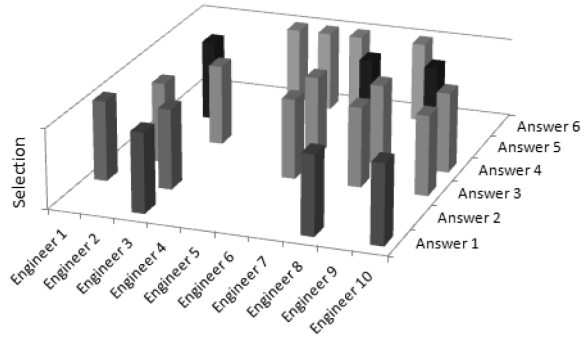


Fig. 5. Distribution of answers to the question about reasons for differences between mixes.

Table 5. Answers of the engineers to question regarding the reason for differences between mixes

No.	Answers to the question: "What, in your opinion, was the reason for obtaining differences between mixes?"
1	Insufficient precision of the engineered system when handling by gestures
2	Presence of visual information in DAW
3	Possibility of mixing without visual support
4	Lack in convenience of the engineered system when handling by gestures
5	Others
6	Hard to say

3.7 Assessment of the aesthetic value of the mixes

The ranking given to particular mixes by the engineers has been analyzed considering medians and presented in a box-and-whisker diagram (Fig. 6) and in Table 6. It was checked that ranks did not correlate with the order of methods. No correlation was found between the ranks given to the mixes from the manner involving direct use of DAW software and the degree of proficiency in handling this software.

The obtained rank distributions for each mixing method have been analyzed in terms of statistic significance using the Friedman test. The test statistics have been given in Table 7. Obtaining $p > 0.05$ did not enable us to disregard the zero hypothesis stating no differences of mean values between the mixing methods.

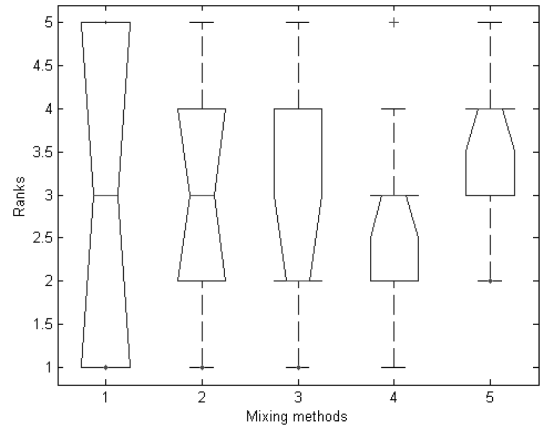


Fig. 6. Box-and-whisker plot for the assessment of aesthetic values of mixes for various mixing methods.

Table 6. Ranks of aesthetic value given by engineers to their own mixes obtained using various mixing methods: Method 1 – handling by gestures / limited GUI, Method 2 – handling by gestures / full GUI, Method 3 – handling by mouse and keyboard / limited GUI, Method 4 – handling by mouse and keyboard / full GUI, and Method 5 – direct use of DAW software, (1 – worse sounding, 5 – best sounding)

	Method 1	Method 2	Method 3	Method 4	Method 5
Eng. 1	5	2	4	1	3
Eng. 2	2	1	5	3	4
Eng. 3	1	4	5	3	2
Eng. 4	3	5	2	2	4
Eng. 5	5	2	1	3	4
Eng. 6	5	1	2	4	3
Eng. 7	5	3	1	2	4
Eng. 8	1	3	4	5	2
Eng. 9	3	4	2	1	5
Eng. 10	1	5	2	3	4

Table 7. Friedman test statistics for the subjective assessment of mixes

SS Effect	df Effect	MS Effect	SS Error
4.45	4	1.1125	95.05
df Error	MS Error	$\chi 2$	p
36	2.64028	1.79	0.7745

3.8 Assessment of gesture dictionary intuitiveness

Before mixing the audio tracks, the engineers were familiarized with the system gesture dictionary. They performed each gesture several times to learn the action it caused. The information about the performed gesture was displayed in the left upper corner of the screen. Additionally, this gesture label was colored red when a gesture which triggered the change of the parameter was performed. The training took approximately ten minutes for each engineer. The engineers had no problems with remembering the gesture dictionary and appreciated its intuitiveness (Fig. 7).

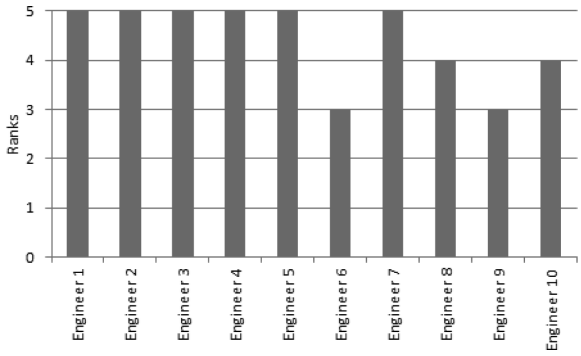


Fig. 7. Intuitiveness ranks given by mixing engineers.

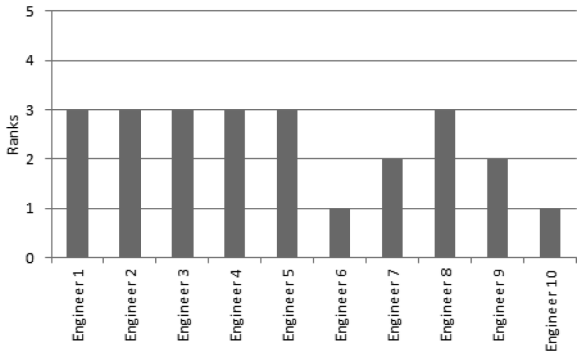


Fig. 8. Convenience ranks given by mixing engineers.

Six engineers assigned the maximum score to the intuitiveness. Two engineers rated the intuitiveness with a score of 3. One of them (engineer no. 6) suggested replacing the V sign with the gesture of a closing hand. Another one (engineer no. 9) indicated a need for introducing more static gestures in order to reduce the number of complex dynamic hand gestures. Two engineers rated the intuitiveness with a score of 4. For one of them (engineer no. 10) the differences between the gestures were too small. Another one (engineer no. 8) stated that the necessity to reform a flat hand in some situations before performing a new gesture lowered the intuitiveness score.

3.9 Assessment of convenience

Grades given by the mixing engineers regarding the convenience of the system when using hand gestures have been presented in Fig. 8. Observations made by the authors during the work of the engineers enabled us to state that low grades were associated mainly with two factors. The first factor consisted in weariness resulting from using hands in a way that prevents them from resting freely, like when using a mouse and keyboard. The second factor was associated with time needed for the system to recognize a gesture after forming the palm into a particular shape. This time equaled approximately 1.5 s and was introduced due to the utilization of an averaging buffer. The length of the buffer corresponded to the 50 elements making up the probability values that a particular shape belonged to one of the gesture classes. Shortening the buffer would reduce the awaiting time at the expense of lowering the gesture recognition efficacy. One of the engineers (engineer no. 9) noted in a free comment box in the questionnaire that improving the system in this aspect would make mixing via hand gestures very convenient.

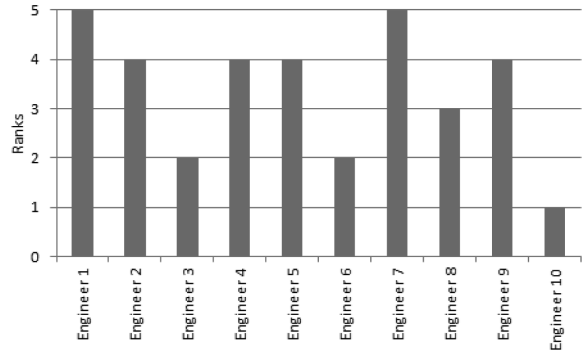


Fig. 9. Ranks given by mixing engineers to precision of parameter value editing when interacting by gestures.

The engineers greatly approved the possibility of controlling two parameters simultaneously. They used this feature mainly for controlling the dynamic compression ratio and threshold at the same time. Similarly, reverb time and mix were also selected this way. The observation was made that the engineers used both hands for controlling two different parameters for the same track, and they did not do it so often for the same parameter of different tracks.

3.10 Assessment of the precision of parameter editing

The degree of precision of editing parameter values has been assessed by the engineers as reflecting that provided by the DAW systems, in particular by the Cubase application, which was managed by the engineered gesture recognition system. However, during the work with the interface in gesture handling mode, it turned out that confirming the set value for the chosen parameter is cumbersome. Namely, changing the palm shape in order to finish editing affected the localization of a hand in the analyzed video stream. Due to the mentioned in Section 3.9 delay between changing the palm shape and recognizing a new gesture it introduced a slight change in the predicted parameter value. This feature was considered an impediment by eight engineers and resulted in a lower precision score (Fig. 9). Two engineers (engineers no. 1 and no. 7) were able to appropriately compensate the change of palm position in the image during a gesture change with subtle movements of the hand. Thus, the parameter values they set reflected the predicted values every time. Among the eight engineers who did not give a maximum score, six rated at least one of the two mixes obtained by mixing via gestures better than the mixes obtained using the DAW software.

3.11 Assessment of the system efficiency

The system efficiency is the parameter which can highly influence the convenience and precision of the parameters edition. Therefore, some tests were performed to measure the execution time of various algorithmic operations. The system run on a computer equipped with Intel Core 2 Duo

Table 8. Performance of the SVM employed in the system for static gesture recognition.

	Left Hand	Right hand
Min. efficacy (%)	66.67	65.83
Max. efficacy (%)	100.00	100.00
Average efficacy (%)	95.68	94.65
Median (%)	98.33	97.50
Average training time (ms)	6435	6598
Average validation time (ms)	197	203

P7350 2.0 GHz processor and DDR2 400 MHz RAM memory. The Windows Vista 32-bit system was used for testing. Due to the multimedia projector operating mode, the screen resolution equaled 1024 × 768 pixels. The video stream captured from the camera was 320 × 240 pixels. During the tests, the users were asked to perform a continuous up-down motion with a flat palm.

The obtained 22 FPS frame rate enabled the mixing engineers to use the system without noticeable latency when performing dynamic gestures and editing parameters value. However, as mentioned in Section 3.9, the averaging buffer used for the probabilities of the proper static gesture recognition, introduced the delay between forming a palm into a particular shape and the actual system reaction. A proposed solution to this shortcoming is a trackbar in the application GUI- used to shorten the buffer.

3.12 Assessment of the gesture recognition reliability

3.12.1 Static gesture recognition efficacy

Eighteen persons took part in the tests of the static gesture recognition efficacy. The SVM-based recognition module was tested for three static gestures presented in Table 1. For some system functions handling, the OK gesture was also added to this set. Each gesture was performed with three different motion trajectories, that is, moving a hand from the left to right side alternately, moving a hand up and down alternately, and moving a hand in a gesticulation of a circle drawing. For each trajectory, 30 camera frames representing a pose were collected, and a leave-one-out cross-validation method was used for testing. A validation set was a collection of samples representing a particular motion trajectory of a particular person. Sets of samples within two other trajectories for other persons constituted a training set. Such a method allowed to examine the generalization of a classifier. In Table 8, some qualities of the classifier derived from the tests, are presented.

It should be noticed that the mentioned averaging buffer causes further increase of static gesture recognition reliability.

3.12.2 Dynamic gesture recognition efficacy

The recognition of dynamic gestures given in Table 1 was examined in tests involving 20 persons. Each of them was asked to repeat each gesture 18 times. Among these 18 repetitions, 10 middle gesture representations were chosen. The four beginning and four ending gestures were rejected as those needed to familiarize with making each gesture or due to the test performer's weariness leading to human mistakes. No special restrictions like moving a hand absolutely straight in a particular direction or forming a particular shape with a palm were imposed. The results are presented in Tables 9 and 10. Zero values have been excluded for better readability.

Table 9. The efficacy of one hand dynamic gesture recognition (see Table 1 for symbols of gestures).

	G4	G5	G6	G7
G4	97.0		1.3	1.7
G5		96.2	2.1	1.7
G6			100.0	
G7	0.8	0.4		98.8

Table 10. The efficacy of both hands dynamic gesture recognition (see Table 1 for symbols of gesture G6 – G11, G12 – hand steady)

	G6	G7	G8	G9	G10	G11	G12
G8			100.0				
G9				100.0			
G10	0.1				99.9		
G11		0.1				99.9	
G12							100.0

High efficacy of dynamic gesture recognition enabled the mixing engineers to use the system without any noticeable shortcomings.

4 CONCLUSIONS

In this article, a novel gesture-based interface for sound mixing has been presented. The novelty of the presented system lies in its possibility to control all mixing operations of the chosen DAW software by hand gestures only. The experiments show that mixing audio signals using hand gestures instead of physical interfaces like a mouse or a keyboard is possible and intuitive. It was proved that visualizing audio parameter values can affect the decision process during sound mixing. Mixing with visual support has led to broadening the panorama and a more intensive use of the shelving equalizer in more than half of the cases. The results of listening tests prove that employing hand gesture interaction in sound mixing produces mixes that are not worse regarding aesthetic value than the ones obtained using DAW software handled by a mouse, keyboard, and MIDI controller. The mixes resulting from mixing via gestures without visual support were more vivid than mixes obtained directly using the DAW software. This appealed to many engineers and as a result they assigned more maximum scores to these mixes than to the ones from Cubase. However, at the same time the vividness was considered unpleasant by some and this resulted in lower minimum scores.

The presented system, thanks to providing the possibility of immersion into the mixing process, could also advance the development of new approaches to sound mixing, with

more emphasis on artistic aspects than in traditional methods. According to the record producer and sound engineer Gareth Jones, "it is very easy to forget about the human factor and artistic aspects in the era of modern studio technology". The need for creating novel solutions has been mentioned by sound engineers [28]. The engineered system seems to fit well into the idea of creating novel touch-less human-computer interfaces, and as suggested by Steve Lillywhite may allow for "listening with your ears, not your eyes" [29] in the mixing process.

5 ACKNOWLEDGMENTS

The research was funded by the project No. POIG.01.03.01-22-017/08, entitled "Elaboration of a series of multimodal interfaces and their implementation to educational, medical, security and industrial applications". The project is subsidized by the European regional development fund and by the Polish State budget.

6 REFERENCES

[1] R. Campbell, "Behind the Gear," *Tape Op - The Creative Music Recording Magazine*, no. 81, pp. 12–13 (Feb / Mar 2011).

[2] J. Congleton, "Sound Fascination," Tape Op - The Creative Music Recording Magazine, no. 81, pp. 14–18 (Feb / Mar 2011).

[3] S. Litt, "Scott Litt," Tape Op - The Creative Music Recording Magazine, no. 81, pp. 20–25 (Feb / Mar 2011).

[4] B. Owsinski, *The Mixing Engineer's Handbook: Second Edition*, Boston: Thomson Course Technology PTR (2006).

[5] G. Jones, "Sonic State Interview: Gareth Jones," 2011. [Online]. Available: http://www.youtube.com/watch?v=AXB8dSnNHLc. [Accessed 11 Nov. 2011].

[6] Steinberg, "CC121 Advanced Integration Controller - Operation Manual," 2011. [Online]. Available: ftp://ftp.steinberg.net/Download/Hardware/CC121/CC121_OperationManual_en.pdf. [Accessed October 24 2011].

[7] T. Carr, "A Multilevel Approach to Selective Attention," in *Cognitive Neuroscience of Attention*, M. Posner (ed.), New York, The Guilford Press, pp. 56–70 (2004).

[8] F. Avanzini, "Interactive Sound," in *Sound to Sense Sense to Sound – A State of the Art in Sound and Music*, D. Rocchesso and P. Polotti, Eds., Information Society Technologies, pp. 302–345 (2007).

[9] A. Dobrucki, P. Plaskota, P. Pruchnicki, M. Pec, M. Bujacz, P. Strumillo, "Measurement System for Personalized Head-Related Transfer Functions and Its Verification by Virtual Source Localization Trials with Visually Impaired and Sighted Individuals," *Journal of the Audio Engineering Society*, vol. 58, no. 9, pp. 724–738 (2010).

[10] S. Merchel, E. Altinsoy, M. Stamm, "Touch the Sound: Audio-Driven Tactile Feedback for Audio Mixing Applications," *Journal of the Audio Engineering Society*, vol. 60, no. 1/2, pp. 47–53 (2012).

[11] B. Kunka, B. Kostek, "Objectivization of audio-video correlation assessment experiments," *Archives of Acoustics*, vol. 37, no. 1, pp. 63–72 (2012).

[12] J. Vroomen, B. de Gelder, "Perceptual Effects of Cross-modal Stimulation: Ventriloquism and the Freezing Phenomenon," *The handbook of multisensory processes*, vol. 3, no. 4, pp. 1–23 (2004).

[13] J. Cohen, G. Aston-Jones, M. Gilzenrat, "A Systems-Level Perspective on Attention and Cognitive Control," in *Cognitive Neuroscience of Attention*, M. Posner (ed.), New York , The Guilford Press, pp. 71–90 (2004).

[14] F. Everest, *Master Handbook of Acoustics*, 4th ed., New York : McGraw-Hill / TAB Electronics (2001).

[15] P. Evjen, J. Bradley, S. Norcross, "The effect of late reflections from above and behind on listener envelopment," *Applied Acoustics*, no. 62, pp. 137–153 (2001).

[16] M. Marshall, J. Malloch, M. Wanderley, "Gesture Control of Sound Spatialization for Live Musical Performance," in *Gesture Based Human Computer Interaction and Simulation*, M. Sales Dias (ed.), Berlin , Springer, pp. 227–238 (2009).

[17] L. Valbom, A. Marcos, "WAVE: Sound and music in an immersive environment," *Computers & Graphics*, vol. 29, no. 6, pp. 871–881 (2005).

[18] F. Berthaut, M. Desainte-Catherine, M. Hachet, "Interacting with 3D Reactive Widgets for Musical Performance," *Journal of New Music Research*, vol. 40, no. 3, pp. 253–263 (2011).

[19] W. Balin, J. Loviscach, "Gestures to Operate DAW Software," *AES 130th Convention*, London (2011).

[20] R. Selfridge, J. Reiss, "Interactive Mixing Using Wii Controller," *AES 130th Convention*, London (2011).

[21] "Wii" Nintendo2011. [Online]. Available: http://www.wii.com. [Accessed October 27 2011].

[22] M. Karjalainen, T. Maki-Patola, A. Kanerva, A. Huovilainen, "Virtual Air Guitar," *Journal of the Audio Engineering Society*, vol. 54, no. 10, pp. 964–980 (2006).

[23] M. Lech, B. Kostek, "Hand gesture recognition supported by fuzzy rules and Kalman filters," *Int. J. Intelligent Information and Database Systems*, vol. 6, no. 5, pp. 407–420 (2012).

[24] M. Lech, B. Kostek, "Fuzzy Rule-based Dynamic Gesture Recognition Employing Camera & Multimedia Projector," *Advances in Intelligent and Soft Computing, Advances in Multimedia and Network Information System Technologies*, vol. 80, pp. 69–78 (2010).

[25] A. Holladay, "Audio Dementia: A Next Generation Audio Mixing Software Application," *118th AES Convention, Barcelona* (2005).

[26] J. Blauert, U. Jekosch, "A Layer Model of Sound Quality A Layer Model of Sound Quality," *Journal of the Audio Engineering Society*, vol. 60, no. 1/2 (January/February 2012).

[27] N. Zacharov, J. Huopaniemi, M. Hamalainen, "Round robin subjective evaluation of virtual home theatre sound systems at the AES 16th international conference," *AES 16th International Conference on Spatial Sound Reproduction* (1999).

[28] M. Morrell, J. Reiss, "Auditory Cues for Gestural Control of Multi-Track Audio," *17th International Conference on Auditory Display (ICAD-2011)*, Budapest (2011).

[29] San Francisco 133 Audio Eng. Soc. Convention Opening Ceremonies, Keynote Speaker: Steve Lillywhite: "Listen with Your Ears, Not Your Eyes" (http://www.aes.org/events/133/productdesign/?ID=3172) (2012)

THE AUTHORS

Michal Lech

Bozena Kostek

Michal Lech received his M.Sc. degree in 2007 from the faculty of Electronics, Telecommunications and Informatics, Technical University of Gdansk. The subject of his thesis concerned developing the application for automatic pitch detection and correction of detuned singing.

As a researcher and software developer he has been involved in various research projects. His scientific interests are associated with image processing and human interaction employing interfaces based on artificial intelligence. He is also interested in studio recording and music production.

Currently he is a Ph.D. student of the Multimedia Systems Department.

●

Bozena Kostek holds professorship at the Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology (GUT), Poland. She is now Head of the Audio Acoustics Laboratory. She is a Fellow of the Audio Engineering Society. She received her M.Sc. degree in Sound Engineering from the Technical University of Gdansk (1983) and her second M.Sc. in Organization and Management in 1986. In 1992, she supported her Ph.D. degree with honors from the Technical University of Gdansk. In March 2000 she supported her D.Sc. degree at the Institute of Research Systems of the Polish Academy of Sciences in Warsaw. In 2005 she was granted the title of professor from the President of Poland.

Her research activities are interdisciplinary, however the main research interests focus on cognitive bases of hearing and vision, music information retrieval, musical acoustics, studio technology, Quality-of-Experience, human-computer-interaction (HCI) as well as applications of soft computing and computational intelligence to the mentioned domains. She has published more than 450 scientific papers, three books and a dozen of book chapters.

In 1991, she helped to form the Polish Section of the Audio Engineering Society, and since then has served as a member of the Committee. In 2003, 2005 and 2009 she was elected Vice-President of the Audio Engineering Society for Central Europe, and in 2007 and 2011 she was elected as Governor of the AES. She is now Editor-in-Chief of JAES.