GESTURE-CONTROLLED SOUND MIXING SYSTEM WITH A SONIFIED INTERFACE

Michał Lech
Multimedia Systems Dept.,
Faculty of Electronics, Telecommunications and Informatics
Narutowicza 11/12, 80-233 Gdańsk, Poland
mlech@sound.eti.pg.gda.pl

Bożena Kostek
Audio Acoustics Laboratory
Faculty of Electronics, Telecommunications and Informatics
Narutowicza 11/12, 80-233 Gdańsk, Poland
bokostek@audioakustyka.org

ABSTRACT

In this paper the Authors present a novel approach to sound mixing. It is materialized in a system that enables to mix sound with 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. To check the hypothesis that the mixing process needs only an auditory display, the influence of audio information visualization on sound mixing and the ergonomics of the system usage in comparison to a mouse and keyboard interface are tested and the results of this study are presented.

1. INTRODUCTION

Sonification, audification, and auditory interfaces are terms that are conceptually interrelated and often exchanged within the auditory display area. It may be found that first definitions of sonification and auditory displays referred to situation in which their task was simply to convey a message in the form of sound, e.g. caution, warning, or danger. From its broader point of view the auditory display field encompasses perception and technology. Another definition from 1997 says: "Sonification is the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation." [1]. The above given arguments support the interpretation that by nature, the area of sonification is interdisciplinary.

In this paper the authors present a concept of an auditory display that facilitates sound mixing. Sound mixing is the process by which multiple recorded sound tracks are combined into one or more channels. The process may be considered as creating sound rather than only editing it. There are two basic ways to mix sound. The first, so called ("out of the box") is using audio mixing consoles with knobs and faders, and the second, so called ("in the box"), is using DAW stations (Digital Audio Workstations) equipped with mixing software. However, in either case the ultimate aim is to obtain a very impressive and realistic sound. To some extent, both mentioned methods rely on visual information. This applies especially to DAW-based mixing, which may be seen as operating against the rule of "listen with your ears, not your eyes". Moreover, a user of a DAW station is limited by the ergonomics of a computer interface (i.e. mouse, keyboard, touch screen). The Authors present a completely novel approach to sound mixing. They materialized it in a system that enables to mix sound with 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. The influence of audio information visualization on sound mixing and the ergonomics of the system usage in comparison to a mouse and keyboard interface are tested and the results of this study are presented.

It can be observed that 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 purely economic. However, many respected sound engineers claim that mixing in the box provides worse results than a mixing desk [2] [3] [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 [2] [3]. There are also audio engineers who believe that the quality of algorithms is not a significant factor. According to their observations, the results are affected mainly by the ergonomics of a mixing interface [4]. When using a mouse and keyboard, the most important limitation is that only one parameter can be handled at any one time. Editing a parameter is also not as convenient as using a physical switch. Thus, various compact sound mixing interfaces have been developed.

Another issue is the visual aspects of DAWs. Many sound mixing engineers claim that modifying audio signals with graphical information representing parameter changes leads to worse aesthetics effects [5]. The reason may be in the common physiology of sensory systems and multimodal perception mechanisms, in which sight plays the primary role [6] [7] [8] [9]. As a result, mixing engineers may be distracted by visual information from their original task which is audio signal processing [5]. 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. As an example, one can mention the ventriloquism effect which occurs unconsciously and regardless of the will of people taking part in tests [7] [10] [11].

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. We would like to claim that sound engineers need only an auditory display in the mixing process. Such an approach could also improve ergonomics in comparison to the computer mouse and keyboard interface, because when using two hands, an engineer can manage two audio parameters simultaneously.
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 paper presents some of the gesture controlled interface solutions applied to the audio domain. Then, the architecture of the developed system along with the main algorithms, GUI and gesture sonification are described. The methodology of the subjective tests involving both mixing engineers and expert listeners is given. Analysis of experimental results obtained is also provided. The paper concludes with a short summary of main findings and aspects discussed in the paper.

2. 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. [12], 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. Another solution in the immersive virtual instrument domain has been proposed by Valbom et al. [13]. 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 a computer screen. As stated by the authors, "to control the system, the user moves 3D sensors (receivers) built into the mice gripped in each hand to move the wands in the display, and to trigger musical objects". To provide 3D immersion the solution employs virtual reality technologies and three-dimensional sound techniques based on a near-field stereo-sound system coupled with a 4.1 surround-sound system. Currently, a new solution was proposed in the immersive instrument domain that deals with the simultaneous control and visualization of musical processes. The project presented by Berthaut et al. [14] is called interacting with 3D reactive widgets for musical performance. They introduced a new hardware control, called Piivert to manipulate the graphical 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 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 [15]. The system utilizes a Wii controller [16]. The motion of the 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. Infrared diodes and cameras are often used as the basis for various object recognition methods [17] [18]. 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 [15]. 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 non-obtrusive sound mixing interface in which gestures are recognized purely on the basis of camera stream processing.

3. THE ENGINEERED SYSTEM

3.1. 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. 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 are needed. The system can be used with either a dual or a multi-channel sound system, as presented in fig. 1. A video presentation of the system is available online [28].

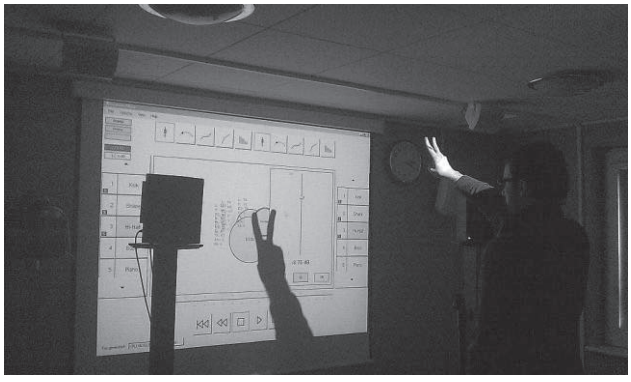


Fig. 1. Placement of system components and the location of user

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. Dynamic gestures are closely 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.

3.2. System architecture

The software of the system is divided into two parts, i.e. the application recognizing gestures and relevant actions, and the application being a gesture dedicated graphical overlay for any DAW software. The communication with the DAW software is based on the MIDI protocol. The graphical 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 (Support Vector Machine) classifiers separately for the left and right hand. This enables the assignment of audio parameters for each hand independently and modifying two parameters simultaneously.

3.3. Algorithms

3.3.1. Image processing

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. Before subtraction, both graphic streams are processed to obtain the most similar characteristics. The camera frame measuring 320 x 240 pixels is perspective corrected. Next, the perspective corrected image is cropped to obtain a view of the projected image only. Furthermore, to eliminate the influence of lighting and the vignette effect introduced by lens, the color of each pixel p_{ij} of the i x j image obtaining pixel p'_{ij}, is adjusted accordingly to Eq. (1),

$$p'_{ij} = p_{ij} + p^{c}_{ij} \tag{1}$$

where $p_{ij}^{\ c}$ given by Eq. (7) is a pixel of one of 5 color calibration images created during the automatic calibration phase. During this process 5 images of which all pixels in RGB color space have values $p^{red} = [255, 0, 0]$, $p^{green} = [0, 255, 0]$, $p^{blue} = [0, 0, 255]$, $p^{white} = [255, 255, 255]$ and $p^{black} = [0, 0, 0]$, respectively, are displayed by the multimedia projector. From each of the images the corresponding camera frame is subtracted, thus giving images consisting of pixels $p_{ij}^{\ r}$, $p_{ij}^{\ g}$, $p_{ij}^{\ b}$, $p_{ij}^{\ wh}$ or $p_{ij}^{\ bk}$, which denote the outputs for red, green, blue, white and black calibration images, respectively, accordingly to Eqs. (2) – (6).

$$p^{r}_{ij} = \begin{cases} p^{red} - p_{ij} & \left| p^{red} - p_{ij} \geq 0 \right. \\ 0 & \left| p^{red} - p_{ij} < 0 \right. \end{cases} \tag{2}$$

$$p^{g}_{ij} = \begin{cases} p^{green} - p_{ij} & \left| p^{green} - p_{ij} \geq 0 \right. \\ 0 & \left| p^{green} - p_{ij} < 0 \right. \end{cases} \tag{3}$$

$$p^{b}_{ij} = \begin{cases} p^{blue} - p_{ij} & \left| p^{blue} - p_{ij} \geq 0 \right. \\ 0 & \left| p^{blue} - p_{ij} < 0 \right. \end{cases} \tag{4}$$

$$p^{wh}_{ij} = \begin{cases} p^{white} - p_{ij} & \left| p^{white} - p_{ij} \geq 0 \right. \\ 0 & \left| p^{white} - p_{ij} < 0 \right. \end{cases} \tag{5}$$

$$p^{bk}_{ij} = \left| p^{black} - p_{ij} \right| \tag{6}$$

$$p^{c}_{ij} = \begin{cases} p^{r}_{ij} \left| \; r_{ij} > g_{ij} + t^{rgb} \wedge r_{ij} > b_{ij} + t^{rgb} \right. \\ p^{g}_{ij} \left| \; g_{ij} > r_{ij} + t^{rgb} \wedge g_{ij} > b_{ij} + t^{rgb} \right. \\ p^{b}_{ij} \left| \; b_{ij} > r_{ij} + t^{rgb} \wedge b_{ij} > g_{ij} + t^{rgb} \right. \\ p^{wh}_{ij} \left| \; r_{ij} > t^{wh} \wedge g_{ij} > t^{wh} \wedge b_{ij} > t^{wh} \right. \\ -p^{bk}_{ij} \left| \; r_{ij} < t^{bk} \wedge g_{ij} < t^{bk} \wedge b_{ij} < t^{bk} \right. \end{cases} \tag{7}$$

The symbols r_{ij}, g_{ij}, b_{ij} in Eq. (7) indicate red, green and blue components of the pixel p_{ij}, respectively. The thresholds t^{rgb}, t^{wh} and t^{bk} are used to distinguish between: red, green, blue components and white images and black ones, respectively. The default threshold values were determined empirically and equal 50, 180, 80, for t^{rgb}, t^{wh} and t^{bk}, respectively.
The projected image, retrieved from the computer, is scaled to ensure the dimensions of the camera image. The processed camera image **p**' is subtracted from the processed projected image pscreen, according to Eq. (8).

$$\mathbf{p}_{out} = \left| \mathbf{p}^{screen} - \mathbf{p}' \right| \tag{8}$$

The effect of the subtraction is converted from the RGB space to a perceptually weighted gray scale, accordingly to Eq. (9). This results in the image of pixels $p_{ij}^{\ gray}$, which accordingly to Eq. (10), consist of identical red, green, and blue components, denoted as $r_{ij}^{\ gray}$, $g_{ij}^{\ gray}$, $b_{ij}^{\ gray}$, respectively.

$$r^{gray}_{ij} = g^{gray}_{ij} = b^{gray}_{ij} = \left[0.299 r^{out}_{ij} + 0.587 g^{out}_{ij} + 0.114 b^{out}_{ij} \right] \tag{9}$$

$$p^{gray}_{ij} = \left[r^{gray}_{ij}, g^{gray}_{ij}, b^{gray}_{ij} \right] \tag{10}$$

Then, the obtained image is converted into a binary image with a threshold equal to 100, accordingly to Eq. (11) and median filtered with a mask of a size equal to 7 by default.

$$p_{ij}^{bin} = \begin{Bmatrix} [0,0,0] & \left| r_{ij}^{gray} < 100 \right. \\ [255,255,255] & \left| r_{ij}^{gray} \geq 100 \right. \end{Bmatrix} \quad (11)$$

3.3.2. Hand tracking

In the obtained image, hands are detected by a contour-based algorithm from the OpenCV [19] library used for the system implementation. The algorithm utilizes contour trees, which first appeared in a paper by Reeb [20] and were further developed by Bajaj et al. [21] and Carr et al. [22]. The method of determining hand position depends on the part of the camera frame in which it appears. If the hand is in the left half of the frame, the position is the right upper corner of the rectangular area in which the hand fits. For the right half of the frame the position is the upper left corner of the hand rectangular area. The presented method for determining hand position in the image is related to the method of dynamic gesture recognition, and is used to eliminate a situation in which the initial phase of performing a gesture with both hands is interpreted as a gesture of one hand.

The change of a hand position within time is interpreted as a dynamic gesture. Hand movements are modeled by motion vectors created based on positions detected in camera frames n and $n + 3$. The optimal interval equals 3 frames according to the adopted method of dynamic gesture recognition, described further on. It was chosen empirically based on efficacy tests for 20 video streams containing recorded motion sequences. Each vector $u_{ij} = [u^x_{ij}, u^y_{ij}]$ is analyzed in the Cartesian coordinate system (Fig. 2.) with regard to velocity and direction.

The velocity is expressed by Eq. (12) and the direction is denoted by an angle φ relative to angle α between the velocity vector and versor of axis y, according to Eqs. (13) and (14).

$$\upsilon_{ij} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{t_i - t_j} \quad (12)$$

$$\varphi_{ij} = \begin{Bmatrix} \alpha_{ij} & , u_{ij}^x \geq 0 \\ 360° - \alpha_{ij} & , u_{ij}^x < 0 \end{Bmatrix} \quad (13)$$

$$\alpha_{ij} = \frac{180° \cdot \arccos \frac{u_{ij}^y}{|\vec{u}_{ij}|}}{\pi} \; [°] \quad (14)$$

For the purpose of reliable hand tracking, a Kalman filter was employed. A predicted state $\hat{s}_{t|t-1}$ of the hand in time t in relation to time t-1 is given by Eq. (15)

$$\hat{s}_{t|t-1} = F_t \hat{s}_{t-1|t-1} + w_{t-1} \quad (15)$$

where F_t is a transition matrix, $\hat{s}_{t-1|t-1}$ is a state in time t-1 and w_{t-1} is the process noise drawn from a zero mean multivariate normal distribution.

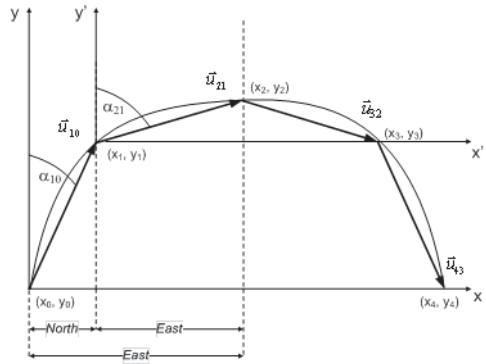


Fig. 2. Motion vectors created for hand movement with circular trajectory

The state of the hand at the given moment, according to Eq. (16), is expressed by (x, y) position, horizontal velocity, given by Eq. (17) and vertical velocity, given by Eq. (18).

$$s_t = \left[x_t, y_t, \upsilon_t^x, \upsilon_t^y\right] \quad (16)$$

$$\upsilon_t^x = \upsilon_t \sin\varphi \quad (17)$$

$$\upsilon_t^y = \upsilon_t \cos\varphi \quad (18)$$

The state in time t is associated to the state in time t-1 by a function of velocity and so the transition matrix takes the following values:

$$F = \begin{bmatrix} 1 & 0 & dt & 0 \\ 0 & 1 & 0 & dt \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (19)$$

where dt, expressed by Eq. (20), is a time modification of the velocity and depends on the camera frame rate f_{FR} and the number of frames based on which a singular motion vector is created. The c parameter is a scaling constant equal to 10 resulting from the adopted speed unit.

$$dt = c \cdot \frac{n_{\tau_0}^{\tau_1}}{f_{FR}} \quad (20)$$

Applying the transition matrix to the state at time step t-1, we obtain the hand state prediction at time step t, given by Eq. (21).

$$\hat{s}_{t|t-1} = \begin{bmatrix} x_{t|t-1} = x_{t-1|t-1} + c \cdot \dfrac{n_{\tau_0}^{\tau_1}}{f_{FR}} \cdot \upsilon_{t-1|t-1}^{x} \\ y_{t|t-1} = y_{t-1|t-1} + c \cdot \dfrac{n_{\tau_0}^{\tau_1}}{f_{FR}} \cdot \upsilon_{t-1|t-1}^{y} \\ \upsilon_{t|t-1}^{x} = \upsilon_{t-1|t-1}^{x} \\ \upsilon_{t|t-1}^{y} = \upsilon_{t-1|t-1}^{y} \end{bmatrix} \quad (21)$$

3.3.3. Dynamic gesture recognition

For the dynamic gesture recognition we use a fuzzy rule-based system. Motion trajectories are described by 30 fuzzy rules. Speed and direction have been chosen as linguistic terms describing a single motion vector. Since the motion trajectories are analyzed in two-vector segments, for the left and right hand, we can distinguish eight linguistic variables, i.e. the left hand speed in interval $t_2 - t_1$, denoted as ${\upsilon_{21}}^L$, the speed of the left hand for interval $t_1 - t_0$, denoted as ${\upsilon_{10}}^L$, the speed of the right hand for time interval $t_2 - t_1$, denoted as ${\upsilon_{21}}^R$, the speed of the right hand for interval $t_1 - t_0$, denoted as ${\upsilon_{10}}^R$, the direction of the left hand for time interval $t_2 - t_1$, denoted as ${\varphi_{21}}^L$, the direction of the left hand for interval $t_1 - t_0$, denoted as ${\varphi_{10}}^L$, the direction of the right hand for interval $t_2 - t_1$, denoted as ${\varphi_{21}}^R$, and the direction of the right hand for interval $t_1 - t_0$, denoted as ${\varphi_{10}}^R$. We have defined 4 fuzzy sets for the speeds and 4 fuzzy sets for the angles defining the direction. Fuzzy sets for the speed, denoted as VS, S, M, L represent linguistic terms: *very small*, *small*, *medium* and *large*, respectively. Fuzzy sets for the directions, denoted as N, E, S, W are linguistic terms: *North*, *East*, *South*, *West*. Triangular membership functions are used for all sets. Set N is defined by two triangular functions determining intervals [0°, 90°] and [270°, 360°]. The trajectories are modeled regarding the naturalness of human motions. Thus, sample fuzzy rules describing the left hand motion from the left to the right side, regarding the possibility of performing the gesture with circular trajectory, take the form of (22 – 24).

$$\forall \begin{pmatrix} \varphi_{10}^L \in N \wedge \varphi_{21}^L \in E \wedge \upsilon_{10}^L \notin S \wedge \\ \upsilon_{21}^L \notin S \wedge \upsilon_{10}^R \in VS \wedge \upsilon_{21}^R \in VS \end{pmatrix} \Rightarrow \\ \Rightarrow g \in G_2 \quad (22)$$

$$\forall \begin{pmatrix} \varphi_{10}^L \in E \wedge \varphi_{21}^L \in E \wedge \upsilon_{10}^L \notin S \wedge \\ \upsilon_{21}^L \notin S \wedge \upsilon_{10}^R \in VS \wedge \upsilon_{21}^R \in VS \end{pmatrix} \Rightarrow \\ \Rightarrow g \in G_2 \quad (23)$$

$$\forall \begin{pmatrix} \varphi_{10}^L \in E \wedge \varphi_{21}^L \in S \wedge \upsilon_{10}^L \notin S \wedge \\ \upsilon_{21}^L \notin S \wedge \upsilon_{10}^R \in VS \wedge \upsilon_{21}^R \in VS \end{pmatrix} \Rightarrow \\ \Rightarrow g \in G_2 \quad (24)$$

The resulting sets representing gesture classes take the form of singletons, adapting a zero-order Takagi-Sugeno model. The output value of the inference system is the output of the rule for which the membership function takes the greatest value. In addition, a threshold equal to 0.5 is set, below which no gesture class is associated with the motion. Thus, the problem of classifying transitional movements as gestures is solved this way.

Employing fuzzy inference in gesture recognition enabled to obtain an average efficacy equals respectively to 96.94% for one-hand, and 99.30% for two-hand gestures. In the previous approach proposed by the authors the recognition module was based on fixed thresholds instead of fuzzy sets, and the results were not fully satisfactory [23].

3.3.4. Static gesture recognition

For static gesture recognition we use Support Vector Machine classifier of a type C-SVC (C-Support Vector Classification) [24]. Its software implementation is based on the LibSVM library [24]. The linear kernel has been chosen as it performed similarly to the RBF (Radial Basis Function) kernel and provided slightly better efficiency during preliminary tests.

The input vectors contain values of pair-wise geometrical histograms representing palm contours. During the phase of classifiers training, each gesture class is represented by 90 histograms, i.e. three 30-frame sets of samples for different motion trajectories, defined for the purpose of examining the classifiers. The training process is performed using a one-versus-all method, i.e. the input vectors are divided into two subsets. The first subset contains all vectors representing the particular gesture and the second subset contains all the remaining vectors for other gestures.

The utilized LibSVM implementation extends the classical SVM method with the possibility to determine the probability of a recognized gesture assignment to a given class. This feature has been employed in the system by setting the shape-gesture association decision threshold to 0.8. The output of the whole classification system is the output of the classifier which returns the maximum positive value of the probability.

3.4. System GUI and Gesture Sonification

Considering the specificity of multimodal perception, all sound mixing operations can be handled with a GUI that does not provide visual information reflecting audio changes or with full graphical representation of sound modifications. The middle part of the application window set in the second of the above-mentioned modes contains circles representing audio sources (fig. 3). 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 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. 3). 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.

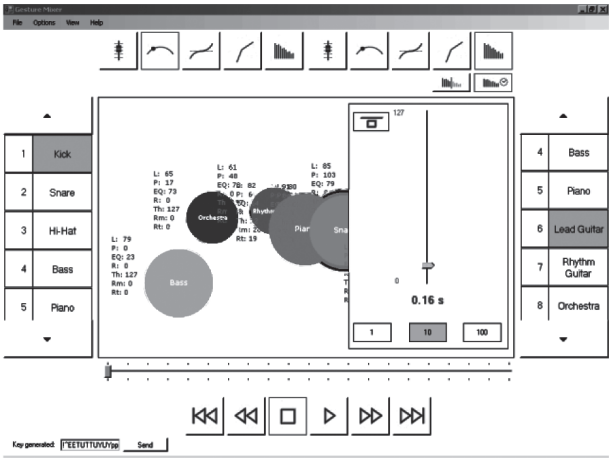


Fig. 3. Graphical user interface of the application

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. 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.

Table 1. Default gesture set

1	✋	7	↓
2	☝	8	⇔
3	✌	9	⇒⇐
4	→	10	⇑⇓
5	←	11	⇓⇑
6	↑		

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

4. EXPERIMENTS INVOLVING SOUND ENGINEERS

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 sound mixing processes were carried out using the engineered interface and the Steinberg Cubase Studio 5 music production system. The experiments were performed for various manners of systems controlling. The excerpts have been subjectively assessed (audio samples available online [29]).

4.1. Sound mixing methodology

Ten professional mixing engineers were involved in the experiments. The task of each engineer was to mix eight audio tracks which significantly differed from each other regarding both musical and signal features. 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 5 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 manner 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 5 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. Among these aspects were gesture dictionary intuitiveness, convenience and precision of a parameter editing. As this paper put emphasis on visual aspects of DAW controlling the mentioned features examination can be found in another paper by the authors [26]. The engineers were also asked to order their own mixes from the best sounding to the worst sounding. The results have been presented in a further section.

4.2. Experiment conditions and methodology of subjective assessment

Both the mixing of audio signals and subjective assessment were conducted in identical conditions. Yamaha MSP5 studio monitors placed on Ultimate Support MS-45B2 stands were used. The distance between the monitors equaled 1.85m. The mixing engineer was situated in the sweet spot.
Subjective evaluation was conducted using a rank order test [27]. The assessed samples were 15-second excerpts of mixes from all 5 mixing methods. The ranking order of mixes from each engineer was analyzed via pair comparison, according to table 3.

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

Pair of mixes	Information provided by pair comparison
1. & 2.	Checking 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 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 greater influence on mixing results has controlling manner (2. > 3.) or presence of visual stimuli (3. > 2.)
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 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)

It is worth noticing that 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.

4.3. 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 graphical mode, described earlier. In fig. A (see Appendix A) sample visualizations for one engineer (engineer no. 3) and all five mixing methods have been presented. 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 statistical significance.

4.4. 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), i.e. 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.
In addition, after a few weeks after the mixing sessions took place the engineers were asked to fill in a supplementary questionnaire. They were supposed to answer a question whether they had closed their eyes during work with the engineered system with regard to the ways of sound controlling. The engineers' answers are presented in table 5. Because of the long period between the mixing sessions and the supplementary questionnaire the engineers could select option "don't remember" in case of difficulties with answering the question. This option could also be selected if an engineer was not sure whether indicated ways of mixing were the only ones in which they had closed their eyes.

Table 5. Answers of mixing engineers to questions about closing eyes in each of the mixing manner

	mixing manners						
Eng. no.	1	2	3	4	5	don't remember	closed eyes
1	Yes						Yes
2		Yes			Yes	Yes	Yes
3	Yes	Yes	Yes	Yes	Yes		Yes
4						Yes	Yes
5						Yes	
6	Yes		Yes			Yes	Yes
7					Yes	Yes	Yes
8						Yes	Yes
9	Yes		Yes				Yes
10	Yes		Yes				Yes

The obtained distributions of answers are interesting. From the above results a tendency to close one's eyes in the case of mixing sound in the limited GUI mode can be observed. The conclusion could be that sound engineers closed their eyes only when there was nothing informative to look at. When changes in sound were supported by trackbars with values the engineers naturally felt obliged to look at the screen. Such a phenomenon reflects perceptual and cognitive processes in which the most informative source has the highest priority during decision making.

4.5. 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. 7). 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. In table 7 ranks of aesthetic value given by engineers no. 3 and 7 are presented. Results for these particular engineers have been chosen for presentation because the subjective evaluation involving independent listeners was performed based on their mixes. The results of the subjective evaluation are given further on.

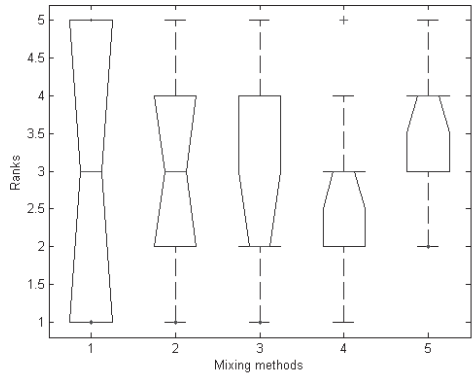


Fig. 7. Box-and-whisker plot for the assessment of aesthetic values of mixes for various mixing methods

Table 7. Ranks of aesthetic value given by engineers no. 3 and 7 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, Method 5 – direct use of DAW software, (1 – worse sounding, 5 – best sounding)

Eng. no. / Method	1	2	3	4	5
3	1	4	5	3	2
7	5	3	1	2	4

The obtained rank distributions for each mixing method have been analyzed in terms of statistical significance using the Friedman test. Obtaining $p > 0.05$ ($\cong 0.77$) did not enable us to disregard the zero hypothesis referring to 'no differences' of mean values between the mixing methods.

5. EXPERIMENTS INVOLVING EXPERT LISTENERS

To check the level of uncertainty in marks given by mixing engineers to their own mixes, in the second stage of the experiments subjective tests involving non-mixing engineer listeners were carried out. These tests were based on pairwise comparison. According to its methodology two test series were used to check reliability of answers. Among 19 persons who took part in the tests 12 were considered expert group. These persons made not more than 3 mistakes (inconsistent indications between series). Also the number of selections of the same sample in both series differed by no more than 1 and it was possible to determine trend of mix preference basing on marks in both series. Basing on the analysis of rating in both series the mixes were ranked according to 5-degree scale (1 – worse, 5 – best). The distributions of ranks are presented in figs. 11 and 12, for mixes of engineers no. 3 and 7, respectively. The greater disparities in ranks of mixes of engineer no. 7 resulted from smaller tone differences between mixes than in mixes of engineer no. 3.

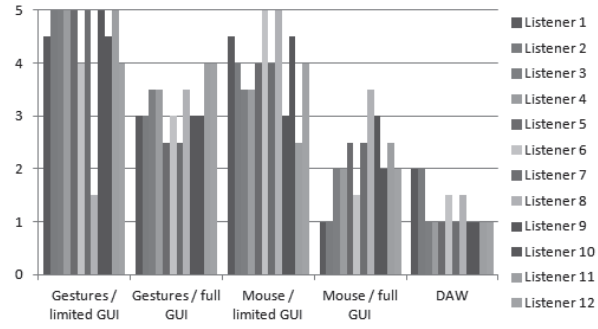


Fig. 11. The distribution of ranks assigned to mixes of engineer no. 3 based on subjective assessment in both series of the pair comparison test

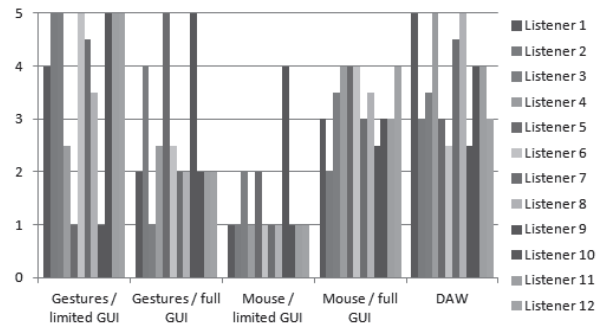


Fig. 12. The distribution of ranks assigned to mixes of engineer no. 7 based on subjective assessment in both series of the pair comparison test

The analysis of the distribution of ranks has been performed based on box-and-whisker plots (figs. 13. and 14) and statistical tests.

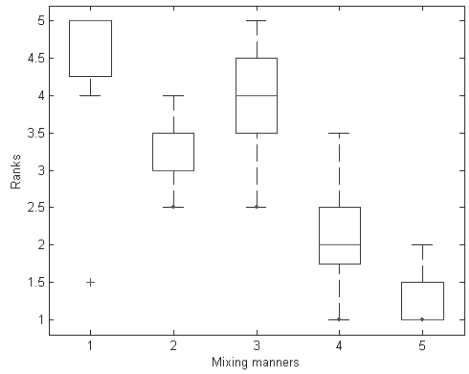


Fig. 13. Box-and-whisker plot for ranks assigned to aesthetic value of mixes of engineer no. 3, for various ways of mixing

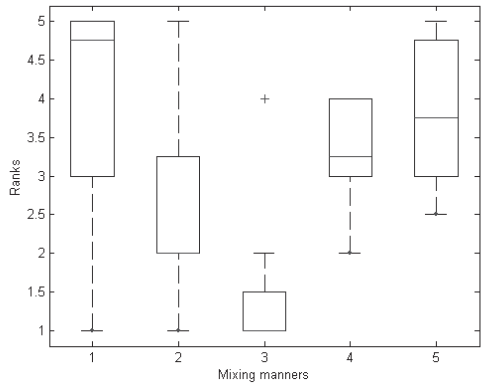


Fig. 14. Box-and-whisker plot for ranks assigned to aesthetic value of mixes of engineer no. 7, for various ways of mixing

The global trend of grades assigned by listeners to the mixes does not reflect the grades assigned by the mixers to their own mixes. Especially, it can be observed in mixes of engineer no. 3 (fig. 13 and table 7). The mix obtained by hand gestures in limited GUI mode was marked as „1" by the engineer, whereas in subjective tests involving listeners it was given the highest number of maximum scores. The reason for this could be the engineer's hearing fatigue after the whole mixing session. This engineer suggested in the questionnaire performing listening tests with other listeners. The trend of ranks assigned by the engineer no. 7 to her own mixes did not reflect the global trend of listeners' ratings in ranks of mixes obtained for manners 2. and 4. The engineer assessed the mix obtained by hand gestures as better than the mix obtained using mouse and keyboard (in both cases for full GUI mode).

The obtained distribution of aesthetic value ranks of engineers no. 3 and 7 was statistically analyzed using Friedman test which is the non-parametric equivalent of ANOVA test (one-way repeated analysis or two-way analysis with single classification), appropriate for ordinal variables. The results are presented in tables 9 and 10. The column headings contain sum of squares between groups (SS Effect), degrees of freedom (df Effect), mean squares (MS Effect = SS/df), sum of squares inside groups (SS Error), degrees of freedom inside groups (df Error), mean square error (MS Error), Friedman's chi-square statistic (χ^2), and p values for the chi-square statistic.

Table 9. Values of Friedman test for ranks assigned by expert listeners to mixes of engineer no. 3

SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	χ^2	p
83	4	20.75	28	44	0.636	35.89	0,00

Table 10. Values of Friedman test for ranks assigned by expert listeners to mixes of engineer no. 7

SS Effect	df Effect	MS Effect	SS Error	Df Error	MS Error	χ^2	p
49.46	4	12.37	67.54	44	1.54	20.29	0,00

Obtained values of test probabilities in Friedman test, in both cases smaller than adopted level of significance equal to 0.05, prove statistically significant differences in distribution of ranks of particular mixes. To check which mixes differed significantly, Wilcoxon signed-rank test was additionally performed. The results for engineer no. 3 are consistent with the results of the analysis of parameter value distribution regarding influence of visualizations for parameter value choice. For both ways of interaction, i.e. employing gestures and mouse, depending on the use of full or limited GUI mode, there were significant differences between mixes.

The distribution of ranks given to the mix of engineer no. 3 is close to the distribution defined by the relation (25) in which pairs defined in table 3 are compared. Obtaining results reflecting such a relation would explicitly show that mixing without eye sight involvement is superior to mixing supported by graphical message. It may also prove that gesture handled interface is more ergonomic than mouse and keyboard controlled system.

$$1. > 2. \,\&\, 1. > 3. \,\&\, 1. > 4. \,\&\, 1. > 5. \,\&\, 2. > 4. \,\&\, \quad (25)$$

$$2. > 5. \,\&\, 3. > 4. \,\&\, 3. > 5. \,\&\, 5. > 4.$$

The only difference between the relation (25) and obtained results is the lower rating of the mix obtained directly using DAW software than the rating of the mix obtained using mouse in the limited GUI mode. The factor which caused this could be different way of presenting information to the user in the engineered system than in DAW software. In particular, such an outcome could be associated with reflecting panorama and equalizer gain in localization of a shape on a screen. Such a way of presenting information clearly appealed to the engineer.

Presented in figs. 15 and 16 histograms of ranks for aesthetic values of mixes of engineers no. 3 and 7, given by listeners, confirm the effect observed in subjective evaluation by mixing engineers. Namely, vividness of mixes obtained by gestures in limited GUI mode caused assigning them the highest number of maximum grades. However, in a few cases the same feature was the reason for assigning the minimum grade to these mixes. Experts who assigned minimum grades, similarly as three engineers mentioned earlier, considered vivid sound as exaggerated.

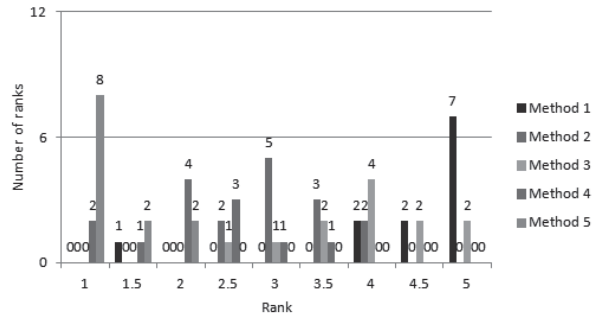


Fig. 15. Histograms of ranks of aesthetic values assigned to the mixes of engineer no. 3, for each of the mixing manners

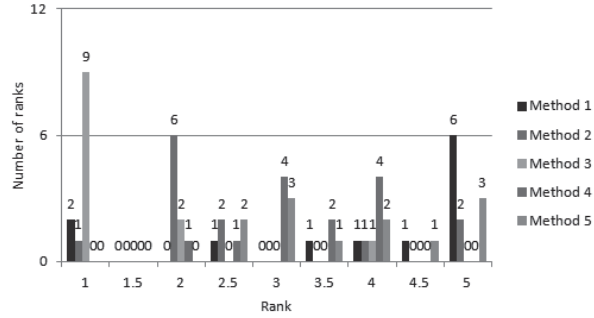


Fig. 16. Histograms of ranks of aesthetic values assigned to the mixes of engineer no. 7, for each of the mixing manners

6. CONCLUSIONS

In the paper, 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 indeed possible. 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.

7. 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.

8. REFERENCES

[1] G. Kramer, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, J. Neuhoff and co-authors, "Sonification report: Status of the field and research agenda," Tech. Rep., International Community for Auditory Display, 1999, Available:http://www.icad.org/websiteV2.0/References/nsf.html. [Accessed March 13, 2013].

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

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

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

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

[6] 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).

[7] 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).

[8] 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).

[9] 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).

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

[11] 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).

[12] 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).

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

[14] 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).

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

[16] Nintendo, "Wii," 2011. [Online]. Available: http://www.wii.com. [Accessed October 27 2011].

[17] Z. Can, W. Jiankang, T. Guofang, "Object Tracking and QOS Control Using Infrared Sensor and Video Cameras," 2006 IEEE International Conference on Networking, Sensing and Control, 2006. ICNSC '06. (2006).

[18] K. Jungsoo, H. Jiasheng, K. Lyons, T. Starner, "The Gesture Watch: A Wireless Contact-free Gesture based Wrist Interface," 11th IEEE International Symposium on Wearable Computers, Boston (2007).

[19] G. Bradski, A. Kaehler, Learning OpenCV: Computer Vision with the OpenCV Library, Sebastopol: O'Reilly (2008).

[20] G. Reeb, "Sur les points singuliers d'une forme de Pfaff completement integrable ou d'une fonction numerique," Comptes Rendus de l'Academie des Sciences, vol. 222, pp. 847-849 (1946).

[21] C. Bajaj, V. Pascucci, D. Schikore, "The contour spectrum," IEEE Visualization 1997.

[22] H. Carr, J. Snoeyink, M. van de Panne, "Progressive topological simplification using contour trees and local spatial measures," 15th Western Computer Graphics Symposium, British Columbia (2004).

[23] 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).

[24] C.-C. Chang, C.-J. Lin, "LIBSVM: a Library for Support Vector Machines," Science, vol. 2, no. 3, pp. 1-39 (2011).

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

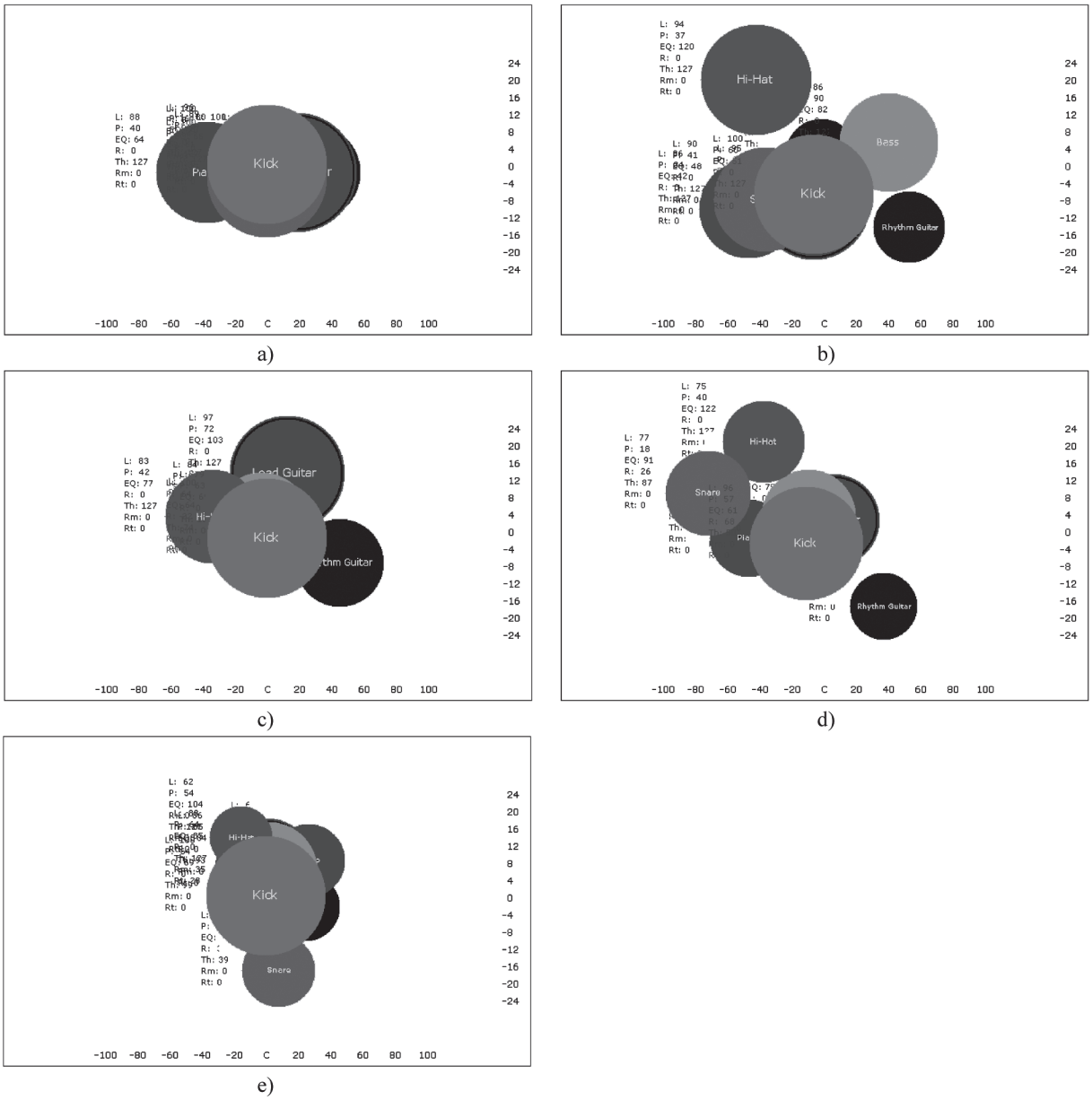
[26] M. Lech, B. Kostek, "Evaluation of the influence of ergonomics and multimodal perception on sound mixing with a novel gesture-based mixing interface," Journal of the Audio Engineering Society, 2013 (*in press*)

[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] Video presentation of the system, http://sound.eti.pg.gda.pl/~kuneck/filmy/AES_2013_mix/Gesture-controlled_MIX.wmv

[29] Audio samples, http://sound.eti.pg.gda.pl/~mlech/samples/

Appendix A



a) b) c) d) e)

Fig. A1. 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, (e) direct use of DAW software