J Intell Robot Syst (2010) 57:371–391
DOI 10.1007/s10846-009-9351-9

# Multimodal Interface Technologies for UAV Ground Control Stations

## A Comparative Analysis

**I. Maza · F. Caballero · R. Molina · N. Peña · A. Ollero**

Received: 1 February 2009 / Accepted: 1 August 2009 / Published online: 13 August 2009

**Abstract** This paper examines different technologies that can be applied in the design and development of a ground control station for Unmanned Aerial Vehicles (UAVs) equipped with multimodal interfaces. Multimodal technologies employ multiple sensory channels/modalities for information transmission as well as for system control. Examples of these technologies could be haptic feedback, head tracking, auditory information (3D audio), voice control, tactile displays, etc. The applicability and benefits of those technologies is analyzed for a task consisting in the acknowledgement of alerts in an UAV ground control station composed by three

This work is partially supported by the CONET Network of Excellence (ICT-224053) funded by the European Commission under FP7, the ROBAIR Project funded by the Spanish Research and Development Program (DPI2008-03847) and the ATLANTIDA Project leaded by *Boeing Research & Technology Europe*.

I. Maza (✉) · F. Caballero · A. Ollero
Robotics, Vision and Control Group, University of Seville,
Avd. de los Descubrimientos s/n, 41092, Sevilla, Spain
e-mail: imaza@cartuja.us.es

F. Caballero
e-mail: caba@cartuja.us.es

A. Ollero
e-mail: aollero@cartuja.us.es

R. Molina · N. Peña
Boeing Research & Technology Europe, Canada Real de las Merinas,
1-3, Bldg 4, 28042 Madrid, Spain

R. Molina
e-mail: roberto.molina@boeing.com

N. Peña
e-mail: nicolas.penaortiz@boeing.com

A. Ollero
Center for Advanced Aerospace Technology (CATEC), Seville, Spain
e-mail: aollero@catec.aero

screens and managed by a single operator. For this purpose, several experiments were conducted with a group of individuals using different combinations of modal conditions (visual, aural and tactile).

**Keywords** Unmanned Aerial Vehicles · Multimodal interfaces

## 1 Introduction

It is known that multimodal display techniques may improve operator performance in Ground Control Stations (GCS) for Unmanned Aerial Vehicles (UAVs). Presenting information through two or more sensory channels has the dual benefit of addressing high information loads as well as offering the ability to present information to the operator within a variety of environmental constraints. A critical issue with multimodal interfaces is the inherent complexity in the design of systems integrating different display modalities and user input methods. The capability of each sensory channel should be taken into account along with the physical capabilities of the display and the software methods by which the data are rendered for the operator. Moreover, the relationship between different modalities and the domination of some modalities over others should be considered.

Using multimodal technologies begins to be usual in current GCSs [7, 13, 14], involving several modalities such as positional sound, speech recognition, text-to-speech synthesis or head-tracking. The level of interaction between the operator and the GCS increases with the number of information channels, but these channels should be properly arranged in order to avoid overloading the operator.

In [19] some of the emerging input modalities for human-computer interaction (HCI) are presented and the fundamental issues in integrating them at various levels-from early “signal” level to intermediate “feature” level to late “decision” level are discussed. The different computational approaches that may be applied at the different levels of modality integration are presented, along with a briefly review of several demonstrated multimodal HCI systems and applications.

On the other hand, the intermodal integration can contribute to generate the illusion of presence in virtual environments if the multimodal perceptual cues are integrated into a coherent experience of virtual objects and spaces [1]. Moreover, that coherent integration can create cross-modal sensory illusions that could be exploited to improve user experiences with multimodal interfaces, specifically by supporting limited sensory displays (such as haptic displays) with appropriate synesthetic stimulation to other sensory modalities (such as visual and auditory analogs of haptic forces).

Regarding the applications in the UAVs field, [10] provides a survey on relevant aspects such as the perceptual and cognitive issues related to the interface of the UAV operator, including the application of multimodal technologies to compensate for the dearth of sensory information available.

The paper is structured as follows. In the next section, different technologies that can be applied in the design and development of a GCS for UAVs equipped with a multimodal interface are summarized in Section 2. Then, Section 3 describes a multimodal testbed developed by the authors along with a set of tests to measure the benefits under different modal conditions. Section 4 provides the results obtained in several experiments performed with a group of individuals and analyzes the impact of the different modalities. Finally, the conclusions and future work section closes the paper.

## 2 Interactions between Operator and GCS

In the design of a GCS, two information flows are usually considered: *from GCS to operator*, presenting information about the UAV, the environment and the status of the mission, and *from operator to GCS* in the form of commands and actuations which are treated as inputs by the GCS software.

But there is a third flow of information which is not usually addressed in the design of the UAV's GCS; the information about the *operator's state* that can be gathered by sensors and processed by the GCS software. This channel could allow to have an adaptive GCS software, which can change the modality and format of the information depending on the state of the operator (tired, bored, etc). Furthermore, the information about the operator can be also used to evaluate and improve the interface. For instance, it is possible to register which screens are mainly used by the operator during a certain type of mission.

Next subsections are devoted to each of these information flows, summarizing several methods and devices which are usually applied.

### 2.1 Information Flow from GCS to Operator

Classical modalities in GCS to operator communications are visual information (mainly in monitors) and sound alerts. Last decades researchers have dedicated a significant effort to define the characteristics of such communications, taking into account the operator's capabilities and maximizing the information showed to the operator. Thus, effective data visualization and distribution is discussed in [26] and [22], where different models to measure the effectiveness of the visualization system are presented. Other researchers include the color, shape or size of the displays in their analysis.

Sound alerts have also been deeply studied, mainly applied to control stations in general. Intensity, frequency or loudness are some of the parameters taken into account to create comfortable and effective sound alarms. References [17] and [16] are good examples of sound alarm studies focused on civil aircrafts.

However, higher computational capabilities and the evolution of the communication systems raised new devices and techniques able to provide more complex information. The next paragraphs describe some of these new approaches.

#### *2.1.1 3D Audio*

Concerning the aural modality, the 3D audio can improve the Situational Awareness (SA) of the operator. Three dimensional audio is based on a group of sound effects that attempt to widen the stereo image produced by two loudspeakers or stereo headphones, or to create the illusion of sound sources placed anywhere in a three dimensional space, including behind, above or below the listener. Taking into account the portability usual requirement for the ground stations, the use of a headset is usually preferred for the operator, instead of a set of speakers around him.

Thus, the objective of the 3D audio interface in a GCS is to provide multiple located sources of sound for the operator (the listener) to improve his SA while performing a given mission. In this way, the operator is able to recognize the presence of an alarm and also the origin of such alarm. This functionality is provided for example by a library called OpenAL [4] (Open Audio Library), which is a free software cross-platform 3D audio Application Programming Interface (API) designed for efficient rendering of multichannel three dimensional positional audio, and distributed under the LGPL license. This library has been used in our system implementation, which is described in Section 3.

### *2.1.2 Speech Synthesis*

Considering also the audio channel, the speech synthesis technology has been also included in the system used in the experiments presented in this paper. Speech synthesis, also known as text-to-speech, is the artificial production of human speech. It can be implemented in software or hardware and basically can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units.

The quality of a speech synthesizer is judged by its similarity to the human voice, and by its ability to be understood. An intelligible text-to-speech program allows operators to listen complex messages, normally related with the state of commands, events or tasks currently carried out in the GCS. An example of these applications in the UAVs context can be found in the WITAS project [21].

A good example of free speech synthesis software is the Festival library [24]. It is a general multi-lingual speech synthesis system originally developed at Centre for Speech Technology Research (CSTR) at the University of Edinburgh. It offers a full text to speech system with various APIs, as well as an environment for development and research of speech synthesis techniques. It offers a general framework for building speech synthesis systems. As a whole it offers full text to speech through a number of APIs: from shell level, through a Scheme command interpreter, as a C++ library, from Java, and an Emacs interface. In the tests presented in this paper, Festival has been used as a C++ library and integrated into our multimodal software application.

### *2.1.3 Haptic Devices*

Haptic technologies interface to the user via the sense of touch by applying forces, vibrations and/or motions to the operator. This mechanical stimulation can be applied to assist in the "creation" of virtual objects (objects existing only in a computer simulation), for control of such virtual objects, and to enhance the remote control of machines and devices (teleoperators).

In the particular case of GCSs, haptic devices add a new communication channel to the operator. The vibration of the device can be used as a stand alone alarm mechanism, or in combination with other sensory channels can increase the information provided to the user. For instance, if the activation of the haptic device is added to a currently played sound alarm, the operator will consider that the priority/criticity of such alarm has been increased.

## 2.2 Information Flow from Operator to GCS

Normally, the operator provides information to the GCS through mouse, touchpad or keyboard. However, other channels can be used to provide information to the software application running in the GCS.

### *2.2.1 Touch Screens*

A touchscreen is a display which can detect the presence and location of a touch within the display area. The term generally refers to touch or contact to the display of the device by a finger or hand. The touchscreen has two main attributes. First, it enables the operator to interact with what is displayed directly on the screen, where it is displayed. Secondly, it lets the operator do so without requiring any intermediate device. Thus, touchscreens allows intuitive interactions between the operator and the GCS application.

Nevertheless, it is important to remark that touchscreen technology is usually poor in resolution if the operator uses his finger, i.e. the minimal size of the objects required to guarantee a proper interaction with the user must be bigger compared to a mouse or a touchpad. This is one of the main constraints to be considered in the design of the graphical interfaces to be used with touchscreens.

### *2.2.2 Automatic Speech Recognition*

Speech recognition (also known as automatic speech recognition or computer speech recognition) converts spoken words to machine-readable input (for example, to key presses, using the binary code for a string of character codes). Speech recognition provides an easy and very effective way to command tasks to GCSs [7].

## 2.3 Operator's State

The operator's state is the third information flow mentioned above. It can be defined as the set of physiological parameters that allows to estimate the state of a human operator: heartbeat, temperature, transpiration, position, orientation, etc. All this information can be used by adaptive systems to improve the operator environment or to reduce the stress/workload of the operator.

There are plenty of studies examining how psychophysiological variables (e.g., electro-encephalogram, eye activity, heart rate and variability) change as a function of task workload (see [3, 18] or [25]), and how these measures might be used to monitor human operators for task overload [15] and/or used to trigger automated processes.

Next sections details some of the current technologies used in the operator's state estimation.

### *2.3.1 2DoF Head Tracking*

As the operator's head concentrates his main sensorial capabilities, a very important tool to acquire the operator's state is the head position and orientation tracking. 2DoF head tracking applications and products are easy to find. Most of these products are based on image processing and marks/spots placed in the users head (or a hat). They also provide the two angle information used to move the mouse from left/right and up/down. The following professional solutions can be highlighted: *Tracker Pro* [8], *Headmouse Extreme* [2] or *SmartNav 4 AT* [11].

A practical application of this technology could be a GCS software that provides the critical alerts on the screen which is being used by the operator when they occur. It can be also applied to evaluate the interaction between the human and the GCS during each mode of operation in terms of which information/screen is more relevant for the operator, etc.

### *2.3.2 2DoF Eye Tracking*

If the GCS is composed by several screens it could be also necessary in many cases to track the head and the eye position in order to determine which screen is being used by the operator. However, products related with 2DoF eye tracking are scarce in the market. All of them are based on computer vision systems that analyzes the images gathered by a camera (normally mounted on the computer monitor). *EyeTech TM3* [23] is a good example.

### *2.3.3 6DoF Head Tracking*

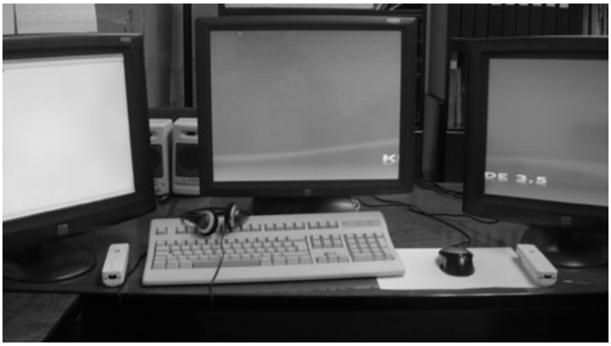
6DoF head tracking moves one step forward and allows estimating the complete position and orientation of the user's head in real time. Most of the existing methods make use of cameras and visual/IR patterns mounted on the operator's head.

*TrackIr* [12], *Cachya* [20] and *FreeTrack* [5] represent the main options in the market. They use a 3D pattern visible in the infrared or visual band to estimate the position and orientation of the operator's head.

### *2.3.4 Body Motion Sensors*

In order to register the behavior of the operator during a mission it could be also convenient to attach small sensors to his body to log the motion data. For example, it is possible to embed a wireless 3-axis accelerometer in each arm of the operator. The data registered can help to determine his current state (bored, tired, etc) and useful information as for example which arm is more used in each mode of operation and GCS configuration.

**Fig. 1** Multimodal technologies based system developed



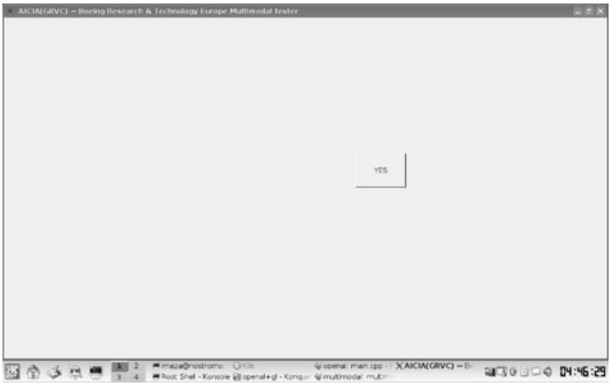

AICIA(GRVC) – Boeing Research & Technology Europe Multimodal tester
YES
04:46:29

**Fig. 2** Multimodal software application graphical interface

## 3 Multimodal Technologies Based System Developed

The applicability and benefits of multimodal technologies has been analyzed for a simple task consisting in the acknowledgement of alerts in an UAV ground control station composed by three screens and managed by a single operator. For this purpose, several experiments were conducted with a group of individuals using different combinations of modal conditions (visual, aural and tactile).

A software application integrating the different modalities has been developed and used in the tests. The experimental results are shown in this section whereas the corresponding analysis and conclusions are detailed in Section 4. This information can be used as a starting point in the design of the multimodal ground control station for UAVs.

### 3.1 System Description

Figure 1 presents the system used to perform the multimodal experiments. This setup emulates a GCS for UAVs in which the operator can interact with the station through the following devices: three touch screens, three wireless haptic devices attached to the right hand, left hand and also on the chest of the operator, one optical mouse, one headset and stereo speakers. In addition, modules for Speech Synthesis and 3D Sound are included into the software application.

The application has been developed under Linux and makes use of different modalities to show the information to the operator. The graphical interface is composed by a single window (see Fig. 2) in which several buttons labeled as "Yes" or "No" appear in random positions. Only one button is present on the screen at any time and each button is displayed until it is pressed or until a programmable timeout

**Table 1** Operator right and wrong actions depending on the type of button which appears in the interface

| Button | Right action | Wrong action |
|---|---|---|
| "Yes" | Press before timeout expires | Do not press before timeout expires |
| "No" | Do not press | Press |

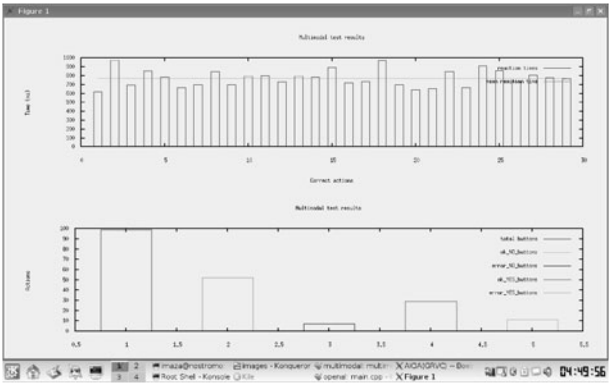


**Fig. 3** Graphical interface with the results of the test

($T_{\text{yes}}$ or $T_{\text{no}}$) expires. The duration of the experiment and the size of the buttons is also programmable.

The mission for the operator is quite simple: press **only** buttons labeled as "Yes" as soon as possible. Then, the right and wrong actions when each button appears on the screen are summarized in Table 1. For some values of the parameters, some wrong actions do not exist, i.e. if $T_{\text{yes}} \to \infty$ there is no possible wrong action for the "Yes" buttons.

Both type of buttons have the same grey color, which is also the same color used in the background of the window. Therefore, when a button appears in the perimeter of the field of view, it is hard to realize for the user that it is there. This bad feature has been intentionally left in the application to emphasize the benefits of other modalities different from the visual one.

Once a test has finished, several performance parameters are computed and showed automatically on the screen. Figure 3 is an example of the interface with the results of a given experiment. In the top subfigure, the reaction time of the operator for each right action (corresponding to "Yes" buttons pressed before $T_{\text{yes}}$) is shown in milliseconds. The mean reaction time ($\overline{T_{\text{yes}}}$) is also represented with an horizontal line. In the subfigure below, the total number of buttons ($n$), and the number of right and wrong actions for each button ($n_{\text{right\_yes}}, n_{\text{right\_no}}, n_{\text{wrong\_yes}}$ and $n_{\text{wrong\_no}}$) are represented with bars. Table 2 shows a summary of the values presented in Fig. 3.

The system developed allows integrating visual, aural and tactile modalities into the GCS. Their integration in the developed software have been carried out as follows:

– Speech synthesis: Once each button appears on the screen, its label is told to the operator.
– 3D audio: Depending on the location of the button on the window (left, right or middle), the source of audio corresponding to its label is generated on the left, on the right or in front of the operator respectively.

**Table 2** Summary of the values represented in Fig. 3

| $T$ (sec) | $\overline{T_{\text{yes}}}$ (ms) | $n$ | $n_{\text{right\_yes}}$ | $n_{\text{right\_no}}$ | $n_{\text{wrong\_yes}}$ | $n_{\text{wrong\_no}}$ |
|---|---|---|---|---|---|---|
| 90 | 773.28 | 99 | 29 | 52 | 11 | 7 |

**Table 3** Summary of the tests designed along with the identifiers that will be used later to make reference to them

| Experiment nr. | Description | Identifier |
|---|---|---|
| #1 | Mouse interface only | Mouse |
| #2 | Touch screen interface only | TS |
| #3 | Touch screen and speech synthesis | TS+speakers |
| #4 | Touch screen and 3D audio | TS+3D |
| #5 | Touch screen and tactile interfaces | TS+vibrator |
| #6 | Touch screen, 3D audio and tactile interfaces | TS+3D+vibrator |
| #7 | Touch screen interface test repetition | TS2 |

- Vibrator: The wireless vibrator is activated every time a “Yes” button appears on the screen. Moreover, depending on the location of the button on the window (left, right or middle), the device on the left, on the right or in the middle vibrates respectively.

### 3.2 Tests Performed and Results

Prior to the different modalities tests, several sizes for the rectangular buttons are used in many tests with the touch screens. The goal is to determine the minimum size for the buttons which can “guarantee” a correct operation with the application. This minimum size is estimated to be approximately 2.8 × 2.6 cm.

The tests described in the next subsections have been performed using the multimodal software previously presented. A short video with a summary of the experiments can be found at [9]. The values selected for the parameters of the software application have been the following:

- Full duration of each test: $T = 8$ min.
- Size of the buttons: 3.0 × 2.8 cm for the central screen and 2.8 × 2.6 cm for the left and right screens.
- Timeout period of the buttons: $T_{\text{yes}} \to \infty$ and $T_{\text{no}} = 1.6$ sec respectively.

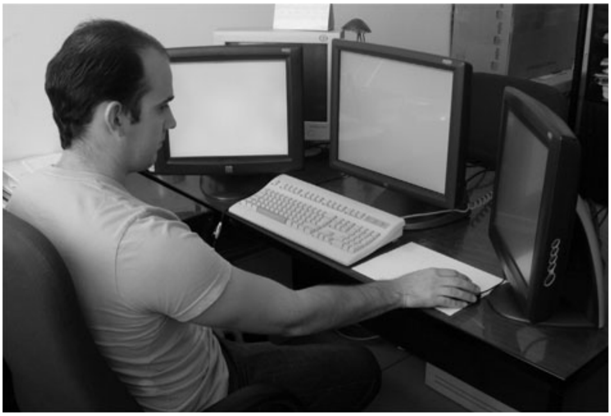


**Fig. 4** In the experiment #1 the operator is only allowed to use the mouse interface

**Table 4** Summary of the results for the experiment #1

| Individual | $n$ | $n_{\text{right\_yes}}$ | $n_{\text{right\_no}}$ | $n_{\text{wrong\_no}}$ | $\overline{T_{\text{yes}}}$ (ms) | $\sigma$ (ms) |
|---|---|---|---|---|---|---|
| #1 | 334 | 166 | 168 | 0 | 1288.2 | 375.3 |
| #2 | 316 | 165 | 151 | 0 | 1462.3 | 537.7 |
| #3 | 325 | 162 | 163 | 0 | 1377.2 | 349.0 |
| #4 | 341 | 172 | 169 | 0 | 1239.7 | 373.1 |
| #5 | 358 | 178 | 179 | 1 | 1098.7 | 322.6 |
| #6 | 368 | 195 | 173 | 0 | 1066.3 | 256.3 |
| #7 | 350 | 170 | 180 | 0 | 1151.9 | 303.7 |
| #8 | 342 | 172 | 170 | 0 | 1230.7 | 382.6 |
| #9 | 357 | 171 | 186 | 0 | 1093.6 | 300.2 |

On the other hand, the tests have been done by nine people with ages between 20 and 30 years old (3 women and 6 men), registering their performance and opinions.

Table 3 shows a summary of the seven tests designed for the multimodal station. Each of those tests is detailed in the next subsections, including the results obtained by the different individuals.

### *3.2.1 Experiment #1: Mouse Interface*

In this test, the operator can only use the mouse to press the "Yes" buttons appearing on the screens (see Fig. 4). His reaction time and the number of right/wrong actions are measured.

The results obtained by each individual are detailed in Table 4. It should be pointed out that all of them were used to work with the mouse.

### *3.2.2 Experiment #2: Touch Screen Interface*

This test is like the previous one, but using the touch screen interface. It allows to compare both input technologies in order to evaluate which one is better suited for the station considered. The more efficient input method (mouse or touch screen) will be used in the following experiments.

Table 5 shows results better than those obtained with the mouse interface. In order to quantify the benefit of the touch screens, the percentage of reduction in the mean

**Table 5** Summary of the results for the experiment #2

| Individual | $n$ | $n_{\text{right\_yes}}$ | $n_{\text{right\_no}}$ | $n_{\text{wrong\_no}}$ | $\overline{T_{\text{yes}}}$ (ms) | $\sigma$ (ms) |
|---|---|---|---|---|---|---|
| #1 | 343 | 162 | 180 | 1 | 1210.2 | 442.2 |
| #2 | 338 | 174 | 164 | 0 | 1271.2 | 339.9 |
| #3 | 348 | 167 | 181 | 0 | 1171.9 | 279.9 |
| #4 | 359 | 185 | 174 | 0 | 1110.7 | 305.6 |
| #5 | 356 | 168 | 188 | 0 | 1097.9 | 251.9 |
| #6 | 362 | 190 | 172 | 0 | 1105.4 | 290.7 |
| #7 | 369 | 185 | 184 | 0 | 1032.7 | 242.7 |
| #8 | 371 | 189 | 182 | 0 | 1021.9 | 286.1 |
| #9 | 371 | 195 | 175 | 1 | 1042.5 | 265.8 |

**Table 6** Summary of the results for the experiment #3

| Individual | $n$ | $n_{\text{right\_yes}}$ | $n_{\text{right\_no}}$ | $n_{\text{wrong\_no}}$ | $\overline{T_{\text{yes}}}$ (ms) | $\sigma$ (ms) |
|---|---|---|---|---|---|---|
| #1 | 364 | 182 | 182 | 0 | 1065.5 | 373.5 |
| #2 | 348 | 165 | 183 | 0 | 1160.5 | 268.4 |
| #3 | 363 | 179 | 183 | 1 | 1068.9 | 254.1 |
| #4 | 370 | 184 | 186 | 0 | 1030.8 | 241.7 |
| #5 | 371 | 182 | 189 | 0 | 1001.2 | 232.2 |
| #6 | 372 | 181 | 191 | 0 | 991.0 | 194.9 |
| #7 | 389 | 200 | 188 | 1 | 921.7 | 206.1 |
| #8 | 393 | 216 | 176 | 1 | 943.5 | 277.6 |
| #9 | 381 | 202 | 179 | 0 | 981.2 | 266.9 |

reaction time ($\overline{\Delta\overline{T_{\text{yes}}}}$) and also in the standard deviation of the reaction times ($\overline{\Delta\sigma}$) have been computed for the whole population:

$$\overline{\Delta\overline{T_{\text{yes}}}} = +9.33\%, \ \overline{\Delta\sigma} = +19.73\% \tag{1}$$

Therefore, the touch screen interface has been determined to be better suited for the intended application than the mouse. Then, the touch screens is the input system adopted for the following experiments. Nevertheless, it should be pointed out that both with the mouse and the touch screens, the head of the operators was constantly moving from one screen to another searching for buttons. Then, the required effort to achieve low reaction times was quite high.

### *3.2.3 Experiment #3: Speech Synthesis*

In this experiment, once each button appears on the screen, its label is told to the operator through the speakers. Therefore, two modalities (visual and aural) are involved simultaneously and the potential benefits can be analyzed.

In the interviews after the tests, it was mentioned that the workload is reduced with the speech synthesis as far as the operator can be relaxed until the "Yes" message is received. Then, it was observed that the head was more or less static if several "No" buttons appeared consecutively. Once a "Yes" message was heard, the operator moved his head from one screen to another searching for the "Yes" button (Table 6).

**Table 7** Summary of the results for the experiment #4

| Individual | $n$ | $n_{\text{right\_yes}}$ | $n_{\text{right\_no}}$ | $n_{\text{wrong\_no}}$ | $\overline{T_{\text{yes}}}$ (ms) | $\sigma$ (ms) |
|---|---|---|---|---|---|---|
| #1 | 364 | 168 | 196 | 0 | 1020.4 | 244.5 |
| #2 | 344 | 157 | 187 | 0 | 1173.5 | 183.8 |
| #3 | 376 | 209 | 165 | 2 | 1048.7 | 250.3 |
| #4 | 382 | 193 | 188 | 1 | 953.2 | 241.1 |
| #5 | 374 | 182 | 192 | 0 | 973.8 | 190.2 |
| #6 | 376 | 190 | 186 | 0 | 982.7 | 184.0 |
| #7 | 388 | 208 | 180 | 0 | 954.6 | 243.0 |
| #8 | 380 | 173 | 207 | 0 | 891.6 | 181.4 |
| #9 | 380 | 183 | 196 | 1 | 939.3 | 254.3 |

**Table 8** Summary of the results for the experiment #5

| Individual | $n$ | $n_{\text{right\_yes}}$ | $n_{\text{right\_no}}$ | $n_{\text{wrong\_no}}$ | $\overline{T_{\text{yes}}}$ (ms) | $\sigma$ (ms) |
|---|---|---|---|---|---|---|
| #1 | 364 | 181 | 183 | 0 | 1062.0 | 279.5 |
| #2 | 352 | 157 | 195 | 0 | 1094.9 | 154.2 |
| #3 | 367 | 182 | 185 | 0 | 1041.4 | 235.3 |
| #4 | 392 | 216 | 175 | 1 | 959.7 | 244.9 |
| #5 | 372 | 168 | 204 | 0 | 956.4 | 226.2 |
| #6 | 384 | 201 | 183 | 0 | 958.1 | 190.5 |
| #7 | 375 | 191 | 184 | 0 | 1001.6 | 226.6 |
| #8 | 382 | 188 | 192 | 2 | 948.5 | 202.3 |
| #9 | 379 | 176 | 202 | 1 | 922.1 | 233.2 |

### *3.2.4 Experiment #4: 3D Audio Interface*

This test is like the previous one, but adding the 3D audio technology. Depending on the location of the button on the screens (left, right or middle), the source of audio corresponding to its label is generated synthetically on the left, on the right or in front of the operator respectively through the headset. The goal is to evaluate the potential benefits of the 3D audio w.r.t. the conventional audio.

The results obtained are shown in the Table 7 and compared with the speech synthesis alone, it can be seen than the performance is better. In fact, it could be observed during the experiments that the individuals pointed their head directly on the right screen after hearing the “Yes” message. Then, the workload was lower due to two different factors:

- No need to pay attention while hearing “No” messages.
- Once a “Yes” button appeared, no need to search for the button from one screen to another (focus immediately on the screen with the “Yes” button instead).

### *3.2.5 Experiment #5: Tactile Interfaces*

In this case, three wiimotes are used along with the touch screens. The devices are attached to the left and right arms, and also on the chest. The wiimote vibrator is

**Table 9** Summary of the results for the experiment #6

| Individual | $n$ | $n_{\text{right\_yes}}$ | $n_{\text{right\_no}}$ | $n_{\text{wrong\_no}}$ | $\overline{T_{\text{yes}}}$ (ms) | $\sigma$ (ms) |
|---|---|---|---|---|---|---|
| #1 | 375 | 198 | 177 | 0 | 1017.3 | 336.8 |
| #2 | 360 | 170 | 190 | 0 | 1061.0 | 215.8 |
| #3 | 368 | 197 | 171 | 0 | 1068.5 | 258.5 |
| #4 | 390 | 189 | 200 | 1 | 881.9 | 209.2 |
| #5 | 387 | 191 | 196 | 0 | 910.7 | 204.4 |
| #6 | 387 | 196 | 191 | 0 | 929.2 | 158.8 |
| #7 | 384 | 202 | 182 | 0 | 963.8 | 220.7 |
| #8 | 393 | 200 | 193 | 0 | 879.3 | 197.4 |
| #9 | 389 | 209 | 180 | 0 | 953.7 | 241.4 |

**Table 10** Summary of the results for the experiment #7

| Individual | $n$ | $n_{\text{right\_yes}}$ | $n_{\text{right\_no}}$ | $n_{\text{wrong\_no}}$ | $\overline{T_{\text{yes}}}$ (ms) | $\sigma$ (ms) |
|---|---|---|---|---|---|---|
| #1 | 360 | 174 | 185 | 1 | 1073.6 | 325.1 |
| #2 | 355 | 189 | 166 | 0 | 1154.7 | 304.7 |
| #3 | 345 | 171 | 174 | 0 | 1202.6 | 299.3 |
| #4 | 362 | 176 | 186 | 0 | 1067.9 | 304.5 |
| #5 | 366 | 197 | 169 | 0 | 1092.4 | 380.6 |
| #6 | 364 | 171 | 193 | 0 | 1025.0 | 257.1 |
| #7 | 365 | 172 | 193 | 0 | 1014.0 | 232.2 |
| #8 | 369 | 173 | 191 | 5 | 1023.2 | 281.7 |
| #9 | 384 | 197 | 186 | 1 | 953.0 | 232.0 |

activated every time a "Yes" button appears on the screen. Moreover, depending on the location of the button on the window (left, right or middle), the wiimote on the left, on the right or on the chest vibrates respectively.

Table 8 shows values which are quite similar in mean to those obtained in the last experiment with the 3D audio interface. The reason is that the kind of benefits that the vibrators provide are essentially the same provided by the 3D audio:

- No need to pay attention while there is no vibration.
- Once a vibrator is activated, no need to search for the button from one screen to another (focus immediately on the screen with the "Yes" button instead).

### *3.2.6 Experiment #6: Integrated 3D Audio and Tactile Interfaces*

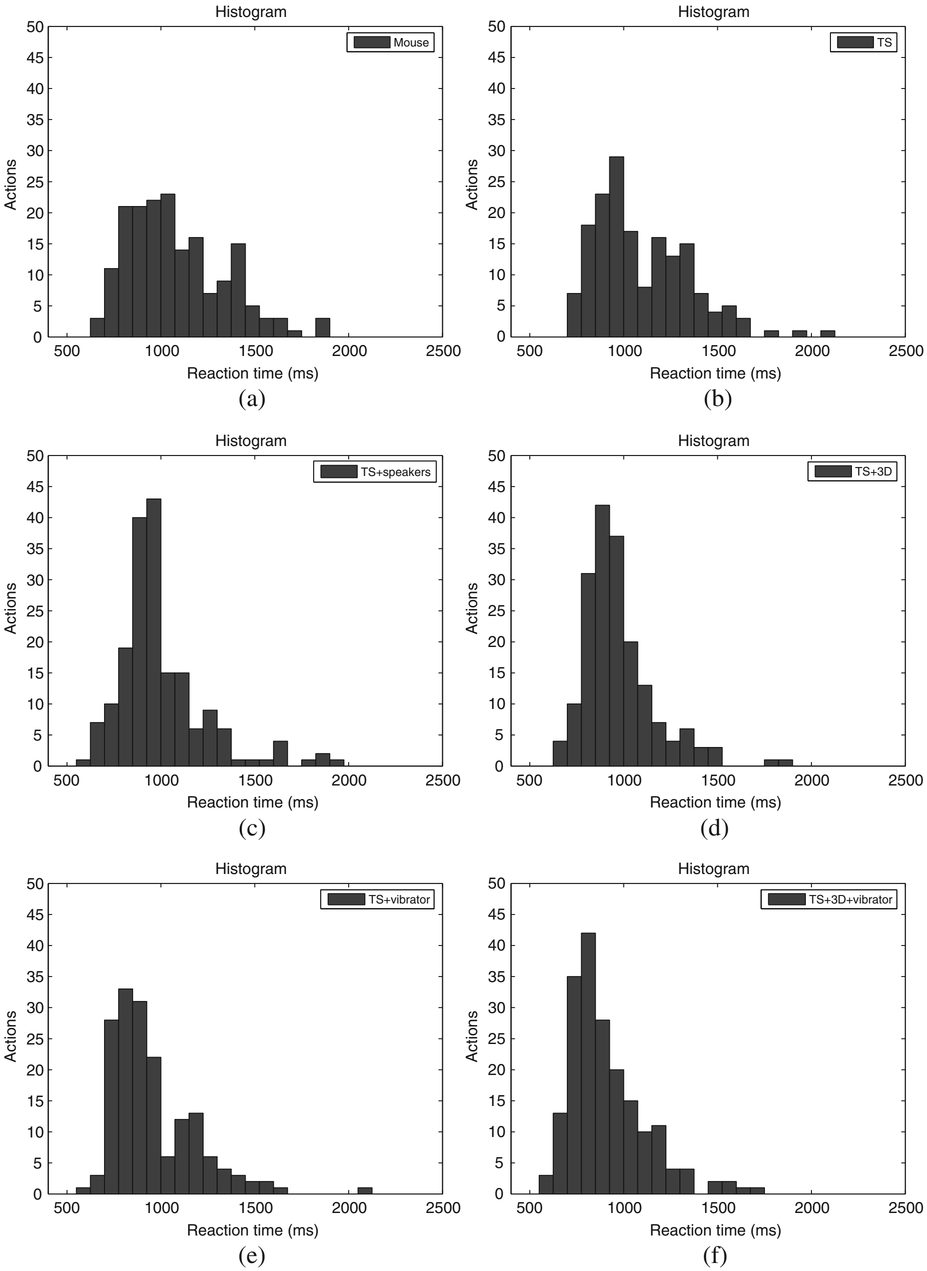
This test is a combination of the modalities involved in the last two experiments. The operator receives redundant information from the 3D audio and tactile interfaces. Then, depending on the location of the button on the screens (left, right or middle):

- the source of audio corresponding to its label is generated synthetically on the left, on the right or in front of the operator respectively through the headset, and
- if the button is a "Yes", the wiimote on the left, on the right or on the chest vibrates respectively.

**Table 11** Summary of the results for individual #5

| Experiment | $n$ | $n_{\text{right\_yes}}$ | $n_{\text{right\_no}}$ | $n_{\text{wrong\_no}}$ | $\overline{T_{\text{yes}}}$ (ms) | $\sigma$ (ms) |
|---|---|---|---|---|---|---|
| Mouse | 358 | 178 | 179 | 1 | 1098.7 | 322.6 |
| TS | 356 | 168 | 188 | 0 | 1097.9 | 251.9 |
| TS+speakers | 371 | 182 | 189 | 0 | 1001.2 | 232.2 |
| TS+3D | 374 | 182 | 192 | 0 | 973.7 | 190.2 |
| TS+vibrator | 372 | 168 | 204 | 0 | 956.4 | 226.1 |
| TS+3D+vibrator | 387 | 191 | 196 | 0 | 910.7 | 204.3 |
| TS2 | 366 | 197 | 169 | 0 | 1092.4 | 380.5 |

In the Table 9, it can be observed that the results are slightly better than those presented in the previous two experiments. Therefore, it seems that the redundant information from the audio and tactile interfaces contributes to improve the performance of the operator.



(a)

(b)

(c)

(d)

(e)

(f)

**Fig. 5** Individual #5: Histograms with the number of correct actions in each reaction time interval for the different experiments (**a**–**f**)

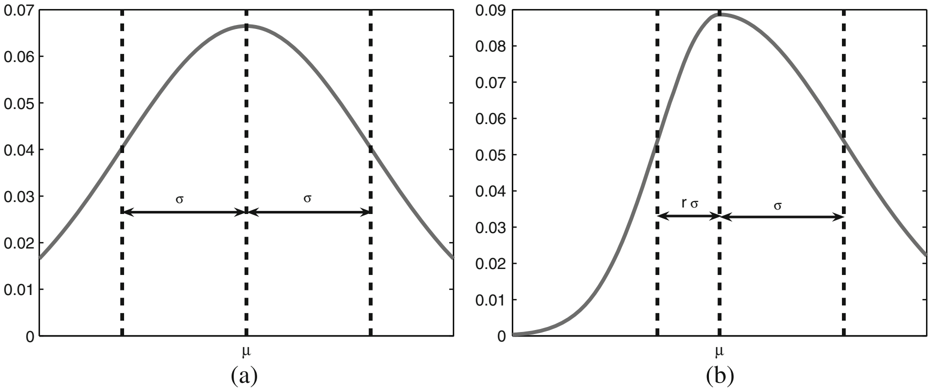


**Fig. 6** Univariate Gaussian (**a**) and univariate asymmetric Gaussian (**b**)

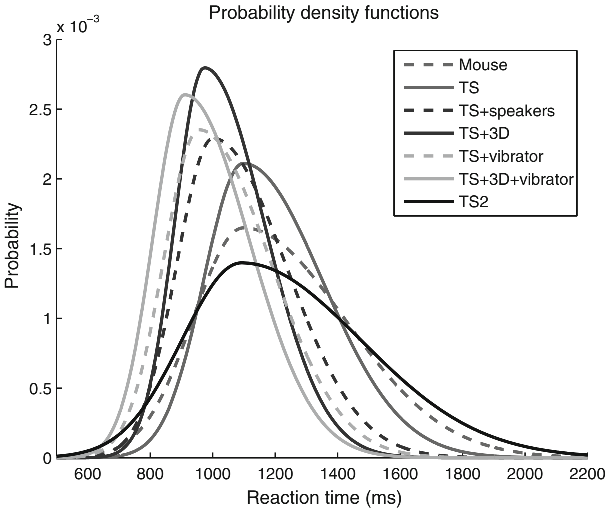
### *3.2.7 Experiment #7: Touch Screen Interface Repetition*

The goal of this test is to check if the learning process of the user has any impact on the results. To fulfill this purpose, the individual is requested to repeat the test using only the touch screen interface after completing all the previous experiments. Comparing the results obtained in Table 10 with those corresponding to the second experiment (see Table 5), no significant improvement from the learning process arises.

## 4 Analysis of the Results

In order to compare in a more exhaustive manner the different technologies involved in the experiments, we will focus on the results from one individual. Table 11 shows several performance parameters of individual #5 in all the experiments.

**Fig. 7** Individual #5 reaction time probability density functions (using an approximation based on the *univariate asymmetric Gaussians* (UAGs) with $r = 0.5$



On the other hand, Fig. 5 contains six histograms corresponding to the first six experiments (from Exp. #1 to #6) with the number of correct actions in several reaction time intervals. From those histograms the idea was to find a probability density function that could approximate them. The approach adopted is depicted in the next subsection.

### 4.1 Probability Density Functions

Taking into account the histograms from the experiments and due to the nature of the measured values, it seems reasonable to use Gaussian distributions as an analytical approach for the results. However, the shape of the histograms computed is not symmetric with respect to the mean value (the decrease at the left is more abrupt than at the right of the mean value). Therefore, it has been considered that the probability model of the *asymmetric Gaussians* (AG) [6], which can capture temporal asymmetric distributions, could outperform Gaussian models.

Let $\chi$ be the random variable associated to the reaction times measured in the experiments presented before. To indicate that a real-valued random variable $\chi$ is normally distributed with mean $\mu$ and variance $\sigma^2 \geq 0$, we write

$$\chi \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{2}$$

The continuous probability density function of the normal distribution is the Gaussian function

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3}$$

where $\sigma > 0$ is the standard deviation and the real parameter $\mu$ is the expected value.

We now introduce an asymmetric Gaussian (AG) model with the following distribution:

$$\varphi_{\mu,\sigma^2,r}(x) = \frac{2}{\sigma(r+1)\sqrt{2\pi}} \begin{cases} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ if } x > \mu, \\ \exp\left(-\frac{(x-\mu)^2}{2r^2\sigma^2}\right) \text{ otherwise} \end{cases} \tag{4}$$

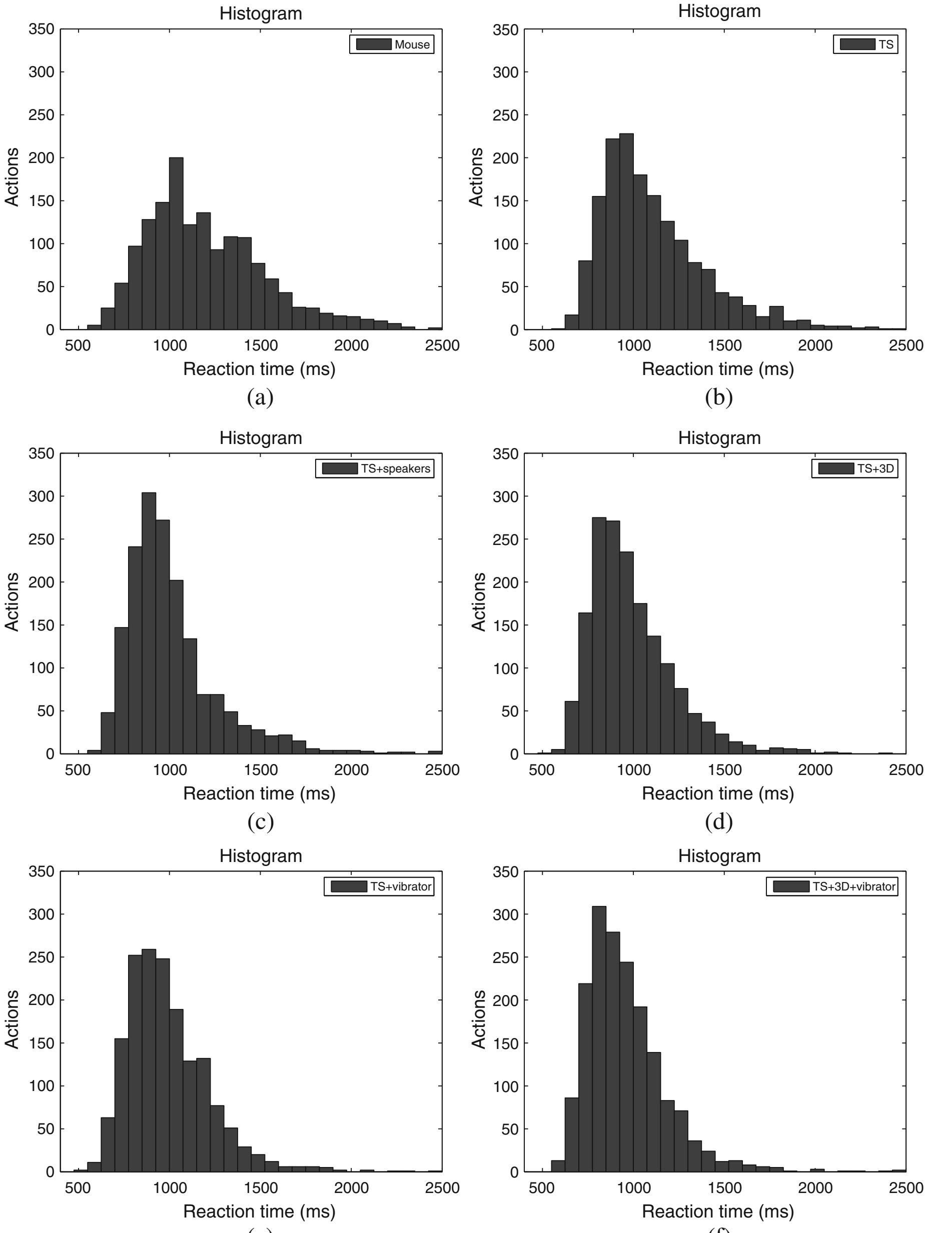
with $\mu$, $\sigma$ and $r$ as parameters. We term the density model (4) *univariate asymmetric Gaussian* (UAG). It is shown that UAG have an asymmetric distribution by the Fig. 6b, where the density function is plotted. In addition, UAG is an extension of a Gaussian since UAG with $r = 1$ is equivalent to the Gaussian distribution.

Then, the next step was to approximate each histogram by an UAG distribution. For example, for the histograms in Fig. 5, the values of $\mu$ and $\sigma$ have been already

**Table 12** Summary of the improvements in mean with respect to the results of Experiment #2 (TS): percentage of reduction in the mean reaction time ($\overline{\Delta\overline{T}_{\text{yes}}}$) and also in the standard deviation of the reaction times ($\overline{\Delta\sigma}$)

| | Mouse | TS | TS+speakers | TS+3D | TS+vibr | TS+3D+vibr | TS2 |
|---|---|---|---|---|---|---|---|
| $\overline{\Delta\overline{T}_{\text{yes}}}$ (%) | −9.33 | 0 | 8.89 | 11.18 | 10.97 | 13.78 | 4.41 |
| $\overline{\Delta\sigma}$ (%) | −19.73 | 0 | 13.91 | 24.93 | 24.46 | 23.68 | 1.04 |

computed (see Table 11). Selecting $r = 0.5$ and plotting the UAGs corresponding to the first six experiments in the same figure, it is possible to compare easily the different modalities tested (see Fig. 7).



(a) (b) (c) (d) (e) (f)

**Fig. 8** Histograms with the number of correct actions in each reaction time interval for the whole population during the different experiments (**a**–**f**)

## 4.2 Comparative Results among Technologies

After collecting the full set of data from all the individuals in all the experiments, it has been processed in order to obtain a general comparison among the technologies involved in the experiments.

In a first step, the improvement in the mean reaction time and in its standard deviation has been computed for all the individuals participating in the experiments (see Table 12). This improvement has been expressed in percentage and computed with respect to the results in the Experiment #2, in which only the touch screens were used (a negative value in the percentage means that the performance was worse).

It can be seen that the progressive introduction of better multimodal technologies from the first experiment to the sixth one improves the performance of the operator. On the other hand, when equivalent technologies are used (i.e. 3D audio or vibrators), the results obtained in mean are quite similar (although each individual could show preference for one of them).

It should be pointed out that there is a "minimum" response time due to the limitations of the operating system and the electronic components and interfaces involved in the system. This minimum response time has been estimated to be approximately 100 ms. Then, if we remove this interval from the computed mean reaction times, the percentages of improvement presented in Table 12 would have higher values.

The histograms for the whole population in the experiments from #1 to #6 are shown in Fig. 8. Comparing this figure with the histograms of the individual #5, it can be seen that when the number of samples increases, the shape of the histograms is more similar to the UAG distribution adopted for the analysis.

Then, using the values of $\mu$ and $\sigma$ for the whole population and with $r = 0.5$ the UAG distributions for the experiments from #1 to #6 are computed and plotted together in Fig. 9. This figure allows to compare the impact of each modality for the whole population at a glance. The Gaussians move from right to left as we use better

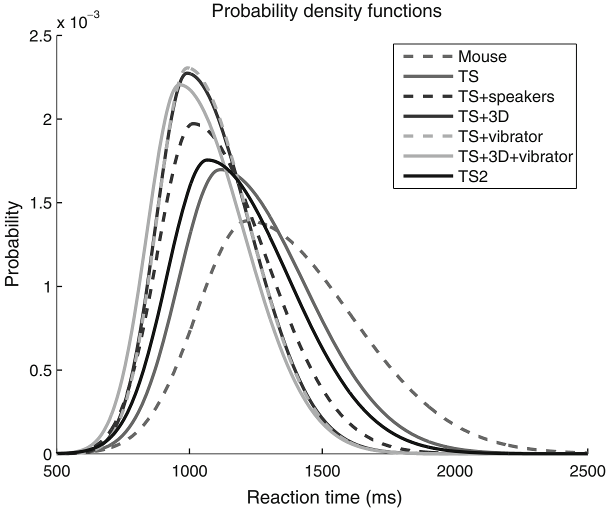


**Fig. 9** Reaction times probability density functions for the whole population in the different experiments

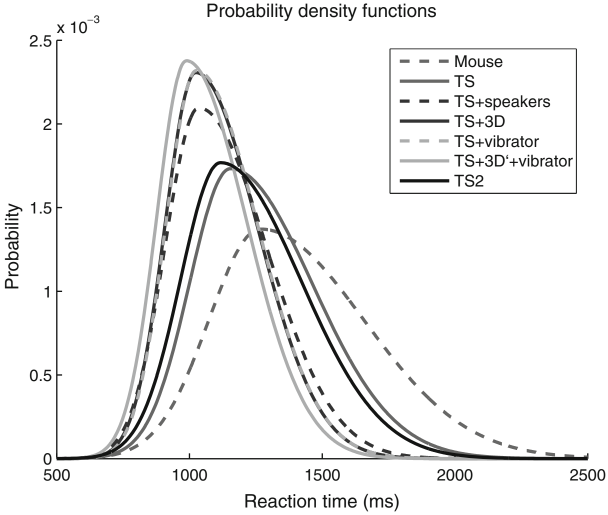


**Fig. 10** Reaction times probability density functions for the whole population in the different experiments considering only the transitions from one screen to another

modalities in the interface, because the mean reaction times are lower. Moreover, the shape of the Gaussians is narrower also from right to left as far as the standard deviation is lower.

Finally, during the experiments it was also registered the screen where each "Yes" button was pressed by the operator, allowing us to compute the reaction times when a transition from one screen to another happened. Using this information, the UAG distributions for the reaction times of the transitions were calculated (see Fig. 10). The Gaussians are slightly displaced to the right with respect to Fig. 9 as expected (the mean reaction times are higher for the transitions) and the benefit of adding modalities is clearer.

## 5 Conclusions and Future Developments

Multimodal display techniques may improve operator performance in the context of a Ground Control Station (GCS) for Unmanned Aerial Vehicles (UAVs). Presenting information through two or more sensory channels has the dual benefit of addressing high information loads as well as offering the ability to present information to the operator within a variety of environmental constraints.

This paper has explored different technologies that can be applied in the design and development of a GCS for UAVs equipped with a multimodal interface. The applicability and benefits of those technologies has been analyzed for a task consisting in the acknowledgement of alerts in an UAV ground control station composed by three screens and managed by a single operator. The system integrated visual, aural and tactile modalities and multiple experiments have shown that the use of those modalities has improved the performance of the users of the application.

Regarding the multimodal application used to obtain the results presented in this paper, there are several possible improvements. One of them would be to compute the exact position of each button on the screen when it is pressed. It will allow to estimate the stochastic relation between the reaction times, the different modalities and the distance between buttons.

On the other hand, the wiimote devices were used in the experiments as wireless vibrators to signal the alarms. But their internal accelerometers can also provide information about the motion of the arms of the operator during the mission, allowing to measure the level of stress for instance.

Finally, it could be interesting to integrate a head-tracking system for the operator in the platform. This system will allow to compute an estimation of the screen at which the head of the operator is pointing at. This information can be used to show each alarm in the screen where the attention of the user is focused, and evaluate its benefits for the operation. Additionally, it can be used along with other body sensors to evaluate the state of the user (level of attention, stress, etc.).

**Acknowledgements** The authors would like to thank the *Boeing Research & Technology Europe* company for their financial and technical support.

## References

1. Biocca, F., Jin, K., Choi, Y.: Visual touch in virtual environments: an exploratory study of presence, multimodal interfaces, and cross-modal sensory illusions. Presence: Teleoperators and Virtual Environments **10**(3), 247–265 (2001)
2. Origin Instruments Corporation: Headmouse extreme. http://www.orin.com/access/headmouse/ (2009)
3. Craven, P., Belov, N., Tremoulet, P., Thomas, M., Berka, C., Levendowski, D., Davis, G.: Foundations of augmented cognition, chap. cognitive workload gauge development: comparison of real-time classification methods, pp. 75–84. Springer, New York (2006)
4. Creative Labs: OpenAL: cross-platform 3D audio library. http://www.openal.org/ (2009)
5. Free Software Foundation: FreeTrack. http://www.free-track.net/english/ (2009)
6. Kato, T., Omachi, S., Aso, H.: Asymmetric gaussian and its application to pattern recognition. In: Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, pp. 405–413. Springer, London (2002)
7. Lemon, O., Bracy, A., Gruenstein, A., Peters, S.: The WITAS multi-modal dialogue system I. In: Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH), pp. 1559–1562. Aalborg, Denmark (2001)
8. Madentec: Tracker Pro. http://www.madentec.com/products/tracker-pro.php (2009)
9. Maza, I., Caballero, F.: Video summarizing the experiments reported in this paper. http://grvc.us.es/JINT_multimodal (2009)
10. McCarley, J.S., Wickens, C.D.: Human factors implications of UAVs in the national airspace. Tech. Rep. AHFD-05-5/FAA-05-1, Institute of Aviation, Aviation Human Factors Division, University of Illinois at Urbana-Champaign (2005)
11. NaturalPoint: SmartNav 4 AT. http://www.naturalpoint.com/smartnav/ (2009)
12. NaturalPoint: TrackIR 4. http://www.naturalpoint.com/trackir/02-products/product-TrackIR-4-PRO.html (2009)
13. Ollero, A., Garcia-Cerezo, A., Gomez, J.: Teleoperacion y Telerrobotica. Pearson Prentice Hall, Englewood Cliffs (2006)
14. Ollero, A., Maza, I. (eds.): Multiple heterogeneous unmanned aereal vehicles, chap. teleoperation tools, pp. 189–206. Springer Tracts on Advanced Robotics. Springer, New York (2007)
15. Orden, K.F.V., Viirre, E., Kobus, D.A.: Foundations of Augmented Cognition, chap. Augmenting Task-Centered Design with Operator State Assessment Technologies, pp. 212–219. Springer, New York (2007)
16. Patterson, R.D.: Guidelines for Auditory Warnings on Civil Aircraft. Civil Aviation Authority, London (1982)
17. Peryer, G., Noyes, J., Pleydell-Pearce, K., Lieven, N.: Auditory alert characteristics: a survey of pilot views. Int. J. Aviat. Psychol. **15**(3), 233–250 (2005)

18. Poythress, M., Berka, C., Levendowski, D., Chang, D., Baskin, A., Champney, R., Hale, K., Milham, L., Russell, C., Seigel, S., Tremoulet, P., Craven, P.: Foundations of Augmented Cognition, chap. Correlation between expected workload and EEG indices of cognitive workload and task engagement, pp. 75–84. Springer (2006)
19. Sharma, R., Pavlovic, V.I., Huang, T.S.: Toward multimodal human-computer interface. Proc. IEEE **86**(5), 853–869 (1998)
20. Cachya Software: Cachya. http://www.cachya.com/esight/overview.php (2009)
21. Stanford University: WITAS Project multi-modal conversational interfaces. http://www-csli.stanford.edu/semlab-hold/witas/ (2009)
22. Sweller, J.: Visualisation and instructional design. In: Proceedings of the International Workshop on Dynamic Visualizations and Learning (2002)
23. EyeTech Digital Systems: EyeTech TM3. http://www.eyetechds.com/index.htm (2009)
24. University of Edinburgh: The festival speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival/ (2009)
25. Wilson, G., Russell, C.: Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. In: Human Factors, pp. 635–643 (2003)
26. Zhu, Y.: Advances in Visual Computing, chap. Measuring Effective Data Visualization, pp. 652–661. Springer, New York (2007)