CrossMark

ORIGINAL RESEARCH

Context acquisition in auditory emotional recognition studies

Davide Carneiro[1,2] · Ana P. Pinheiro[3,4] · Paulo Novais[2]

Abstract This paper describes an environment to assess auditory emotional recognition based on a mobile application. The primary aim of this work is to provide a valuable instrument that can be used both in research and clinical settings, responding to the strong need of validated measures of emotional processing, especially in Portugal. The secondary aim is to acquire and study the participants' interaction behavior with the technological device (e.g. touch patterns, touch intensity), in search for a relationship with medical conditions, cognitive impairments, auditory emotional recognition capacities or socio-demographic indicators. This will establish the basis for the prediction of such aspects as a function of an individual's interaction with technological devices, potentially providing new diagnostic tools.

Keywords Ambient intelligence · Soft sensors · Behavioral biometrics · Auditory emotional recognition

✉ Davide Carneiro
dcarneiro@di.uminho.pt

Ana P. Pinheiro
ana.pinheiro@psi.uminho.pt

Paulo Novais
pjon@di.uminho.pt

1 CIICESI, ESTGF, Polytechnic Institute of Porto, Felgueiras, Portugal

2 Algoritmi Centre/Department of Informatics, University of Minho, Braga, Portugal

3 School of Psychology, University of Minho, Braga, Portugal

4 Faculty of Psychology, University of Lisbon, Lisbon, Portugal

1 Introduction

Conversations between individuals, especially those that are undertaken face-to-face, are very rich in content. Indeed, the effectiveness of the conversation process depends not only on what is actually said (i.e. the words) but also on other equally important aspects such as our body language (e.g. gestures, body posture, physiological signs) and the characteristics of the speech.

In fact, speech is an acoustically rich signal that provides significant information about the talker. Therefore, the ability to convey emotions (production and perception) is fundamental for the success of the communication (Juslin and Laukka 2003).

Auditory emotional recognition refers precisely to the ability of a listener to infer emotion from speech. When studying auditory emotional recognition, the standard perception paradigm is to have listeners choose which one of several emotion words best characterizes linguistically neutral utterances made by actors attempting to portray various emotions (Bachorowski 1999). One important aspect here is that of emotional valence, which encodes the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an emotion.

In practice, the most traditional approach involves setting up the experimental trials, controlling for stimulus presentation and timing through software such as Presentation[1] (Neurobehavioral Systems, Inc., Albany, CA, USA) or Superlab[2] (Cedrus, San Pedro, CA). The few behavioral

---

[1] Presentation is a stimulus delivery and experimental control program typically used in neuroscience and behavioral research. https://www.neurobs.com/.

[2] Superlab is an environment for setting-up and running experimental studies, providing accuracy and reaction time measures. http://www.superlab.com/.

measures that are often the focus of those sudies (e.g., accuracy rates, reaction time) are usually obtained by recording the participants' responses directly via the software, or by using a paper-and-pencil approach.

This typical approach is often time-consuming, prone to errors (e.g. when the results are passed from the paper to the computer), and dependent on the availability of the software and equipment in the context in which the behavioral data are required. Furthermore, in these traditional approaches, one frequent issue is the lack of motivation of the participants to perform the task, as well as the fatigue arising from the experimentally controlled procedure that typically requires a high number of stimuli. Also, for special populations such as children and older adults, answering to the questions in paper may be particularly problematic, an issue that is exacerbated when the subjects present attention deficits or motor coordination problems.

This paper introduces an innovative instrument to assess auditory emotional recognition that can be used both in research and clinical settings, centered on a Tablet. The user interacts with a mobile application to provide feedback about the auditory stimuli. To do so, the participant selects which one of several emotion words (arranged in buttons and set by the expert when defining the study) best characterizes the emotion felt.

Although the prototype is being developed in the context of vocal emotional recognition, it may be easily adapted to other areas. Compared with more traditional assessments, this application will provide a faster and more dynamic way of assessing vocal emotional recognition in healthy subjects as well as in clinical populations. Moreover, this application incorporates concepts from context-aware computing, allowing a real-time analysis of the participants' behavior in terms of their interaction with the computer.

The proposed environment will accomplish the following goals.

– To provide a reliable assessment instrument that might be used to test healthy individuals of different ages, and clinical populations characterized by impaired vocal emotional processing. This instrument responds to the strong need of standardized assessment instruments in Portugal.
– To provide an engaging assessment instrument that might be easily used in different contexts and that might detect, prevent and counteract the reduction of motivation and attention, as well as the increase of fatigue, which are frequently observed during the administration of traditional assessment instruments, particularly in younger or older participants.
– To provide an assessment instrument that might be used with individuals from different ages and backgrounds, and potentially in different countries;
– To develop an assessment instrument that is intuitive and easily configurable by non-experts, significantly facilitating the process of study design and data collection and analysis.

The remaining of the paper is organized as follows. After introducing the background and motivation of the work, the current section also addresses relevant aspects from the field of Human-Computer Interaction and how previous work of the research team was integrated, namely to identify fatigue or distraction. Section 2 describes the client-server architecture of the environment, detailing the funcionalities developed and how they were implemented in both the server and the mobile application client. Section 3 details the features that characterize the behavior of the participant, their importance and the process of acquiring the relevant data. Section 4 details the first case-study in which this system will be used and the process of classifying the participant's state. Finally, Sect. 5 presents the conclusions of this work.

1.1 Context awareness and human–computer interaction

Context-aware Computing refers to systems that are aware of their user's state and surroundings and that may adapt accordingly with the goal of improving some aspect of the user's actions (Perera et al. 2014). Knowledge about the environment or the user may be relatively simple (e.g. the network to which the device is connected, the devices in the proximity) or may be more complex and even built from assumptions about the user's current situation.

In its inception, research on user context was focused mostly on user location (Dey 2001). The hype revolved around applications that would provide personalized services according to their user's location. Nevertheless, in the last years the notion of context has been significantly widened and now refers not only to *where* the user is but also to *who* the user is, *what* the user is doing and *when* he is doing it, possibly inferring *why* he is doing it.

These novel and increasingly complex contextual models provide the support for applications that are able to adapt interfaces, improve information retrieval techniques, target services more efficiently or use implicit user-interaction techniques (Bolchini et al. 2007). In fact, a very promising and growing research field at the moment stems precisely from the intersection of Context Awareness and Human-Computer Interaction. The key idea is to search for contextual features in the interaction of the user with a given device (e.g. computer, smartphone). The first evident advantage of this approach is that it requires no additional hardware for acquiring contextual features. On the contrary, traditional approaches tend to rely on sensors or other

specific hardware, which tends to increase system complexity and costs. Then, it also imposes no changes on the individual's routine and can be used continuously throughout the day, significantly improving and facilitating extensive data collection.

A growing number of studies have recently highlighted the potential of this intersection in a range of different applicational fields. Epp et al. (2011) looked at its potential for inferring human emotions from the rhythm of the individual's typing patterns on a standard keyboard. The author showed that the influence of emotions on typing behavior is significant to the point of supporting the training of emotion classifiers. On a similar approach, Vizer (2009) presented some preliminary results on the influence of stress on keystroke behavior. Wang et al. (2014) used smartphones together with questionnaires to study students' stress, concluding that the fluctuation of the students' levels of stress during the academic term can be estimated from their interaction with smartphones. Hernandez et al. (2014) pointed out how measuring stress in the workplace in a practical way remains a challenge and presented an approach based on a pressure-sensitive keyboard and a capacitive mouse. By doing that he showed that stress can be estimated from the pressure exerted by people on these peripherals.

A similar approach is embedded in this work, which improves the proposed application with very interesting additional features. First of all, it allows the continuous study of the participant's behavior during the study, providing previously ignored features which may prove interesting for the aim of the study (e.g. do participants with different conditions interact differently with the smartphone?; can we predict these conditions from interaction patterns?). The continuous analysis of the behavior will also be important for a more practical and immediate application: the detection of fatigue, boredom or distraction of the participants. These phenomena are frequently observed during the administration of traditional assessment instruments, particularly with younger or older participants, and can have a negative impact on the results of the study. The following subsection briefly depicts previous work of the research team in this field.

1.2 Previous work

In the last years we have been building expertise in the fields of Context-aware Computing and Human-Computer Interaction. Specifically, we address these issues from a behavioral standpoint, in which the behavior of the user towards the technological device is seen as a very important aspect of the interaction itself (Carneiro et al. 2014): much like our body language or our facial expressions are what give a true meaning to what we say and do. Therefore,

wee look at users' interactions as a bridge to their inner state, as we look at our interlocutors' facial expression in search for their feelings. Our ultimate aim is to design software and hardware that can not only be sensitive to the user's state but also respond appropriately according to the environment's objectives.

Specifically, we have been focusing on two main trends. On the one hand, we have been studying how stress influences our interaction with the computer and with smartphones, especially in an educational context, to conclude that it is possible to train stress classifiers for both types of device (Castillo Montoya et al. 2012; Carneiro et al. 2015). On the other hand, in a similar vein, we have also been studying the effects of fatigue, especially in the interaction with the computer. Among other aspects, we have studied the role of cognitive workload, the fluctuation of fatigue throughout the day or the emergence of distraction and the drop in motivation that usually comes with it (Pimenta et al. 2015a, b).

This second strand is, in the context of the present work, more important. In fact, we are currently integrating previous work on mental fatigue detection in this environment. The main advantage is that we can analyze the participants' interaction patterns in real-time in order to assess aspects that may have a critical importance to the interpretation of research results. Specifically, the application will continuously look for signs of mental fatigue or distraction identified in previous work (e.g. increased reaction time, longer touches). This will allow the controller of the experiment to intervene if necessary, namely by making a pause, preventing the participant to loose focus or motivation.

The use of these previously studied interaction features will be combined with new ones, namely those related to this new domain of application. Namely, we aim to investigate if stimuli conveying certain emotions result in touches with different characteristics (e.g. duration, intensity, area) or in differences in the way the user holds the device (e.g. amount of movement, acceleration). This may open the door to the development of emotional mobile applications that do not require the use of physiological sensors.

The inclusion of these characteristics in this scope has, to our knowledge, never been addressed before. It represents, therefore, one of the most innovative and promising aspects of this work.
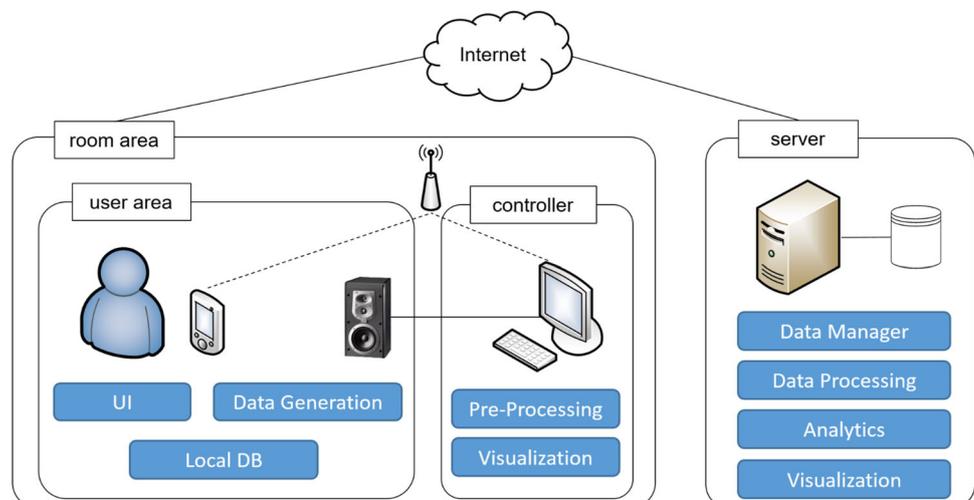
2 Architecture

The architecture of the proposed system was designed to support the case-study described in Sect. 4. It was designed according to the client-server model (Oluwatosin 2014). The client is the mobile application running in the tablet

that collects all the relevant data. The client also implements all the graphical interfaces that allow the participants to provide feedback about the emotions conveyed by each stimulus heard and communicates wirelessly with the local computer in order to play those stimuli. Anytime a new audio clip must be played (e.g. when the participant moves forward in the task) the client sends a request to the local computer with the reference of the clip to be played. All the audio clips are thus stored in the local computer, which makes the mobile application lighter and more easily configurable. The server receives all the collected data and stores them in a MongoDB database for a posterior analysis.

From a logical standpoint, four main components can be identified in the architecture (Fig. 1):

– *User Area* This is the room where the participant is during the study, including the tablet used for interaction and a Logitech 5.1. loudspeaker system connected to the computer of the controller. The tablet contains the following main modules:

  – *UI* The user interface with which the participant interacts during the study;
  – *Data Generation* Includes all the mechanisms that generate data, including operational data (e.g. user actions in the application) and behavioral data (e.g. interaction events such as touches and their characteristics);
  – *Local DB* A mysql database exists in the mobile device that maintains the configurations of several studies. These configurations are managed from the controller's computer.

– *Controller* The room where the controller is, adjacent to the user area, including the computer that controls the audio clips playing in the loudspeaker system and

shows real-time information about the participant's progress and state. The controller is composed of the following modules:

  – *Pre-Processing* The data is received from the tablet in real time by the controller. This module aggregates the data and performs some pre-processing before sending the aggregated data in blocks, at regular intervals, to the server;
  – *Visualization* Provides visual feedback to the controller about the participant (e.g. task progress, performance indicators);

– *Room Area* Depicts the whole physical environment where case studies take place (user area and controller), equipped with a wireless network that ensures the constant connection between the tablet and the controller's computer;
– *Server* The remote computer that receives and maintains all the data collected during the case studies, detailed in Sect. 3. Multiple environments for data collection may exist for one server. The server has the following main modules:

  – *Data Manager* This module implements an interface for accessing the MongoDB database. All other modules interact with it in order to access the database (i.e. read/write);
  – *Data Processing* Processes blocks of data, generating high-level information describing interaction patterns or emotion recognition performance;
  – *Analytics* Contains algorithms for conducting larger automated analyses of data, namely including all the participants of a given study;
  – *Visualization* Provides graphics to support data visualization and analysis (e.g. histograms, box-and-whisker charts).

Fig. 1 Architecture of the data acquisition framework

The proposed architecture has several key aspects that are worth highlighting. First of all, it allows for a distributed and concurrent collection of data. In other words, several participants may be simultaneously participating in different rooms. This significantly reduces the time required to conduct extensive studies. The fact that the server is logically and physically detached from the data collection environment allows for studies to be carried out in any environment, as long as there is an internet connection and the necessary hardware (i.e. tablet, loudspeakers and local computer).

In order to create a new data collection environment it is only necessary to install the software in the computer and in the tablet, which takes only a few minutes.

We also believe that this approach will be significantly more engaging, especially for studies with young participants, which will result in decreased fatigue and frustration and increased commitment during the studies. Nonetheless, the extent of this effect will be studied in more detail in the future.

However, the most innovative and promising aspect are the features provided by the application which were until now disregarded in similar approaches. These features, described in detail in Sect. 3, will allow taking into consideration additional variables while studying auditory emotional recognition or other aspects, namely behavioral and physical measures.

2.1 Server side

The server side of the architecture, implements a group of functionalities that aim to improve the usability of the system from the controller's point of view. These functionalities are described next. They are supported by a local database and the local file system in which stimuli are kept.

2.1.1 Stimuli management

This group of functionalities allows the controller to manage the available stimuli. While this particular work focuses on auditory stimuli, other types of stimuli are often used in different studies. Thus, the system provides support for different types of stimuli. Stimuli are added in two steps: (1) the controller moves the file of the desired stimulus to a location in the computer's file system; and (2) the controller adds the new stimulus to the application (and consequently to the database), assigning it an unique identifier, a name, a description, a type (e.g. sound clip, movie, image) and, optionally, key words. These key words allow a more efficient search mechanism (e.g. search all the stimuli that have the keywords "emotion" and "positive valence").

After this moment, the stimulus is available to be used in the application. Its properties can be edited at any time.

The controller can also conveniently access the stimulus from within the application (e.g. visualize an image, play a sound) without the need to go through the operating system's file explorer.

Finally, the controller can also delete the stimulus. In order to delete it from the application, it is enough to remove it using the provided graphical interface. To find the stimulus to be deleted, the controller can use several mechanisms (e.g. an advanced search tool, a drop down list). The controller can then choose whether to remove the stimulus' file from the local file system as well.

2.1.2 Research design

This is, undoubtedly, the most important group of functionalities regarding the server side. Here, the controller is able to set-up a new study. Each study is briefly characterized by a unique identifier, a name and a description, as well as the creation date and the controller' ID.

After its creation, the controller can proceed to configure the study. Briefly, each study is composed of a list of tasks (each task generally consisting of a stimuli and a group of options). Each task contains the following elements:

- *sequence_no* A sequence number that encodes the order in which tasks are presented to the participant;
- *stimulus_id* The unique identifier of the stimulus that will be presented to the participant in this task;
- *duration* If the task has a duration, it is specified here in milliseconds. Otherwise it contains a null value;
- *repeat* Contains a boolean value that determines if the participant is allowed to repeat the stimulus (e.g. to play the sound again);
- *options* Besides an identifier that is unique in the task and determines the order in which the options will appear to the participant, each option contains the following elements:

  - *type* The type of the option to be presented to the user: either an image or a text;
  - *content* If the option is of text type, this field contains the text to be presented to the participant as an option. If it is an image the content is the unique identifier of the resource file, which is stored in the server's database and retrieved in real-time to be shown to the participant;

The controller can manage all these aspects through a convenient graphical interface. Therefore, he/she may add tasks to the task list, with the associated stimuli, duration (if applicable) and repetition policy. For each task the controller can also conveniently add the possible answers either in the form of text (a case in which the controller

```
1 ▾ {
2 ▾   "stimuli_list": [
3 ▾     {
4         "sequence_no": 0,
5         "stimulus_id": "0pySVeRt1Z",
6         "duration": 10000,
7         "repeat": true,
8 ▾       "options": [
9 ▾         {
10 ▾          "a": {
11              "type": "1",
12              "content": "Happy"
13            }
14          },
15 ▾        {
16 ▾          "b": {
17              "type": "1",
18              "content": "Sad"
19            }
20          }
21        ]
22      },

23 ▾     {
24          "sequence_no": 1,
25          "stimulus_id": "0pySVeRt1Z",
26          "duration": null,
27          "repeat": false,
28 ▾        "options": [
29 ▾          {
30 ▾            "a": {
31                "type": "2",
32                "content": "6WlLkZnZBn"
33              }
34            },
35 ▾          {
36 ▾            "b": {
37                "type": "2",
38                "content": "5XrZnvKLFE"
39              }
40            }
41          ]
42        }
43      ]
44   }
```

inputs the text of each answer) or in the form of images (a case in which the controller picks the resource id from the database). These options can be of mixed types, i.e., it is possible to have, at the same time, text and images as possible answers in a single task. By default, these options are presented to the participant in the order they are added by the controller. The controller may, however, rearrange this order or choose to randomize it.

When the controller finishes the configuration of the study, this will be saved in the database and, at this point, he/she may exit the application. When the controller wants to run a specific study (a process which is described in detail further below) a JSON (JavaScript Object Notation) representation of the study is generated, to be sent to the client. JSON is a lightweight data-interchange format that is easy for humans to read and write as well as for machines to parse and generate. It is based on a subset of the JavaScript Programming Language.

This JSON representation of the study is sent to the client application, which generates the corresponding Android activities in real-time to implement the study, including the graphical interfaces for the participant to interact with. This is, in fact, one of the most innovative points of the system: there is no need for the intervention of a Mobile Application specialist as the total configuration of the mobile application for each individual study is carried out by the controller through an intuitive desktop application. A demonstrative example of the code generated to implement a very simple study with two tasks is depicted in Fig. 2. The first task (stimulus_id = 0pySVeRt1Z) has a maximum duration of 10 s and the user may repeat the stimulus. The participant must answer by choosing one of two texts ("Happy" or "Sad"). The second task has no

duration and the user cannot repeat the stimulus. In this case, the participant must answer by selecting one of two images.

2.1.3 Data export

Another interesting feature is the Data Export, which implements a transparent access to the collected data. Here, the controller is able to select the study and participants for which data are to be exported. As previously mentioned, all the collected data are stored in a MongoDB database. This set of functionalities implements an intuitive interface for the database that, upon selection of the necessary parameters, will provide as output one CSV file (Comma-Separated Values) for each study and each participant. These files contain a log of all the events that took place during the study. These events include not only those specific to the study (e.g. moving to a specific task, playing or repeating stimulus, answer of a participant), as well as all the data collected to compile the interaction features described further below (e.g. touches on the screen, acceleration on the device).

This allows a detailed analysis of the participant's behavior throughout all the study and for each individual task. For example, the controller is able to determine, for each task, the time spent, the number of repetitions of the stimulus, the movement of the handheld device, the intensity of the touch (e.g. does the user reacts differently when he is presented with a stimulus of a given emotional valence?) among many others.

The presentation of the collected data in this format significantly facilitates the access to the stored information, and does not require the involvement of an expert in

Databases. It also facilitates data analyses procedures: since data are extracted in a format that is consistent between studies, the analysis procedures will be similar from study to study.

2.1.4 Server management

This set of features is a relatively simple but useful one. It implements the management of authorized users, and allows adding, editing or excluding users with the role of controller, which are authorized to create and manage studies and collections of stimuli. Users are necessary to track the actions carried out (e.g. who created or edited a given study, who administrated it, who added a given set of stimuli).

This part of the application also implements the management of existing studies. With it, it is possible to visualize, edit or delete existing studies.

2.2 Client side

The client-side part of the system is relatively simpler, being composed of two main parts, described in detail in the two following sub-sections. The first part is responsible for all the operational aspects, whereas the second implements the transparent collection of behavioral data.

2.2.1 Operational aspects

This part of the mobile application is responsible for the interaction with the device and for the configuration of the application according to the data received from the server. The most important aspect here is undoubtedly the automatic creation and configuration of all the graphical interfaces based on the message detailing the study design received from the server in JSON format.

It works as follows. When the mobile application is started, it attempts a connection to the server. If the connection is established, the application enters a stand-by mode while the controller assigns a particular study to this client. This is done in the server-side. When this is done, the mobile application receives a message with the JSON describing the study and generates all the corresponding graphical interfaces. Nonetheless, the actual collection of the data will only start when this process is over and when the controller explicitly orders it, from the server. In other words, data collection only starts when both controller and participant are ready to start.

After the task is initiated by the controller, the process of data collection is implemented as follows (Fig. 3). The first stimulus is presented to the participant, accompanied by the possible answers shown in the tablet. If the stimulus is an auditory one, the client sends a message to the controller
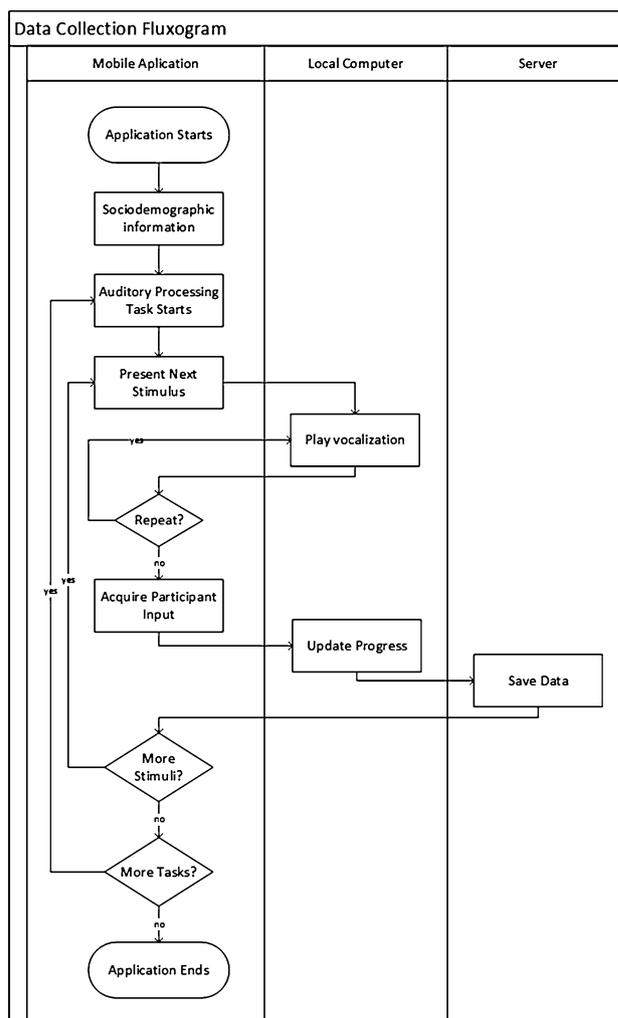
Fig. 3 Fluxogram depicting the process flow that implements the data acquisition procedure

containing the unique identifier of the stimulus to be played in the sound system. If the stimulus is a visual one, the client retrieves the image from the server based on the resource identifier. A similar process takes place with the response options in the case of images. The participant responds to each stimulus using the options shown in the tablet. The tablet sends, in background, the participant's answer and the interaction features to the local computer, to update the data shown in real-time. The tablet also sends the same data to the server, so that it is stored in the database. The mobile application then moves forward to the following task, if there is one.

2.2.2 Collection of interaction features

Another very important aspect in the mobile application is the collection of interaction features. These features, described in more detail in Sect. 3, allow the controller to

have an in-depth view of the participant's behavior during the study in an unprecedented way. As such, they will allow the detection of significant changes in the state of the participant and that were not considered in previous approaches. In particular, we are interested in determining if the emotional valence of the stimulus has an effect on the participants' interaction patterns with the device. This would potentially allow to develop mobile applications that are sensitive to the user's perceived emotion. Moreover, we are also interested in detecting signs of fatigue, frustration or stress, which are very important to ensure the validity of the studies conducted. This possibility has already been explored successfully in previous studies (Carneiro et al. 2014, 2015; Castillo Montoya et al. 2012; Pimenta et al. 2015a).

In order to collect data describing the interaction of the users with the device, the application is designed with a transparent layer that is placed on top of all graphical interfaces. This layer captures all user interactions, registers the necessary data (e.g. intensity, area) and then forwards the touch event to the next layer, allowing the application to perform normally and to react to user's input. This constitutes a transparent and non-intrusive way of acquiring this data, which is critical not to affect the results of the study.

3 Acquisition of interaction-related features

One of the most innovative aspects of this environment is the collection, for the first time in studies of this kind, of interaction-related features. In previous work we have established that aspects such as stress or fatigue significantly influence our interaction with technological devices (e.g. computers, tablets, smartphones) (Carneiro et al. 2014, 2015; Castillo Montoya et al. 2012; Pimenta et al. 2015a). However, we have only done so with healthy individuals, in environments such as the workplace or the classroom.

In these milieus we found that mental fatigue slows our interaction down: our clicks become a few milliseconds slower as well as our key presses. The movement of the mouse also becomes generally slower. We have also established that stress influences the acceleration measured on our handheld devices when we hold them, as well as the intensity of our touches or their duration.

We now aim to use this kind of approach in the domain of auditory emotional recognition, with populations with potentially very different characteristics. All this information, previously not considered in such studies, may reveal very interesting cues about the participant's inner state and the influence of emotional valence at new levels (e.g. physical, behavioral).

The proposed instrument provides access to a set of interesting additional features, that make it significantly more appealing than traditional ones. In what concerns the process itself, from an operational point of view, the following can be pointed out.

– Graphical interfaces can be adapted to different participants according to aspects such as age (e.g. more appealing interfaces for children, more accessible interfaces for elderly or people with cognitive or physical impairments).
– Participating in the study becomes a more engaging process and potentially less tiresome. This will result in more accurate and valid results.
– Collecting and storing data can be a fully automatized process, without the interference of human experts, without using paper and without the need for a human expert to input the answers of the participant into a computer.
– The process becomes highly dynamic and configurable. Applications for studying different aspects can be rapidly developed and implemented by using the existing infrastructure;
– The progress of each participant can be followed in real-time, both in the control room and on the server-side.
– The collected data are available for analysis as soon as the participant ends the task or experiment.
– Data analysis processes can be implemented that automatically present the intended results upon data collection.

From the point of view of vocal emotional processing, this approach provides both accuracy rates and reaction time measures for:

– General auditory perception.
– Vocal discrimination.
– Emotional vocal discrimination and recognition, as a function of gender, age and type of emotional category.

The way these measures are constructed is further detailed in Sect. 4.

Let us now focus on the features acquired from Human-Computer Interaction which constitute, in our opinion, one of the most innovative aspects of the presented work. Indeed, we rely on the use of a tablet as an interface with the aim of collecting additional information about the participants. Specifically, we are interested in studying the participants' interaction patterns with the tablet in search of underlying individual factors (e.g. sociodemographic indicators, cognitive/sensory impairments, differences in emotional processing).

To acquire interaction-related features we make use of the event listeners provided by the Operating System of the

tablet. An event listener is an interface in the `View` class that contains a single callback method that will be activated by the Android framework when the View to which the listener has been registered is triggered by user interaction with the item in the UI. For the purpose of this work, the application uses the `onTouch()` callback method, which is activated when the participant performs an action qualified as a touch event, including a press, a release, or any movement gesture on the screen. Thus, in each touch of the participant on an item of the UI several touch events are fired: one when the finger of the participant first touches the screen (identified by the action event `ACTION_DOWN`), several while the participant is touching (depending on the duration of the touch), and one when the finger releases the screen (identified by the action event `ACTION_UP`).

Each of these events has information about the intensity of the touch (via the `getPressure()` method), quantifying the pressure exerted on the screen, and about the position of the event. Moreover, when each event is fired the application registers it with its timestamp. This allows tracking the evolution of a touch in terms of its intensity over time, as described below.

The following features are extracted from the participant's interaction with the tablet:

– *Touch duration* The duration of a touch is defined as the difference between the timestamps of the events `ACTION_UP` and `ACTION_DOWN`. We expect that touches of participants of different ages or with different conditions may have different duration. Moreover, and as determined in previous work (Carneiro et al. 2013), we expect touch duration to increase with fatigue or to decrease with stress. Therefore, this may represent a very relevant variable to assess the state of the participant in real-time. Moreover, it may be used to better understand the results.
– *Touch intensity* The intensity of a touch event depicts the force exerted by the participant's finger when touching the device.. Given that each touch event includes a pressure and that each touch fires several touch events, it is possible to analyze the mean intensity of each touch as well as the variation of the intensity throughout all the touch.
– *Touch area* The area of the touch is obtained, as in the previous feature, from each touch event. Similarly, it is possible to analyze the mean area of each touch as well as its evolution during the touch.
– *Intensity/area curve* These two features can be seen as composite features, derived from sequences of the two previously mentioned features. They provide additional information regarding the behavior of the participant.

– *Acceleration* The acceleration, measured in the three axis by the built-in accelerometer, provides insights on the way the participant holds and interacts with the tablet. In previous work, we demonstrated that participants with higher stress levels tended to move their hands and their body less than individuals with lower stress levels (Carneiro et al. 2013). We expect that this feature proves to be equally important in the context of this study.

The features described in this section constitute a significant improvement when compared to the traditional approach that relied on paper-and-pencil and was heavily dependent of human experts. The last group of features, in particular, may prove especially interesting and promising in what concerns the characterization of the participant's behavior and state during the study. Equally important is the fact that these features can be acquired non-intrusively and in a way that is completely transparent to the participants, just by analyzing their interaction patterns with the device.

We expect that, among other aspects, these features may point out fatigue, frustration or stress during the participation of the subjects in the experiments. We will also search for correlations between some of these features and emotional recognition impairments.

4 Case-study

The environment developed will be tested and validated with a specific case study, that will allow to evaluate its suitability. This section describes the case study.

The participant sits alone in a room listening to audio clips, which he/she must classify according to the emotions conveyed. While in the past this process was carried out manually, we now aim to automatize it and make it more engaging for the participant. For this reason, the participant will have access to an Android tablet, which will provide feedback about each audio clip. The controller sits on an adjacent room, monitoring the participant in real-time and ready to intervene if necessary (Fig. 4).

The local computer will contain stimuli in the form of nonverbal vocalizations conveying different emotional expressions that correspond to the six basic emotions (anger, sadness, fear, disgust, happiness, surprise). The vocalizations will be uttered by five male and five female speakers with training in acting, and from different age groups (childhood 8–10 years; adolescence 14–16 years; adulthood 30–32 years; older adulthood 60–62 years). Recordings will be made in a quiet room with an Edirol R-09 recorder and a CS-15 cardioid-type stereo microphone, with a sampling rate of 22 KHz and 16-bit
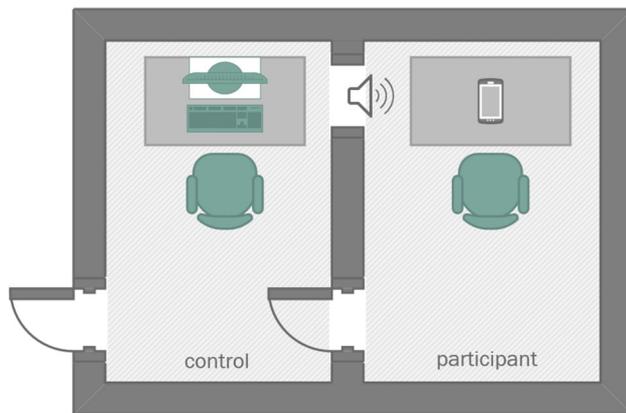
Fig. 4 Layout of the environment: the participant sits alone in a room, listening to vocalizations through a sound speaker and interacting with the tablet, which is wirelessly connected to the computer in the control room

quantization. With these stimuli, we will be able to evaluate three major dimensions of voice processing: gender identification; age identification; emotion discrimination and recognition. The stimuli will be first validated with a sample of healthy participants (n = 50). The stimuli will be thoroughly described in terms of their physical features (e.g., pitch and intensity) using Praat software. Stimuli will be pseudorandomized and presented in consistent order across subjects. The final set of stimuli will include vocalizations from 10 children, 10 adolescents, 10 adults and 10 older adults (five male and five female individuals in each age group).

For each subject, the first part of the process will consist in using the mobile application to fill in relevant sociodemographic information, such as gender, age, and education. After that, four short tasks evaluating simple auditory processing will be presented, including:

– *Tone matching task* This task is designed to measure pitch perception. Subjects are presented with pairs of 100 ms tones with 500 ms inter-tone intervals and are asked to indicate whether they sounded the same or different. In order to minimize learning effects, three different base tones are used: 500, 1000, or 2000 Hz. Half the stimuli pairs are identical and half include manipulated tones, which differed in frequency by 1, 2.5, 5, 10, 20, or 50 % of their counterpart.
– *Intensity task* This task is designed to measure sound intensity. Subjects hear 2 stimuli and are asked to indicate which one was softer than the other. Stimuli difference begins at 13.8 dB from standard stimulus and changes adaptively, based on subject's performance, to a maximum of 19.5 dB.
– *Auditory Timing Task* This task measures time perception. Two sounds will be played in succession. Subjects

are asked to indicate which stimulus was shorter. One stimulus is always 96 ms, while the other ranges from 112 to 608 ms. An additional auditory "masked" task using a distractor noise in between tones acts a control comparison to examine effects of attentional impairments on task performance.
– *Rhythm Task* The rhythm task is an auditory task designed to measure rhythmic perception. Subjects hear 2 5-note rhythms and are asked to indicate which has an extra gap. The monotonic version uses 1000 Hz tones, while the polytonic version presents tones in random pitches from an octave divided into seven steps.

After a brief general auditory assessment, participants will complete a brief training regarding the tasks on vocal emotional processing. The aims of the tasks will be carefully described.

– *Non-emotional vocal discrimination* Participants have to discriminate between questions or statements, by indicating if linguistic prosody is the same or different (16 trials).
– Vocal emotional discrimination: pairs of vocalizations will be presented in the same or different emotional tone, and the participants will be instructed to judge if the two sounds are the same or if they are different using "yes" or "no" buttons.
– *Vocal emotional recognition* Participants will be asked to identify the emotion (six possible options) of the portrayal. For each emotion, five males and five female trials will be presented, for each age category (childhood, adolescence, adulthood, older adulthood). Stimuli will be pseudorandomized. Participants will be instructed to listen to the vocalization and to recognize the emotion as quickly and as accurately as possible. Next, they will be asked to judge the stimulus intensity on a scale of 1 to 10 (no intense-extremely intense).
– *Conflicting vocal emotion* Participants will listen to the seven types of vocalizations (neutral and 6 emotions) paired with congruent or incongruent facial expressions, and they will have to indicate if they are congruent or not using "yes" or "no" buttons.

4.1 Classification of the participant's state

In this work we are also concerned with the analysis of the participant's state. Specifically, we want to detect signs of mental fatigue, stress, lack of attention or frustration. This is important to assess the validity of the results attained, as well as to design better studies. For example, we may be able to determine the maximum length of the study before the participants start to feel bored or even fatigued, which may significantly influence the outcomes of the study.

Moreover, we also want to determine if emotional valence has an effect on interaction patterns. For example, do participants tend to press the screen of the tablet harder when presented with an emotional stimulus with positive valence? Or, do participants tend to move less when presented with a vocal stimulus with negative valence? Or also, do participants with impaired vocal emotional processing interact differently with handheld devices? The answers to these questions will potentially open the door to the development of better mobile applications, with increased sensitivity and reactivity to the users' state.

In order to achieve this, we will train models that characterize the interaction patterns of both healthy and clinical populations, under different circumstances. Specifically, we will create datasets of interaction for these populations when under stress, fatigue and when under a baseline normal condition. This will support the training of classifiers to distinguish between these states, and that can be used to classify the state of the participants in real-time.

In previous work, we have already validated this approach. Specifically, we analyzed the intensity curve of touches in the screen of a handheld device (Castillo Montoya et al. 2012). For the purpose, we developed a task in which participants performed mental calculations and memorization of results, using the handheld device to provide their task-related input. Participants did this in two different settings, with and without stressors. The stressors used in the task included social pressure of other participants, self-assessment (through continuous update of the participant's performance) and annoying sound and vibrations on the device (Carneiro et al. 2013).

This resulted in two different datasets describing the interaction of the participant with the device. These datasets included all the features detailed before, as well as the score of the participants in the task, acting as an additional measure of the effects of stress on the participant.

In this case-study, the datasets will include additional features describing the performance of the participant in the auditory emotional recognition task. New so-called operational features will also be included, that are extracted from the progress of the task. These features include aspects such as the time spent to classify each auditory stimuli, the number of necessary repetitions (if allowed), among others. Much like the interaction features, these features are compiled by the server based on the events sent by the client application.

After conducting a statistical analysis of the datasets, the intensity curve was selected to be analyzed in more detail, as a preliminary validation of the approach. To tackle this problem, we explored the fact that the intensity from all the touches follows a similar shape: a convex curve that grows to a maximum point and then decreases. Thus, the approach was to fit a second polynomial degree curve to each intensity curve. For this purpose, we used a fit function of a statistical software which finds a least-squares fit to a list of data.

An example of this approach is depicted in Fig. 5b. Given that the second degree polynomial curves are of the type $y = ax^2 + bx + c$ we can compare the parameters of the curve of each touch: similar values of a, b and c indicate similar curves, thus similar touches.

Based on these observations we moved on to select and train a classifier able to distinguish between the two groups of touches, which takes as input three numeric attributes a, b and c (the independent variables), and a nominal attribute that describes the state of the user at the time of the touch as stressed or not stressed (the dependent variable).
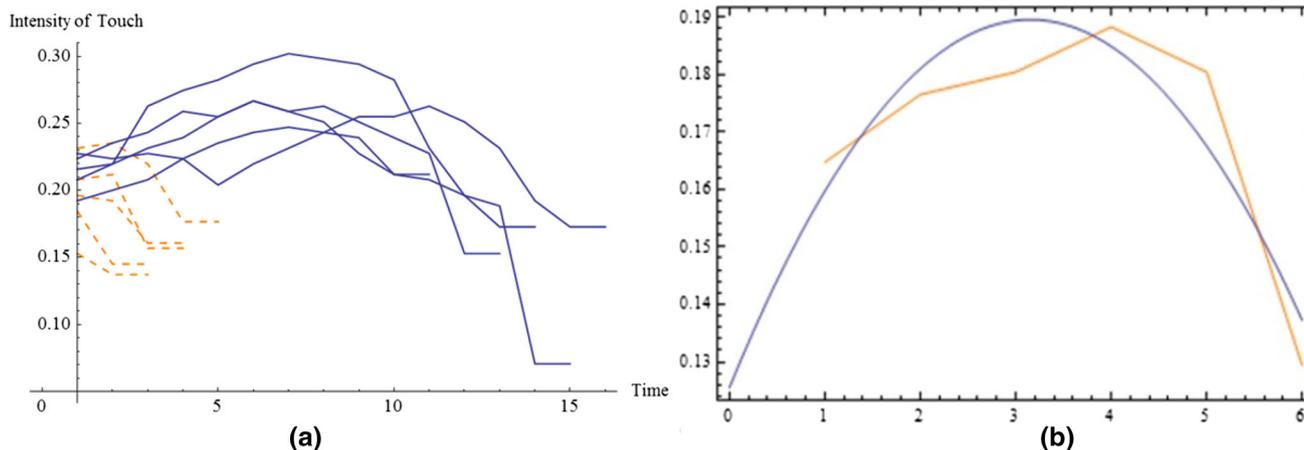
Fig. 5 **a** 10 different touches from participants: touches can be composed of a different number of touch events. The orange dashed lines depict touches classified as "calm" whereas the *blue solid lines* depict touches classified as "stressed". **b** Fitting a polynomial curve (*blue curve*) to a given touch (*orange line*)
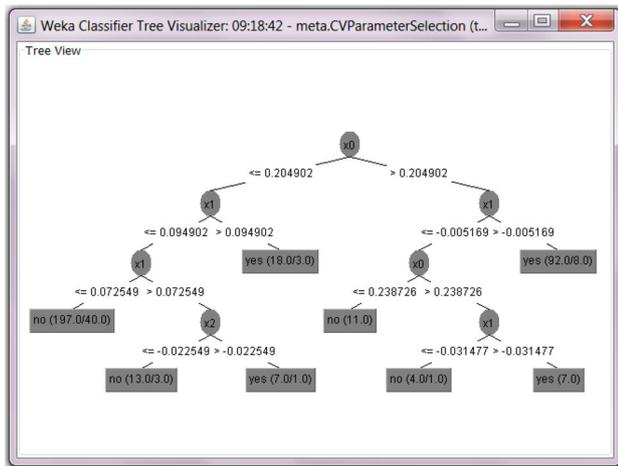
Fig. 6 Decision tree to classify a touch as "stressed" or "not stressed" from its shape

Since selecting the optimal parameters for an algorithm may be a rather time-consuming process, we used a meta-classifier provided by weka that allows to optimize a given base-classifier.

Specifically, we used the `weka.classifiers.meta.CVParameterSelection`. After finding the best possible configuration of parameters, the meta-classifier then trains an instance of the base classifier with these parameters and uses it for subsequent predictions. The meta-classifier was used with lower bound 0.01, upper bound 0.5 and 10 optimization steps.

When using the J48 classification tree as the base classifier for the meta-classifier, the model is able to correctly classify 271 out of the 349 instances, which amounts to 77.6504 %. The Kappa statistic for this model is 0.5434 and the value of the ROC area is 0.796. The constructed tree has a size of 15 nodes and a total of 8 leaves (Fig. 6). In this tree, attributes `x0`, `x1` and `x2` correspond to the values of `a`, `b` and `c` of the polynomial curve, respectively. Given this, it is possible to use the rules of this tree to build a classifier capable of distinguishing between stressed and calm touch curves.

When the SMO function is used to build a classifier, the results achieved are similar. In fact, the correctly classified instances amount to 79.9427 % (279 out of 349), the value of the Kappa statistics is 0.5809 and the value of the ROC area is 0.781. These results also show that a classifier can be trained with these data to distinguish between stressed and calm touches. Given that the results of both classifiers are similar, we decided to use the J48 tree since it can easily be used by our system to classify touches in real time.

After this preliminary validation of the approach, we will now include the remaining features into the process, in order to make it multi-modal and thus more accurate.

Ultimately, we aimat developing extensive and broad interaction models that characterize the participant's interaction with the handheld device in a very detailed way. Moreover, we aim to use these classifiers in real time so that the controller may conduct the study in a way that increases its validity, as well as the comfort and engagement of the participant.

5 Conclusion

Auditory emotional recognition assessments are nowadays highly dependent on human experts and on pen-and-pencil. They can thus become a tedious process for participants, especially in the case of children who get easily bored and, consequently, distracted. This may have a significant negative effect on the results of the assessment. Moreover, a large part of the expert's time is spent introducing the feedback of the participant into the computer, also with an additional risk of human errors that may compromise the results.

In this paper, we presented an approach to automatize and improve this type of assessments. The developed environment respects all the identified requirements of traditional approaches. It presents, however, a group of additional advantages. First of all, the subject interacts with a mobile application, through a tablet, whose graphical interfaces can be personalized according to aspects such as age, gender or limitations (e.g. vision problems). This will constitute a more engaging approach than pen-and-pencil approaches, more appealing to participants and thus less tedious. We expect subjects to be more motivated and engaged throughout the experiment or assessment session, with more accurate results.

However, the use of the tablet as an interaction resource also includes other interesting aspects. Specifically, we extract a set of interaction features that describe, to a large extent, how the participant handles the tablet. In previous work and in different contexts, we determined that the interaction with technological devices correlates with aspects such as stress or fatigue. On the one hand, we expect to see the same correlations here, allowing the controller of the assessment to intervene if the subjects show significant signs of fatigue, distraction or stress. On the other hand, we also expect to find different interaction patterns in subjects with different characteristics or clinical conditions. This approach could thus be a step towards the extraction of potentially interesting new features in the context of auditory emotional recognition assessments, not taken into consideration until now.

Finally, and from an operational point of view, the proposed approach will facilitate this type of assessments by automatizing the whole process of task design, task

implementation and data collection and analysis. Moreover, it will allow the analysis of raw data and high-level information as soon as the study ends, making the process more dynamic and efficient.

Therefore, the approach described here will not only facilitate the implementation of this type of studies and similar ones, but it will also significantly increase the validity of the research developed in this scientific domain. A promising avenue for future studies with human subjects is, thus, open.

References

Bolchini C, Curino CA, Quintarelli E, Schreiber FA, Tanca L (2007) A data-oriented survey of context models. ACM Sigmod Record 36(4):19–26

Carneiro D, Novais P, Gomes M, Oliveira PM, Neves J (2013) A statistical classifier for assessing the level of stress from the analysis of interaction patterns in a touch screen. Soft computing models in industrial and environmental applications, Springer, pp 257–266

Carneiro D, Novais P, Neves J (2014) Conflict resolution and Its context: from the analysis of behavioural patterns to efficient decision-making. Law, Governance and Technology Series, vol 18. Springer

Carneiro D, Novais P, Pêgo JM, Sousa N, Neves J (2015) Using mouse dynamics to assess stress during online exams. Hybrid Artif Intell Syst, Springer, pp 345–356

Castillo Montoya JC, Novais P, Fernandez Caballero A, Carneiro D (2012) Multimodal behavioural analysis for non-invasive stress detection. Expert Syst Appl 39(18):13376–13389

Dey AK (2001) Understanding and using context. Personal ubiquitous Comput 5(1):4–7

Epp C, Lippold M, Mandryk RL (2011) Identifying emotional states using keystroke dynamics. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp 715–724

Hernandez J, Paredes P, Roseway A, Czerwinski M (2014) Under pressure: sensing stress of computer users. Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, pp 51–60

Jo-Anne Bachorowski (1999) Vocal expression and perception of emotion. Curr Dir Psychol Sci 8(2):53–57

Oluwatosin HS (2014) Client-server model. IOSR J Comput Eng (IOSR-JCE) 16(1):67

Patrik PN, Laukka P (2003) Communication of emotions in vocal expression and music performance: different channels, same code? Psychol Bull 129(5):770

Perera C, Zaslavsky A, Christen P, Georgakopoulos D (2014) Context aware computing for the internet of things: a survey. Commun Surv Tutor IEEE 16(1):414–454

Pimenta A, Carneiro D, Novais P, Neves J (2015) Detection of distraction and fatigue in groups through the analysis of interaction patterns with computers. Intell Distrib Comput VIII, Springer, pp 29–39

Pimenta A, Carneiro D, Novais P, Neves J (2015) A discomfort-sensitive chair for pointing out mental fatigue. Ambient Intell Softw Appl, Springer, pp 57–64

Vizer LM (2009) Detecting cognitive and physical stress through typing behavior. CHI'09 Extended Abstracts on Human Factors in Computing Systems, ACM, pp 3113–3116

Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, Zhou X, Ben-Zeev D, Campbell AT (2014) Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM, pp 3–14