Future Generation Computer Systems 89 (2018) 646–658

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Using behavioral features in tablet-based auditory emotion recognition studies

Davide Carneiro [a,c,*], Ana P. Pinheiro [b], Marta Pereira [b], Inês Ferreira [b], Miguel Domingues [b], Paulo Novais [c]

[a] CIICESI, ESTG, Polytechnic Institute of Porto, Portugal
[b] Faculdade de Psicologia, Universidade de Lisboa, Lisbon, Portugal
[c] Algoritmi Centre/Department of Informatics, University of Minho, Braga, Portugal

HIGHLIGHTS

- We present a distributed system that digitalizes auditory emotion recognition interventions.
- We provide new variables that enrich this kind of interventions.
- We analyze the relationship between Emotion, Age, Gender and Human–Computer Interaction.
- We train a gender predictor based on behavioral features.

ARTICLE INFO

Article history:
Received 14 December 2017
Received in revised form 3 May 2018
Accepted 10 July 2018
Available online 18 July 2018

Keywords:
Auditory emotion recognition
Human computer interaction
Real-time analytics
Machine learning

ABSTRACT

The recognition of emotions in spoken words is one of the most important aspects in human communication and social relationships. Traditional approaches to the study of vocal emotional recognition involve instructing listeners to choose which one of several words describing emotion categories best characterize linguistically neutral utterances or vocalizations uttered by actors portraying various emotional states. To this end, generic experiment control software is usually used, which has some disadvantages. In this paper, we present a system that digitalizes the whole process involved in understanding how people perceive and understand vocal emotions, improving data collection, processing and analysis. Moreover, this system provides a new group of features that allows a more comprehensive characterization of the behavioral dimension underlying vocal emotional recognition. In this paper we describe this system and analyze the relationship between emotional perception, gender, age and Human–Computer Interaction.

1. Introduction

Face-to-face conversations are very rich contexts, in which not only verbal meaning is conveyed but also non-verbal cues are communicated [1], that is, information that is conveyed by gestures, postures, intonation or speech rhythm. These aspects are essential for an efficient process of verbal communication between speaker and listener, and allow one to perceive the emotional state, intentions or personality traits of the other. The ability to convey and to accurately and rapidly decode emotions is fundamental for the success of communication and social interactions [2].

This paper introduces an innovative instrument to assess auditory emotional recognition that can be used both in research and clinical settings, focused 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 conveyed by the voice. The participant also classifies the valence, authenticity and intensity of the emotion that was expressed. While developed specifically for the field of auditory emotion recognition, the system can be easily adapted to other domains. Compared with more traditional assessments, this application aims to 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 [3], Ambient Intelligence [4] and Behavioral Biometrics [5], providing an innovative and interesting plethora of new variables that will significantly enrich these studies.

1.1. Related work

This multidisciplinary work brings together research from different fields, including computer science (namely human–

* Corresponding author at: CIICESI, ESTG, Polytechnic Institute of Porto, Portugal
E-mail address: dcarneiro@estg.ipp.pt (D. Carneiro).

https://doi.org/10.1016/j.future.2018.07.013
0167-739X/

computer interaction) and psychology. In this section we review some of the related work in these fields. While there are many works in the field of auditory emotion recognition in the field of psychology, these are all from a purely psychological perspective. Similarly, many researchers have studied human–computer interaction (although not so many have studied it with older users). However, this literature review shows that these two fields have never been brought together in the past.

From a psychological perspective, auditory emotion recognition refers to the capacity of a listener to infer emotions from sounds in the environment, including the voice. Studies in the last decades have consistently demonstrated differences in the processing of neutral vs. emotional cues. For example, compared to neutral cues, emotional vocal stimuli tend to capture more attention resources (e.g. [6,7]) to be associated with faster responses (e.g. [8]), and to increased arousal ratings (e.g. [9,10]. It is worth noting that neutral and emotional vocal cues are distinguished very rapidly in the brain, with some studies indicating differences already within 50 ms after stimulus onset (e.g. [11]). Further, there is also evidence demonstrating that vocal emotions are more accurately decoded by female than by male listeners (e.g. [9,10]), particularly negative vocal sounds [10], which highlights the need to account for individual differences in emotion perception.

Moreover, the existing evidence indicates that emotions conveyed by the voice are perceived categorically (e.g. [12]), and suggests that some emotional categories are more easily identified than others (e.g. [9,10,13]). For example, vocal surprise and fear are typically associated with low accuracy rates and confused with each other due to their similar acoustic profile (e.g. [10]).

When studying auditory emotion recognition, the standard perception paradigm is to instruct listeners to choose which one of several emotion words best characterizes semantically neutral utterances or nonverbal vocalizations uttered by actors portraying different types of emotions [13,14]. In addition, listeners may be asked to classify the sound in several affective dimensions, such as its valence (a continuum ranging from *unpleasant* to *pleasant*), arousal (from *calm* to *aroused*), and dominance (from *controlled* to *in control*) [5]. Other important dimensions include the intensity of the emotion that was communicated, as well as its authenticity (genuine vs. acted emotions are processed differently [15]). Common approaches in emotion research involve setting up the experimental trials, as well as controlling for stimulus presentation and timing through software such as Presentation (Neurobehavioral Systems, Inc., Albany, CA, USA) or Superlab (Cedrus, San Pedro, CA). The measures that are often the focus of those studies (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 transferred from the paper to the computer), and dependent on the availability of the software and equipment in the context where behavioral data are required.

The other field that supports this research is that of computer science, including disciplines such as Human–Computer Interaction [16]. Human–Computer Interaction seeks to study the relationship between humans and technological devices, albeit the focus of this work on mobile devices with tactile screens. Nonetheless, this discipline aims at the design of interactive computer systems that are efficient and easy to use, to which contribute (among others) task complexity, cognitive skills of the user (especially the temporal aspects of interaction) and physical skills (namely psycho-motor performance).

These factors are even more relevant when the users are older people, who are generally more prone to have diminished cognitive and physical skills. When designing ICT for older users, one key issue is to understand the impact of their abilities and restrictions, as the ageing process causes important changes in perceptual and motor skill capabilities. However, the inclusion of older people within the design cycle for information technology is until now limited to aspects such as usability or the graphical aspects of user interface.

Some authors investigated how touchscreen devices have affected the usability of interactive consumer products by older adults [17]. This work was conducted with older adults to explore their perceptions of touchscreen interfaces and to understand existing usability issues and barriers to their adoption. The research was conducted with only four participants and each was required to carry out common tasks on mobile phones which they were unfamiliar with. The main conclusion is that older adults find it easier to use touchscreens than more traditional interfaces. In [18], on the other hand, the authors focus on the design implications when developing devices for older people, with a focus on the exponential growth of the elderly population that suffers from age-related disabilities. In their work, the authors provide a set of guidelines in order to achieve accessibility in mobile interfaces for older people.

In [19], the authors survey the needs and wishes of the elderly regarding mobile applications and tablets, using a questionnaire. They summarize different methods integrated into a user-centered design approach to develop design concepts for a tablet computer. In [20] the authors also address Human–Computer Interaction considerations in applications and hardware in the domain of smart living for elderly.

Most of the research in existing literature adopts fairly similar approaches: they focus on the design needs or on how user experience must be adapted for the elderly. Which is, undoubtedly, also a necessary effort. However, and as this literature review shows, little is known about the psychomotor performance of technological devices for older adults [21]. Focus must be moved from the devices/applications and their development to the elderly users. That is, how does interaction change with age? How is it affected by specific cognitive or physical conditions? These aspects need to be further studied by such multidisciplinary projects.

Hence, the current work represents an opportunity to understand the impact of these changes and to characterize (in terms of HCI) the population that will constitute, in the near future, the majority of technology users. To this end, we will follow an approach that this research team has developed to model, in the past, the interaction of people with technological devices. Specifically, we have shown how stress [22] and emotions [23] can be quantified from our interaction with handheld devices. This will allow for the development of the first model describing the older adult's interaction with technological devices.

Summarizing, the following conclusions can be drawn from the literature review conducted:

- Memory remains plastic even in an older age and can be improved through cognitive training strategies;
- Some of the existing methods for memory training are non-invasive but are intrusive, and cannot be broadly used;
- Methods focused on the use of mnemonics may improve memory in specific tasks, but have limited influence on daily activities;
- Memory training techniques should be personalized and engaging to improve the resulting outcomes;
- Little is known about the psycho-motor performance of technological devices for older adults.

2. Architecture

In Auditory Emotion Recognition studies, there are usually two actors involved: the participant, whose capacity to evaluate subjective dimensions of an emotional sound is being assessed, and the

researcher, who is playing the stimuli or monitoring the software that does so, as well as registering the participant's responses. In this new approach, the same two roles exist. However, there is now a looser coupling between them. That is: now, the researcher does not need to be in the physical presence of the participant as the participant interacts with a tablet to classify emotions. Moreover, there was previously a relationship of 1:1 between these two elements: one researcher dealt with one patient at a time. Now, one researcher may be collecting data from multiple patients at the same time.

There are several main components in the architecture, as illustrated in Fig. 1. The mobile application, implemented in Android, exists in the user area, in which all user interactions takes place. The application receives the details of each study as determined by the researcher in the Controller room and dynamically generates the corresponding user interfaces. It also acts as a data-generation device, collecting not only the participants' responses but also data describing their behavior during the study. These data are collected in a transparent way and sent to the remote database if it is available and/or if there are no network restrictions. If this is not possible, all collected data are saved in CSV files in the Controller's computer, to be later uploaded into the database. Multiple instances of the Room area can exist simultaneously, i.e., this approach supports multiple simultaneous data collection procedures.

The computer in the controller area communicates directly with the mobile application during the studies. The researcher uses this computer to configure new studies or manage existing ones as well as to start a new data collection procedure in a specific device. The computer is connected to a Logitech 5.1 Surround Sound System, which is used to presented the auditory stimuli as requested by the mobile application. The data collected can be visualized in real time in this computer by the researcher.

The server is the computer where data is stored after the conclusion of the data collection procedure. It includes tools for data processing, analysis and visualization. Specifically, data can be stored by adding it directly to the Mongo database (by the client application), when this is available remotely. However, this is not always possible, namely due to network constraints. Thus, alternatively, the client application stores the collected data in CSV files when it cannot connect to the database. These files are then manually copied by the researcher to the server, which uses Hadoop to process them and store the resulting data in the database.

While the Hadoop Distributed File System could have been used for storing the files permanently, we choose to use Mongo instead since it facilitates data representation, access and processing, especially through Mongo's aggregation framework, which is a higher-level mechanism for data processing than Hadoop's MapReduce. Besides the Mongo database (used for storing raw and processed data) and the Hadoop ecosystem (used for simple data processing/transformation tasks), there is also a Spark installation that is mostly used for Machine Learning tasks and higher-level data processing.

Finally, the Analytics component allows for data, in their different levels, to be visualized and interpreted by the researcher. These levels include raw data, processed data (as detailed in Section 3.2) or different aggregations (e.g. touch intensity by gender, correctly classified emotions by age group). Some of the visualizations generated by this component are depicted in Section 4.

Communication between the Controller's computer and the tablet is implemented by means of a sockets API. Messages are exchanged in JavaScript Object Notation (JSON) format. This format is a lightweight, text-based, language-independent data interchange format. It was derived from the ECMAScript Programming Language Standard. JSON defines a small set of formatting rules for the portable representation of structured data. JSON can represent four primitive types (strings, numbers, booleans and null) and two structured types (objects and arrays). This is convenient since data (collected data and existing studies) is stored in a Mongo database, which is a document-based storage in which documents are stored as JSON objects.

These messages are used, for instance, by the mobile application to request the controller computer to play a given auditory stimulus, or by the controller computer to send the configuration of a study to the mobile application.

The following two sub-sections describe, in more detail, the functionalities of the control computer and the mobile application.

2.1. Life-cycle

From a high-level perspective, the life-cycle of the developed system is as follows:

1. The researcher creates or selects an existing study and uploads it to the intended mobile device using the client application in the Controller area;
2. The participant begins by providing the necessary personal (e.g. sociodemographic) information;
3. The participant goes through a training phase, in which instructions on how to proceed are provided, while simulating the tasks to be performed during the study. The aim is that the participant gets familiarized with the process and with the type of stimuli that will be presented;
4. Upon finishing training, the participant may choose to repeat the previous step;
5. The participant starts the actual experimental task. Each study is composed of one or more iterations. In each iteration:
   (a) The participant hears a stimulus (a nonverbal vocalization — e.g. laughs, growls) played by the sound speakers. She/he may repeat it any number of times by clicking on a button in the graphical interface;
   (b) The participant classifies the stimulus according to the perceived emotion: disgust, fear, anger, sadness, neutral, surprise, relief, amusement, pleasure, or triumph;
   (c) The participant classifies the stimulus according to its perceived valence, using a 9-point likert scale (1 = extremely unpleasant; 9 = extremely pleasant);
   (d) The participant classifies the stimulus according to its intensity, using a 9-point likert scale (1 = not intense at all; 9 = extremely intense);
   (e) The participant classifies the stimulus according to its authenticity, using a 9-point likert scale (1 = extremely unauthentic/posed; 9 = extremely authentic).

2.2. Control

In what concerns the control computer, there are three main functionalities worth highlighting (Fig. 2).

The first functionality is the management of existing studies or protocols, which is implemented by the first tab of the graphical interface. The first list in this tab shows the studies that currently exist in the server while the second shows the details (i.e., stimuli being used) of a study after it is selected. The date of the creation of the study as well as its description are also shown below. It is possible to delete the selected study or send it to a specific instance of the mobile application, running at a given IP address. The interface also allows filtering studies or stimuli within studies by name, and it reproduces the auditory stimuli locally whenever one is clicked. Finally, the interface also includes functionalities to check the communication with the client application.

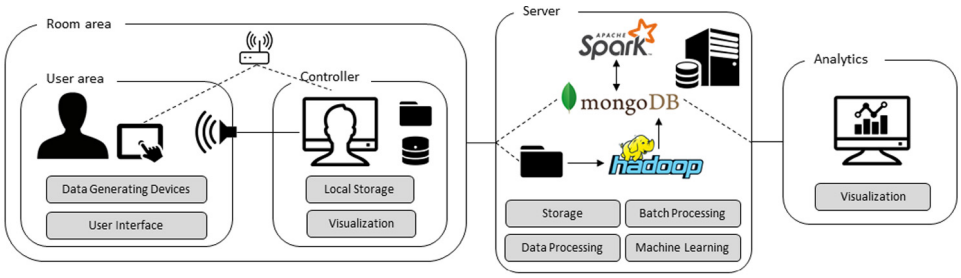


Fig. 1. Architecture of the proposed solution.

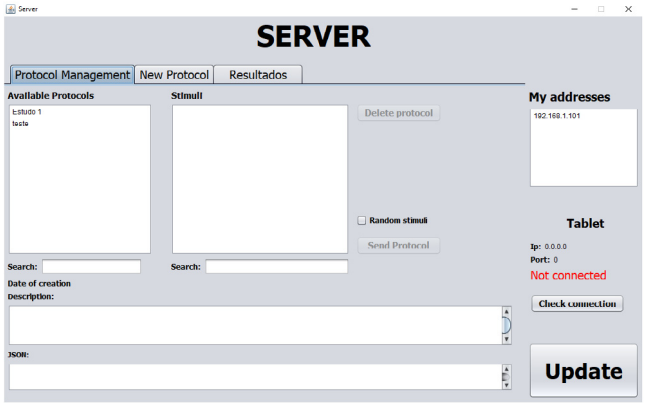


Fig. 2. Graphical Interface of the control computer with its three main functionalities in three different tabs: study management, study creation, results.

The second functionality, implemented in the second tab, is the creation of new studies or protocols. It allows selecting from a wide list of available stimuli in the database those that will constitute a new specific protocol or study, with a given name. By clicking on each auditory stimulus, the researcher may listen to it. For convenience, stimuli may be searched by name. Once a new protocol is created, it is stored in the server's database and becomes available to be used.

Finally, the third tab implements functionalities that allow the researcher to visualize the data. The researcher can visualize data that are being collected in real time or data that were collected in previous studies. Given the extent of the collected datasets, filters are also available to allow a more efficient visualization of the required data. Section 2.2.1 provides a more detailed description of the structure of the generated datasets.

2.2.1. Data collection

The dataset generated during each study describes a group of very different variables, both personal, operational and behavioral.

Personal variables are collected at the beginning of the participation and are used to identify and characterize the participant including, among others, a unique id, gender, date of birth, number of years of education or job. Operational variables describe certain events such as the start of a study or advancing to the next stimulus. Finally, behavioral variables describe the behavior of the user (in terms of Human–Computer Interaction) throughout the study.

Briefly, the dataset is composed of a temporal sequence in which each element describes a specific event or action, at a given time. In this sequence, each element can have one of five types: touch, event, stimulus, emotion or likert scale. Depending on its type, each element holds different data, as follows:

- touch, timestamp, source, X, Y, P, S — it describes the event of a touch on the screen of the mobile device, at a given time. It identifies the source of the touch (e.g. a specific control, the background), its coordinates on the screen and the average values of pressure and area during the touch;
- event, timestamp, description — it describes a specific event in a given time, such as starting the instruction activities or advancing to the study;
- stimulus, timestamp, id — denotes the moment in which the participant in the study proceeds to a new stimulus, i.e., the first moment in which the stimulus is played via the speakers;
- emotion, timestamp, type, #repetitions — it denotes that in a given moment, the participant classified the current stimulus as conveying a particular emotion. It also describes the number of repetitions of the stimulus before registering the response;
- likert, timestamp, type, value, #repetitions — it denotes the classification of the current stimulus in a likert scale of a specific value (i.e. valence, intensity or authenticity) with a certain value. It also describes the number of repetitions of the stimulus before registering the response.

2.3. Mobile application

The mobile application, developed for android, targets devices with large screens in order to facilitate interaction. It was designed to be simple to interact with (of particular relevance for special populations, such as children and older adults), and to minimize influence on results, namely by keeping text and colors to a minimum. Note that the graphical interfaces depicted in this section are in Portuguese since the mobile application was developed for Portuguese participants.

The application receives the protocol via socket in JSON format and dynamically generates all the corresponding graphical interfaces. Two examples are depicted in Figs. 3 and 4.

Fig. 3 depicts the graphical interface used by the participant to classify the emotion expressed while listening to the stimulus. The stimulus can be repeated any number of times by clicking on the sound speaker image. By clicking in one of the emotion categories, the application advances to the following activity.

Fig. 4 depicts the graphical interface in which the participant is able to classify the valence of a given vocalization using a 9-point likert scale. As in Fig. 3, the stimulus can be repeated by clicking on the sound speaker. After adjusting the desired value, the participant advances by clicking on a given button.

As the user interacts with the mobile application, it registers the previously mentioned data and sends them, in real time, to the controller computer. Data are aggregated and sent via socket, in JSON format, at each new stimulus or at the end of the protocol in the case of the last stimulus.

Fig. 3. Graphical interface for the participant to select the emotion portrayed by the vocal stimulus heard: sadness, fear, disgust, anger, surprise, relief, pleasure, amusement, triumph, or a neutral expression (in Portuguese).

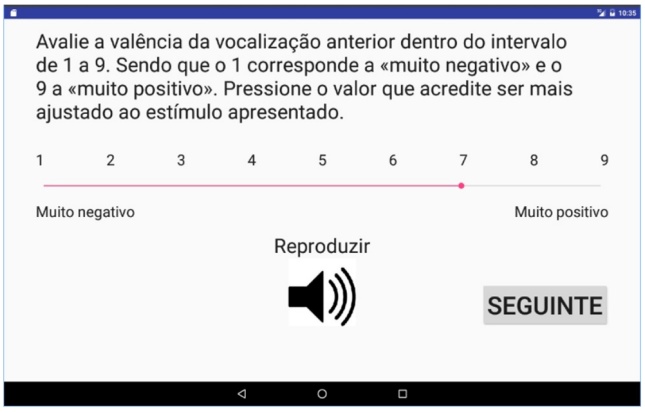


Fig. 4. Graphical interface with a 9-point likert scale for the assessment of stimulus valence by the participant (in Portuguese).

3. Case study

In this section we describe a Case Study in which this system was used in the Faculty of Psychology of the University of Lisbon, Portugal, where the application is now being used. In this case study, the participant sits alone in a room, listening to the stimuli and interacting with the tablet to provide the responses, while the researcher is in an adjacent control room, eventually looking at the participant's responses and behaviors in real time (Fig. 5). The computer in the control room is connected to the sound system in the participant's room and communicates wirelessly with the participant's tablet, so that the auditory stimuli are reproduced when needed and that the data collected is transmitted in real time to the server.

For validation purposes 39 individuals (19 male, 20 female) were selected to participate in a protocol with 50 stimuli, with 5 exemplars of each emotion (sadness, fear, disgust, anger, surprise, relief, pleasure, amusement, triumph, plus a neutral expression). The average age of the participants is 24.69 years ($\sigma = 8.3$). A Samsung Galaxy Tab 3 with a screen of 10.1" was used for user interaction.

The main goal of this Case Study was not to actually study vocal emotion recognition in a given population but rather to validate the developed system by highlighting the new types of collected data as well as the new features they support, when compared to traditional approaches. Nonetheless, interesting results are provided in Section 4.

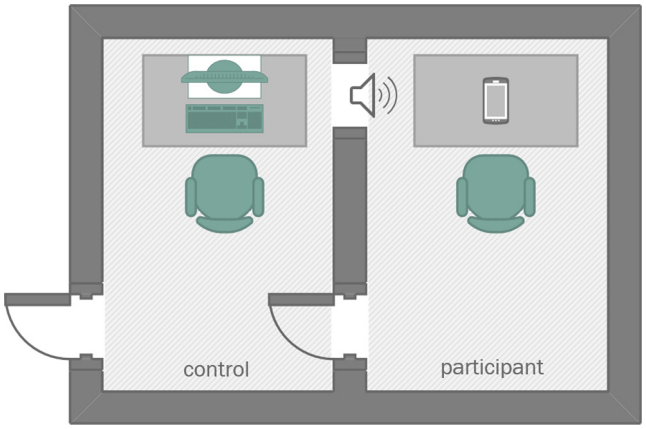


Fig. 5. Layout of the environment: the participant sits alone in a room, listening to vocal sounds through a sound speaker and interacting with the tablet, which is connected to the computer in the control room via wireless.

3.1. Dataset

The data collected in this Case Study resulted in two main MongoDB collections: one describing the participants (e.g. id, age, occupation) and another describing all the events of each individual participation. The first collection has 39 documents, one for each participant. The second collection has 61.933 documents whose schema varies according to the type of event. There are five types of events for which data is generated:

- Touch — describes a specific touch of a user on the screen of the tablet;
- Event — describes events such as a participant entering the welcome screen or finishing the training phase;
- Stimulus — the moment in which a specific auditory stimulus was reproduced;
- Emotion — when a participant classifies a specific stimulus in terms of a discrete emotion;
- Scale — when a participant classifies a stimulus using a likert-scale, in terms of valence, intensity or authenticity.

There are certain attributes that are common to all these types of events:

- _id — identifier of the document;
- userid — identifier of the participant that produced the data described in this document
- date — the date in which this data was produces
- type — the type of the event. One of *Touch*, *Event*, *Stimulus*, *Emotion* or *Scale*;
- is_training — whether the data was generated during the training phase or not;
- current_stimulus — the identifier of the current stimulus (if any) being evaluated by the participant.

Of the 61.933 documents, 43.641 describe touch events (one for each touch of each participant on the screen of the tablet). Each document describing a touch event has the following additional attributes:

- target — identifies the element of the UI in which the touch happened. Used for calculating touch accuracy;
- x,y — the coordinates of the touch on the screen (in pixels);
- intensity — the average intensity of this specific touch;
- area — the average area of this specific touch.

There are also 3.607 documents describing the start of a new stimulus, that is, the moment in which the participant advances to the following stimulus and it is reproduced for the first time in the speakers. Each of these documents has an additional field that identifies the stimulus that was now played.

An equal number of documents exist describing the classification of a given stimulus, by a given participant, at a certain moment. The following attributes are also stored for these documents:

- stimulus — identifier of the stimulus being classified;
- emotion — the emotion classified by the participant after hearing the stimulus;
- repetitions — the number of repetitions of the auditory stimulus that the participant needed to provide this classification;
- correct — whether the classification provided is correct or not. This is determined automatically since the auditory stimuli used are classified.

Another 10.821 documents (three for each stimulus played) describe the classification of each stimulus by the participant, but now in terms of valence, authenticity and intensity. These documents contain the following additional attributes:

- stimulus — identifier of the stimulus being classified;
- scale — the type of scale being classified. One of *Valence*, *Intensity* and *Authenticity*;
- value — the value in a 9-point likert-scale attributed by this participant to this stimulus in this scale.
- repetitions — the number of repetitions of the auditory stimulus that the participant needed to provide this classification;

The remaining 257 documents describe events in the mobile application such as entering the welcome screen, starting the instructions or starting/ending the training phase. There is a single additional attribute used in these documents, which contains the identifier of the event so that the actions of the participant during the study can be reconstructed.

3.2. Data processing

Whenever new data is inserted into the database, some processing tasks are carried out. First, data are filtered so that events and touches that took place during the training phase for each user are discarded. Next, some aggregation operations are performed and its results stored in other collections. This is done so that data analysis and visualization under different perspectives become easier. Moreover, this also prevents these otherwise frequent aggregation operations from being run whenever this type of data are needed. The following aggregations are constructed, which allow views and analyses such as those depicted in Section 4:

- Touch events grouped by participant. This allows, for example, to visualize or analyze changes in touch intensity or touch duration throughout the study for each individual participant;
- Touch events grouped by participant at 5 min intervals. While the previous feature contains values of all the touches, this contains the results averaged at each 5 min interval so that the natural variation of this type of data is smoothed. This allows for more intuitive visualizations of how data is evolving throughout time;
- Touch events grouped by emotion. This allows to study, for instance, how touch intensity varies according to the emotion conveyed by each of the stimulus, regardless of the participant;
- Touch events grouped by emotion and participant. This feature is more specific than the previous one in that it allows to study how the interaction patterns of each individual may be affected by each of the emotions conveyed by the stimuli.

Data generated in these aggregations are also labeled according to the gender of the participant and according to their age group (younger or older than 45). This allows to analyze data based on gender or age group to determine if there are differences in interaction or in emotion recognition capacities in these groups of participants.

4. Results

This section analyzes the collected data and presents some results. The two first subsections are dedicated to results concerning Operational and Behavioral Features, respectively. The third and last subsection shows how a Neural Network can be trained with this type of behavioral data to distinguish between male and female individuals with an interesting success rate.

4.1. Operational features

In this section we include features that are extracted from the actions of the participant during the protocol, and their timing.

One of the potentially interesting features is the number of repetitions of each stimulus. Indeed, each participant may decide to repeat the current stimulus any number of times they want. An increased number of repetitions may indicate a larger difficulty in perceiving the conveyed emotion, hence the interest of this feature.

First, our analysis focused on the number of repetitions of the vocal stimuli as a function of gender. That is, do women or men need less repetitions of each auditory stimulus to perceive the emotion conveyed and complete the task?

According to the collected data, each stimulus was played 1.96 times ($\sigma = 1.64$) for each male participant. For female participants this value decreased to 1.63 ($\sigma = 1.29$). Differences were statistically significant (Kolmogorov–Smirnov test, p-value $= 1.87^{-5}$). This may indicate that women are more efficient at perceiving emotions and thus need to repeat the stimulus less times to do so.

A similar analysis was carried out based on the age group, considering two groups: the group of participants younger than 45 and the group of participants who were 45 or older at the date of their participation. On average, young participants played each stimulus 1.70 times ($\sigma = 1.36$), whereas older participants played each one 3.22 times ($\sigma = 2.31$). This may indicate an increased difficulty of older people in perceiving and classifying vocal stimuli. Differences were statistically significant (Kolmogorov–Smirnov test, p-value $= 4.097^{-14}$).

We also analyzed the number of repetitions as a function of emotion category. That is, are there certain emotions that are more difficult to be recognized? Similarly, we examined whether gender and age influenced emotion recognition.

Concerning gender (Fig. 6), we observed that female participants repeated stimuli fewer times, as discussed above. It also results clear from this figure that *Disgust* is the emotion that is apparently easier to classify for both male and female participants, as the number of repetitions for this emotion is close to 1, as opposed to all the other emotions that present values higher than 2 for both genders. No major gender-based differences were observed when comparing different emotions.

When performing an age-based analysis, we observed that, as previously mentioned, older participants repeated stimuli more frequently than younger ones. The difference was larger when classifying *Neutral* vocalizations that were, according to this data, the type of vocal stimuli that was associated with decreased recognition accuracy. *Amusement*, on the other hand, was the emotion

Table 1
Average number of stimuli played by emotion, gender and age.

	Male		Female		Young		Older	
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
Achievement	2.51	1.57	1.84	0.58	2.04	1.00	4.10	2.97
Anger	2.60	2.24	1.90	0.76	2.11	1.47	4.18	2.57
Disgust	1.39	0.60	1.08	0.17	1.20	0.46	1.62	0.42
Amusement	2.00	1.51	1.61	0.45	1.63	0.83	4.17	2.31
Fear	2.72	1.77	1.95	0.88	2.14	1.24	5.07	2.29
Neutral	2.76	2.38	2.15	1.31	2.19	1.55	6.55	0.31
Pleasure	2.12	1.28	1.72	0.91	1.87	1.14	2.60	0.50
Relief	2.19	1.28	1.51	0.86	1.72	1.02	3.80	1.31
Sadness	2.35	2.43	1.75	0.89	1.96	1.84	3.13	1.41
Surprise	2.63	2.02	1.91	1.43	2.11	1.61	4.60	3.39
$\bar{x}$	2.33	1.71	1.74	0.82	1.90	1.22	3.98	1.75

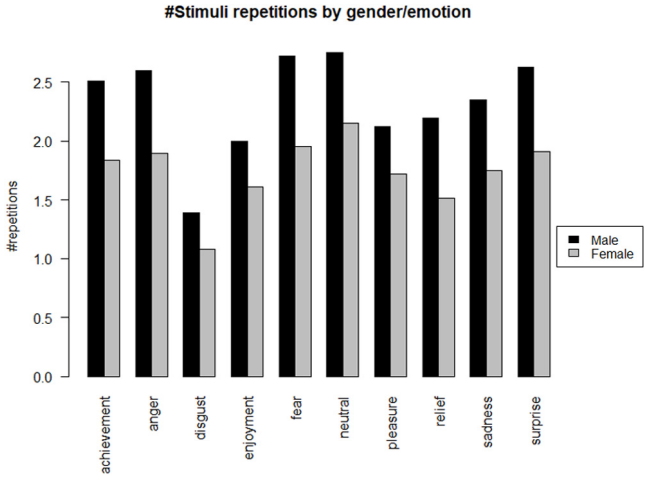


Fig. 6. Number of repetitions of stimuli conveying each of the emotions, grouped by gender.

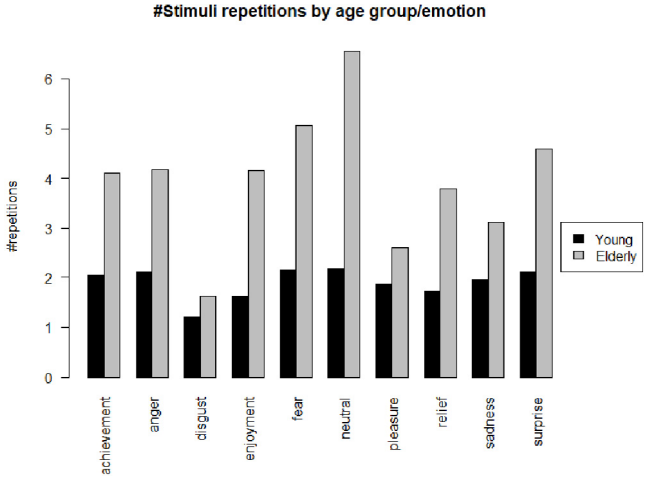


Fig. 7. Number of repetitions of stimuli conveying each of the emotions, grouped by age.

that was easier to distinguish, independently of the age group. Indeed, this emotion was associated with less accuracy differences between younger and older participants. Table 1 summarizes this data. (See Fig. 7.)

The collected data also allows the analysis of the percentage of recognition accuracy for each emotion type to examine which users or group of users are better at decoding emotions in vocal stimuli. While an individual analysis of each participant could be carried out, we focused on the analysis of data by age group and gender (Fig. 8).

In an analysis by gender, it was possible to conclude that male participants correctly classified, on average, 68% of the stimuli. Female participants, on the other hand, correctly classified 71% of the vocal stimuli. When considering age, younger participants correctly classified 69% of the stimuli while older participants correctly classified 68%. Although the observed differences are in line with current evidence, they were not statistically significant.

The aggregation of data created by the proposed system also allows an analysis of correctly classified emotions by type of emotion and gender/group age. For example, this allows testing if there are emotions that are generally easier to identify for a certain individual or group of individuals. Fig. 9 shows how the percentage of correctly classified stimuli varies according to gender and type of emotion. For nearly all emotion types, women performed better (as already detailed) than men, with the curious exception of *Neutral* vocalizations, which men seem to detect more accurately. *Fear* represented the more difficult vocal emotion to categorize for both men and women, as opposed to *Amusement* and *Disgust*, for which both men and women achieved nearly perfect scores. The results obtained for the two age groups are similar and are, therefore, not detailed here.

Another interesting feature that can be extracted from the collected data is the time between decisions, which quantifies the time between each decision a participant takes (e.g. repeating a stimulus, advancing to the next stimulus). According to the collected data, male participants took decisions (on average) each 2.7 s, whereas female participants do so at each 3.06 s. When considering the age of the participants, young listeners took on average 2.74 s between decisions, whereas older participants took 4.24 s. A longer time to take decisions may indicate either a more weighted decision making (as probably happens with female participants) or slower cognitive processing (as is probably the case of older participants).

The differences observed when comparing the male/female and young/older groups were statistically significant in both cases (Kolmogorov–Smirnov test, p-value $< 2.2^{-16}$). (See Fig. 10.)

Other so-called operational features could be extracted, or more detailed versions of the described ones, such as individual analysis by participant that would point out the accuracy of each participant when classifying each type of emotion.

4.2. Behavioral features

This section describes results concerning behavioral features, i.e., features that are extracted from the participant's interaction with the tablet. This group of features represents the most innovative aspect of this work as they are used, to the extent of our knowledge, for the first time in this context. Here, we focused mostly on touch intensity and touch area, as well as on how touch

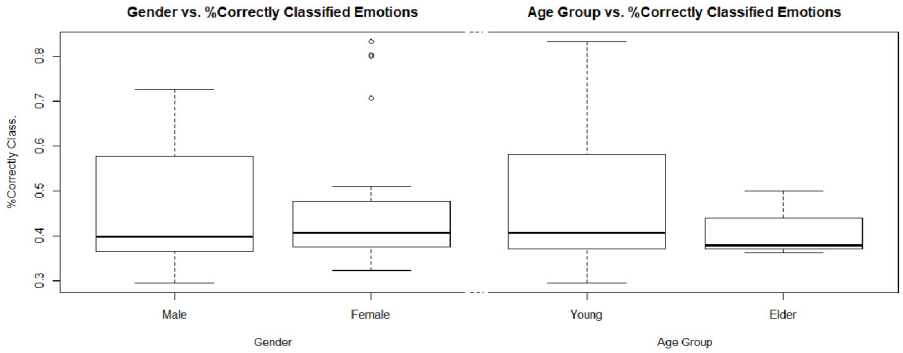


Fig. 8. Percentage of correctly classified stimuli by gender and by age group.

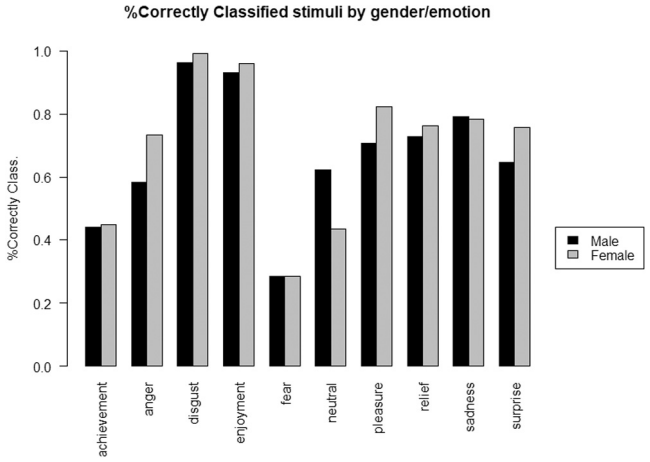


Fig. 9. Percentage of correctly classified stimuli by gender and emotion.

Table 2
General statistics of touch intensity and area for both the different genders and age groups.

	Area		Intensity	
	$\bar{x}$	σ	$\bar{x}$	σ
Male	0.01	0.005	0.11	0.03
Female	0.009	0.005	0.09	0.03
Young	0.01	0.004	0.11	0.03
Elder	0.009	0.005	0.10	0.03

Table 3
Emotions for which interaction patterns were statistically different.

	Ach	Ang	Dis	Enj	Fea	Neu	Ple	Rel	Sad	Sur
Sur			X		X			X	X	
Sad	X	X	X			X	X			
Rel	X		X			X				
Ple	X		X		X					
Neu		X	X		X					
Fea			X							
Enj										
Dis	X	X								
Ang	X									
Ach										

varies according to gender, age group or emotions felt by the participants.

Concerning the influence of age on the interaction patterns, in general it is possible to conclude that older participants tended to press the screen with more intensity and with a larger area of the finger. This is illustrated in Fig. 11 and further detailed in Table 2. The observed differences were statistically significant for both intensity and area features (Kolmogorov–Smirnov test, p-value $< 2.2^{-16}$ for both features).

In what concerns the role of gender in the behavioral features under analysis, the main observation is that men tended to use more intense touches on the screen and also with more area of the finger, as detailed in Fig. 12 and in Table 2. As with the age-based analysis, differences were statistically significant for both intensity and area features (Kolmogorov–Smirnov test, p-value $< 2.2^{-16}$ for both features).

There was a clear and statistically significant difference in terms of interaction with the tablet that can be associated to gender and to age group. This is further explored in Section 4.3, namely by training a Neural Network that is able to distinguish between male and female subjects with an interestingly high degree of accuracy.

We were also interested in the relationship between emotions and Behavioral Biometrics. Specifically, we wanted to determine whether participants touched the tablet differently in response to the different emotions they listened to. Fig. 13 shows the distribution of the intensities of the touches for all participants according to the emotion of the stimulus being classified at the moment of the touch. As in other results addressed before in this document, *Disgust* stands out once again, but now as the emotion that is associated with a generally higher touch intensity. *Surprise*, *Neutral* and *Amusement*, on the other hand, were the emotions for which touch intensity was generally lower. Indeed, the distributions of touch intensities were modulated by the type of vocal emotion.

Table 3 highlights the emotions for which touch intensity was significantly different (i.e. p-value < 0.05). Essentially, this table shows, in a simplistic manner, how interactions differ as a function of emotion. It is interesting to note, for example, that interactions associated to the emotion *Disgust* were those that were more easily distinguished from the remaining ones since its distribution was statistically different from every other emotion except *Amusement*. This confirms, once again, that this emotion significantly alters the participants' interaction with the tablet in a very specific manner.

It would also be interesting to study the relationship between emotion and Behavioral Biometrics for each individual. Indeed, each of us is affected differently by the distinct emotions (and at different levels) as we have different levels of empathy. The collected data can be explored to show such inter-individual differences. As an example, Fig. 14 shows two participants whose interaction patterns with the tablet were significantly different in response to the emotions expressed by the speakers: the participant depicted on the right shows very different distributions of touch intensity for different emotions, as opposed to the participant depicted on the left whose interaction does not seem to be that affected by emotion.

Although it was not explored in the current paper, the relationship between emotional empathy and Behavioral Biometrics

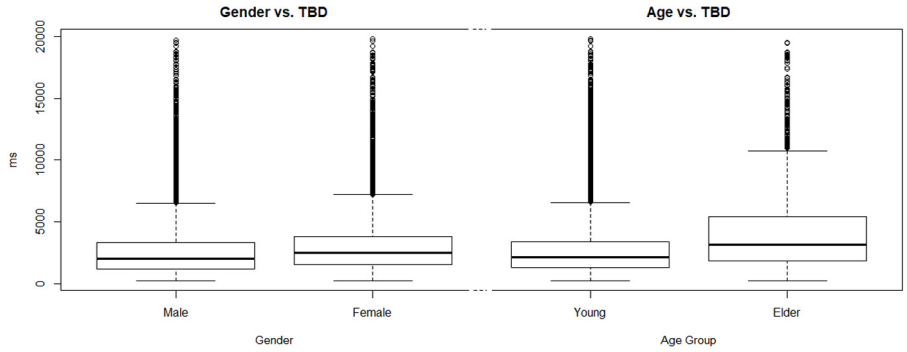


Fig. 10. Time between decisions according to gender and age group.

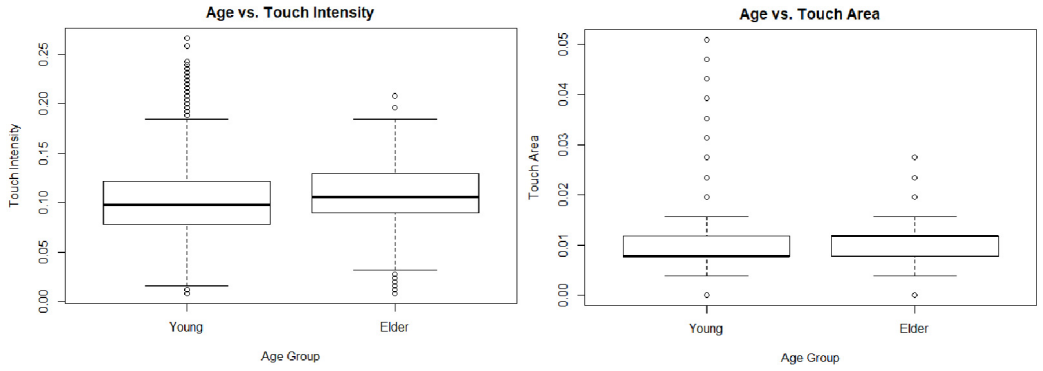


Fig. 11. The influence of age on two Behavioral features: touch intensity and touch area.

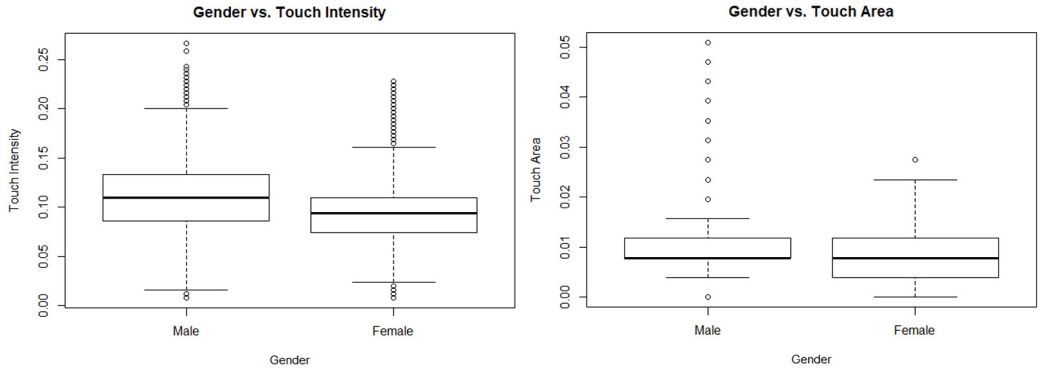


Fig. 12. The influence of gender on two Behavioral features: touch intensity and touch area.

deserves further attention as it may contribute significantly to the development of emotion-aware applications [24].

Other types of analyses are supported by this approach, although we do not explore them any further in this paper. For example, it is possible to analyze how touch intensity varies for each participant, over time. This may allow determining whether there are events or moments that significantly influence the participant's interaction with the device. For example, if stimuli were presented to the users grouped by emotion, this type of analysis could allow to visually assess the differences in behavioral biometrics in each block of stimuli. Moreover, it also allows comparing how different participants (or groups of participants) behave during the study. For example, Fig. 15 compares two participants. During the first 15 min their behavior is similar, significantly changing afterwards. Something caused the male participant's touch intensity to rise significantly after this moment and the female to drop. This type of analyses provide new features that may reveal interesting insights about the individual participating in auditory emotional recognition studies.

4.3. The influence of gender on behavioral biometrics

From the results depicted in Section 4.2, it results clear that there are significant differences in both behavioral and operational features when comparing different groups of the population, namely age-based and gender-based differences. In this section, we explore the relationship between gender and three of the variables for which these differences are more striking: touch intensity, touch area and time between decisions. The goal is to examine if these differences are enough to determine, for example, the gender of the participant. To this end we trained a Neural Network to classify gender from Behavioral Biometrics.

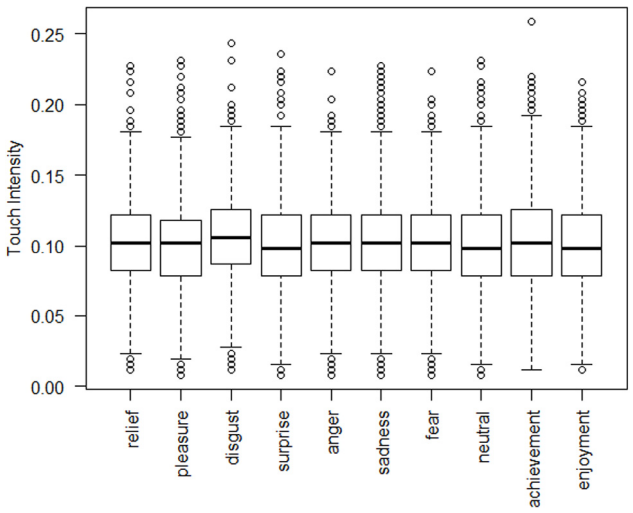


Fig. 13. Distribution of touch intensities while the participants were evaluating vocal stimuli conveying different emotions.

These variables have, however, some inherent variability. For example, the intensity of one's touch on a screen varies significantly from one touch to the next as it is a too fine physical movement to be consciously controlled by the individual. A single touch cannot thus be used to accurately represent or characterize its user's interaction patterns.

For this reason, in this section we aggregated and averaged the data, for each participant, at 5-minute intervals. In this manner, each interval indicates a better representation of the participant's interaction patterns as the average eliminates the variability of each individual touch and shows, in turn, the general behavior of the participant. Since Neural Networks only work with numeric variables, in this dataset the feminine gender is represented by the value 0 whereas the masculine is represented by the value 1.

The dataset used in this Section is thus composed of a total of 250 instances (129 of female participants, 121 of male participants), depicting nearly 21 h of interaction of the 39 different participants. Each instance contains the average, standard deviation and variance of three features (thus 9 variables): touch intensity, touch area and time between decisions. However, in the network, only the average values are used as, after several attempts, these were the ones that provided better results.

Thus, a feed forward Neural Network was trained with three inputs (average touch intensity, average touch area and average time between decisions), two hidden layers and one output: gender (Fig. 16). To evaluate the accuracy of the predictor, the output of the Neural Network is rounded to either 0 or 1 and the resulting vector is compared to the vector of the test set. Fig. 16 shows the results of the predictor before rounding the output: the test set (x-axis) contains only values 0 (female) or 1 (male), while the

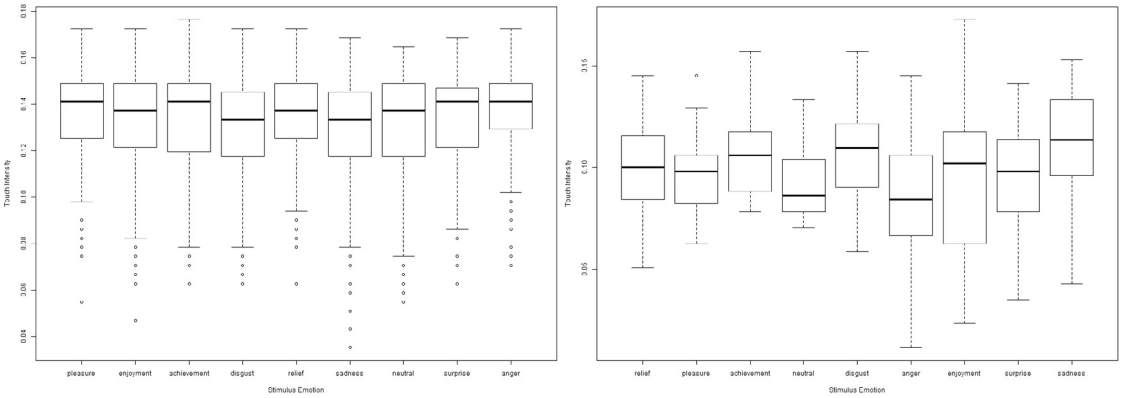


Fig. 14. Inter-individual differences in the influence of emotions on touch intensity for two participants.

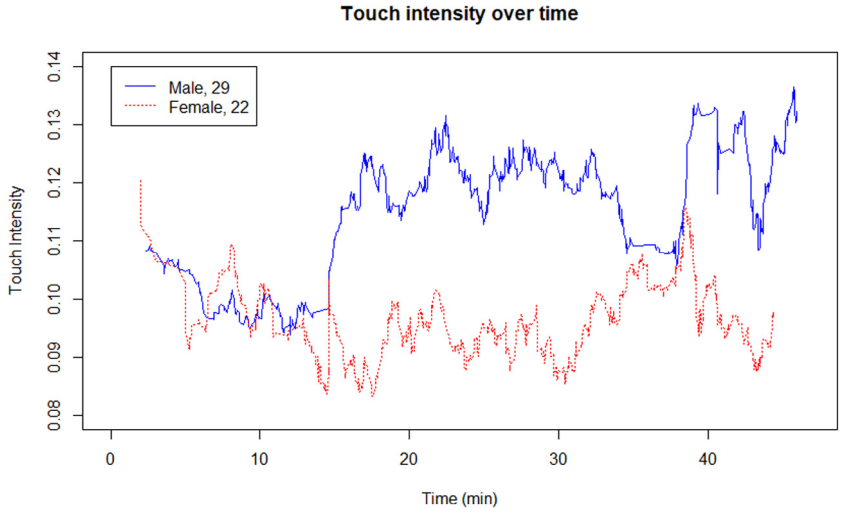


Fig. 15. Evolution of touch intensity over time for two participants.

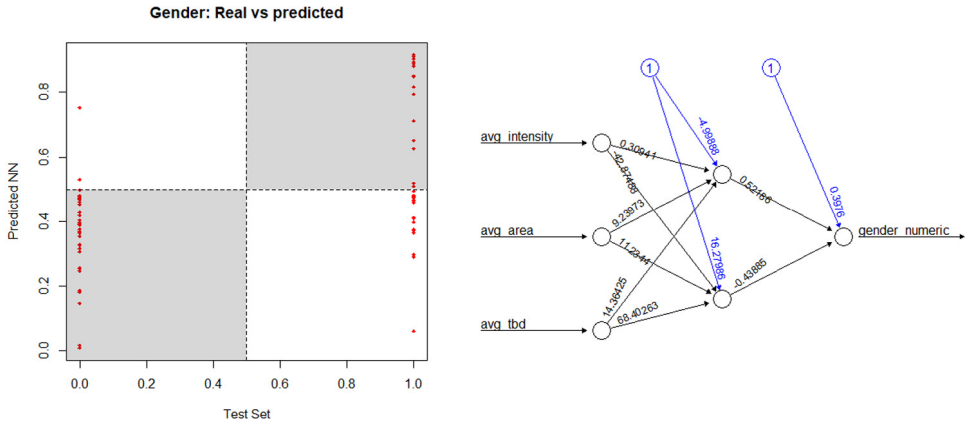


Fig. 16. Neural Network used to predict gender (right) and classification results for the test set (left). Gray areas mark the correctly classified instances (72%).

prediction ranges between 0 and 1 (y-axis). The gray areas depict the correctly classified instances, which amount to 72%.

5. Conclusion and future work

In the current paper an innovative approach was presented to improve auditory emotion recognition studies. The system, composed of a mobile application and a server application, was developed and validated in cooperation with the Faculty of Psychology of the University of Lisbon, where it will be used to assess vocal emotion recognition both in research and clinical settings.

In this paper the developed system was described, with a special emphasis on the functionalities implemented. Moreover, the main innovative aspects of the work were described, accompanied with examples of the information that can be extracted from each study.

In terms of Human–Computer Interaction, it is now clear that user profile (e.g. gender, age) and user state (e.g. emotional state) influence interaction patterns with technological devices, in the same way that they influence the individual's interaction with others. Moreover, we show that this relationship appears to be consistent within the user profile. That is, interaction patterns seem to depend largely on aspects such as age and/or gender.

Specifically, we show that interaction is different for individuals in different age groups. We compare young individuals with elderly ones to conclude that older people touch the screen with increased intensity and area. Interaction is also significantly different according to gender. Collected data showed that male participants touch the screen with more intensify and finger area. Based on these differences, we trained a neural network that is able to classify gender based solely on the observation of the participant's interaction with the device, with an accuracy of 72%.

In terms of auditory emotion recognition and Human–Computer Interaction, conclusions are two-fold. In terms of emotion recognition, data shows which emotions are easier and harder to recognize by the participants. While this does not represent a novel contribution, results are in line with existing knowledge. Namely, we conclude that women are better at perceiving emotion when compared with men, and that younger participants are also better than older ones.

When analyzing the relationship between emotion and Human–Computer Interaction, some interesting conclusions where put forward. Namely, *Disgust*, which is the emotion that is more easily identified by all participants, is also the emotion that more significantly affects touch intensity and area, with both variables being higher than with other emotions. Moreover, data also shows that different participants are affected differently by emotions: some are more susceptible to the influence of different emotions on their interaction patterns. While considerable future work is still necessary, these findings show that it may be feasible to develop emotion-aware devices and applications.

Finally, when comparing the developed system with existing approaches, the following key innovative aspects can be identified:

- Studies are easy to design and share among researchers. The mobile application generates all the necessary graphical interfaces according to the study design, in a transparent way for the researcher;
- Data collection and storage in a structured manner is automatized and requires no Human intervention, improving the efficiency of the process and its validity by eliminating potential Human error;
- Many new variables are now considered that may provide important information about participants' behavior. This may be very important to clarify how each participant is affected by different types of emotions;
- Several studies can be conducted simultaneously as now there is not a dependence on the researcher to play the stimuli and record the participants' responses;
- Data are readily available during and immediately after the collection, facilitating and accelerating its analysis.

We believe that the advantages of the developed system go beyond the enrichment of auditory emotion recognition studies: we are convinced that this approach may lead to the collection of relevant data to create specific user profiles. Indeed, the Faculty of Psychology frequently deals with populations with specific characteristics, including age, gender, intellectual disabilities, psychiatric diagnosis, among others.

The collection of these data, properly contextualized with the participants' features, may allow a deeper characterization of user interaction profiles that can later be used to develop software and hardware sensitive to human emotions or to human characteristics.

In that sense, future work will be conducted with the goal to collect larger sets of data and from different populations, and to improve the server to build additional features that may reveal new insights and improve this type of studies. Namely, we will next focus on studying the influence of age on Behavioral Biometrics with the goal to develop an age predictor for mobile applications based solely on the user's interaction. We will also further the examination of the relationship between Behavioral Biometrics and individual empathy, in an attempt to contribute to the development of emotion aware devices and applications.

Acknowledgments

This work has been supported by FCT — Fundação para a Ciência e Tecnologia, Portugal (PTDC/MHN-PCN/3606/2012) and by COMPETE, Portugal: POCI-01-0145-FEDER-007043 and FCT, Portugal within the Project Scope: UID/CEC/00319/2013.

References

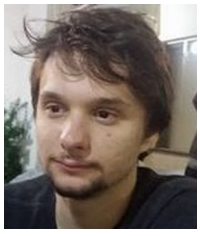
[1] M.L. Knapp, J.A. Hall, T.G. Horgan, Nonverbal Communication in Human Interaction, Cengage Learning, 2013.
[2] P.N. Juslin, P. Laukka, Communication of emotions in vocal expression and music performance: different channels, same code?, Psychol. Bull. 129 (5) (2003) 770.
[3] A. Schmidt, Context-aware computing: context-awareness, context-aware user interfaces, and implicit interaction, in: The Encyclopedia of Human-Computer Interaction, second ed., 2013.
[4] P. Novais, R. Costa, D. Carneiro, J. Neves, Inter-organization cooperation for ambient assisted living, J. Ambient Intell. Smart Environ. 2 (2) (2010) 179–195.
[5] M. Sultana, P.P. Paul, M. Gavrilova, A concept of social behavioral biometrics: motivation, current developments, and future trends, in: Cyberworlds (CW), 2014 International Conference on, IEEE, 2014, pp. 271–278.
[6] A.P. Pinheiro, C. Barros, J. Pedrosa, Salience in a social landscape: electrophysiological effects of task-irrelevant and infrequent vocal change, Soc. Cogn. Affect. Neurosci. 11 (1) (2015) 127–139.
[7] A.P. Pinheiro, C. Barros, M. Dias, S.A. Kotz, Laughter catches attention!, Biol. Psychol. 130 (2017) 11–21.
[8] T. Brosch, D. Grandjean, D. Sander, K.R. Scherer, Cross-modal emotional attention: emotional voices modulate early stages of visual processing, J. Cogn. Neurosci. 21 (9) (2009) 1670–1679.
[9] P. Belin, S. Fillion-Bilodeau, F. Gosselin, The montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing, Behav. Res. Methods 40 (2) (2008) 531–539.
[10] M. Vasconcelos, M. Dias, A.P. Soares, A.P. Pinheiro, What is the melody of that voice? probing unbiased recognition accuracy with the montreal affective voices, J. Nonverbal Behav. (2017) 1–29.
[11] T. Liu, A.P. Pinheiro, G. Deng, P.G. Nestor, R.W. McCarley, M.A. Niznikiewicz, Electrophysiological insights into processing nonverbal emotional vocalizations, NeuroReport 23 (2) (2012) 108–112.
[12] F.R. Balte, A.C. Miu, Emotions during live music performance: links with individual differences in empathy, visual imagery, and mood, Psychomusicol.: Music, Mind, and Brain 24 (1) (2014) 58.
[13] C.F. Lima, S.L. Castro, S.K. Scott, When voices get emotional: a corpus of nonverbal vocalizations for research on emotion processing, Behav. Res. Methods 45 (4) (2013) 1234–1245.
[14] K.R. Scherer, J. Sundberg, L. Tamarit, G.L. Salomão, Comparing the acoustic expression of emotion in the speaking and the singing voice, Comput. Speech Lang. 29 (1) (2015) 218–235.
[15] T.M. Bynion, M.T. Feldner, Self-assessment manikin, in: Encyclopedia of Personality and Individual Differences, Springer, 2017, pp. 1–3.
[16] S.K. Card, The Psychology of Human-Computer Interaction, CRC Press, 2017.
[17] T. Page, Touchscreen mobile devices and older adults: a usability study, Int. J. Hum. Factors Ergon. 3 (1) (2014) 65–85.
[18] J.M. Díaz-Bossini, L. Moreno, Accessibility to mobile interfaces for older people, Procedia Comput. Sci. 27 (2014) 57–66.
[19] N. Jochems, Designing tablet computers for the elderly a user-centered design approach, in: International Conference on Human Aspects of IT for the Aged Population, Springer, 2016, pp. 42–51.
[20] R. Sharma, F.F.H. Nah, K. Sharma, T.S.S.S. Katta, N. Pang, A. Yong, Smart living for elderly: design and human-computer interaction considerations, in: International Conference on Human Aspects of IT for the Aged Population, Springer, 2016, pp. 112–122.
[21] L. Findlater, J.E. Froehlich, K. Fattal, J.O. Wobbrock, T. Dastyar, Age-related differences in performance with touchscreens compared to traditional mouse input, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2013, pp. 343–346.
[22] D. Carneiro, P. Novais, M. Gomes, P.M. Oliveira, J. Neves, A statistical classifier for assessing the level of stress from the analysis of interaction patterns in a touch screen, in: Soft Computing Models in Industrial and Environmental Applications, Springer, 2013, pp. 257–266.
[23] D. Carneiro, A.P. Pinheiro, P. Novais, Context acquisition in auditory emotional recognition studies, J. Ambient Intell. Humaniz. Comput. 8 (2) (2017) 191–203.
[24] R.J.R. Blair, Responding to the emotions of others: dissociating forms of empathy through the study of typical and psychiatric populations, Conscious. Cognit. 14 (4) (2005) 698–718.

Davide Carneiro is an Invited Professor at the Polytechnic Institute of Porto. He is also a researcher at the CI-ICESI centre, of the Polytechnic Institute of Porto, and of the Algoritmi centre, in the Department of Informatics, University of Minho, Braga, Portugal. He holds a Ph.D. from a joint Doctoral Programme in Computer Science of three top Portuguese Universities. He develops scientific research in the fields of Human–Computer Interaction and Context-aware Computing. His main interest lies in acquiring information in a non-intrusive way, from the human's interaction with the computer, namely to assess stress, mental fatigue and emotions. He has participated in several research projects in the fields of Artificial Intelligence, Ambient Intelligence and Online Dispute Resolution. He is the author of several publications in his field of interest, including one authored book, one edited book and over eighty book chapters, journal papers and conference and workshop papers. In 2008 he was awarded the TLeIA08 — a National Award for Artificial Intelligence projects attributed by the Portuguese Artificial Intelligence Association and in 2009 he has been awarded an Academic Merit Scholarship by the Portuguese Government.

Ana P. Pinheiro is Assistant Professor at the Faculty of Psychology — University of Lisbon, where she leads the Voice, Affect, & Speech (VAS) Lab. She holds a Ph.D. in Psychology from the University of Minho, Portugal. Her research interests involve the investigation of brain mechanisms underlying cognitive processes, with the focus on language, speech and social communication, and their interactions with emotion. Her aim is to characterize these processes in a comprehensive manner by gathering evidence from behavioral tasks (understand behavior) and relating them to brain function (e.g., understand underlying neurophysiology). With the goal of making the findings of her research relevant to society, she has been actively involved in research projects that aim to bridge Computer Science and Psychology.

Marta Pereira is currently attending the fourth year of the Master's Degree in Psychology at the Faculty of Psychology of the University of Lisbon (FPUL). Her Master is focused on Applied Social Cognition and she is very interested in Cognitive and Affective Neuroscience.

Inês Ferreira graduated in Psychology by the Faculty of Psychology of the University of Lisbon, and is currently working in her Masters in Clinical Psychology of Health and Disease. Her major academic interests include Neuropsychology, as well as Cognitive and Affective Neuroscience.



Miguel Domingues is a student doing his master's degree on Psychology, in the department of Social Cognition of the Faculty of Psychology of the University of Lisbon. He finished his Bachelor's degree on Psychology in 2017, in the same University. His research interests lie on the fields of cognitive and affective neurosciences, being especially interested in voice processing, the perception of speech and language and visual recognition of letters and words.

Paulo Novais is an Associate Professor with Habilitation of Computer Science at the Department of Informatics, in the School of Engineering of the University of Minho (Portugal) and a researcher at the ALGORITMI Centre in which he is the coordinator of the research group Intelligent Systems Lab, and the coordinator of the research line in "Ambient intelligence for well-being and Health Applications". From the same university he received a Ph.D. in Computer Science in 2003 and his Habilitation in Computer Science in 2011. He started his career developing scientific research in the field of Intelligent Systems/Artificial Intelligence (AI), namely in Knowledge Representation and Reasoning, Machine Learning and Multi-Agent Systems. His interest, in the last years, was absorbed by the different, yet closely related, concepts of Ambient Intelligence, Ambient Assisted Living, Intelligent Environments, AI and Law, Conflict Resolution and the incorporation of AI methods and techniques in these fields. His main research aim is to make systems a little more smart, intelligent and also reliable. He has led and participated in several research projects sponsored by Portuguese and European public and private Institutions and has supervised several Ph.D. and M.Sc. students. He is the co-author of over 230 book chapters, journal papers, conference and workshop papers and books. He is the president of APPIA (the Portuguese Association for Artificial Intelligence) for 2016/2017 and member of the executive committee of the IBERAMIA (IberoAmerican Society of Artificial Intelligence). During the last years he has served as an expert/reviewer of several institutions such as EU Commission and FCT (Portuguese agency that supports science, technology and innovation).