Journal of Ambient Intelligence and Humanized Computing (2020) 11:6125–6140
https://doi.org/10.1007/s12652-020-01983-3

ORIGINAL RESEARCH

Analysis of the interaction between elderly people and a simulated virtual coach

Raquel Justo[1] · Leila Ben Letaifa[1] · Cristina Palmero[2] · Eduardo Gonzalez-Fraile[3] · Anna Torp Johansen[4] · Alain Vázquez[1] · Gennaro Cordasco[5] · Stephan Schlögl[6] · Begoña Fernández-Ruanova[3] · Micaela Silva[4] · Sergio Escalera[2] · Mikel deVelasco[1] · Joffre Tenorio-Laranga[3] · Anna Esposito[5] · Maria Korsnes[4] · M. Inés Torres[1]

Received: 14 October 2019 / Accepted: 17 April 2020 / Published online: 22 May 2020
© The Author(s) 2020

Abstract
The EMPATHIC project develops and validates new interaction paradigms for personalized virtual coaches (VC) to promote healthy and independent aging. To this end, the work presented in this paper is aimed to analyze the interaction between the EMPATHIC-VC and the users. One of the goals of the project is to ensure an end-user driven design, involving senior users from the beginning and during each phase of the project. Thus, the paper focuses on some sessions where the seniors carried out interactions with a Wizard of Oz driven, simulated system. A coaching strategy based on the GROW model was used throughout these sessions so as to guide interactions and engage the elderly with the goals of the project. In this interaction framework, both the human and the system behavior were analyzed. The way the wizard implements the GROW coaching strategy is a key aspect of the system behavior during the interaction. The language used by the virtual agent as well as his or her physical aspect are also important cues that were analyzed. Regarding the user behavior, the vocal communication provides information about the speaker's emotional status, that is closely related to human behavior and which can be extracted from the speech and language analysis. In the same way, the analysis of the facial expression, gazes and gestures can provide information on the non verbal human communication even when the user is not talking. In addition, in order to engage senior users, their preferences and likes had to be considered. To this end, the effect of the VC on the users was gathered by means of direct questionnaires. These analyses have shown a positive and calm behavior of users when interacting with the simulated virtual coach as well as some difficulties of the system to develop the proposed coaching strategy.

Keywords Human behavior analysis · Human–machine interaction · Spanish · Emotional analysis from speech · Language and face

1 Introduction

Despite advances in health care and technology, most of the eldercare is still provided by informal caregivers, i.e. friends and family members. According to predictions, however, this type of care will decrease in the future, for which studies encourage society to concentrate on improving the lifestyle of the elderly, helping them to remain independent for a longer period of time (Willcox et al. 2014). In particular, socio-behavioral and environmental conditions are seen a crucial factor affecting longevity (Kirkwood 2005), which to some extent explains variations found in the aging process, ranging from active and positive to feeble and dependent. We believe that four principles promote active aging, namely dignity, autonomy, participation, and joint responsibility. Information and Communication Technologies (ICT)

✉ M. Inés Torres
manes.torres@ehu.eus

[1] Universidad del País Vasco UPV/EHU, Bilbao, Spain

[2] Universitat de Barcelona and Computer Vision Center, Barcelona, Spain

[3] Osatek/Osakidetza, Bilbao, Spain

[4] Department of Old Age Psychiatry, Oslo University Hospital, Oslo, Norway

[5] Università degli Studi della Campania Luigi Vinvitelli, Caserta, Italy

[6] MCI Management Center Innsbruck, Innsbruck, Austria

are expected to make such principles possible, allowing the elderly to stay active members of the societal community while helping them remain independent and self-sufficient (Brinkschulte et al. 2018).

Consequently, the EMPATHIC (Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly) project (Montenegro et al. 2019; Torres et al. 2019a, b) aims to contribute to technological progress in this area by researching, innovating and validating new interaction paradigms and platforms for future generations of personalized virtual coaches (VC) to promote healthy and independent aging. The project is centered around the development of the EMPATHIC-VC, a non-obtrusive, emotionally-expressive virtual coach whose aim is to engage senior users in enjoying a healthier lifestyle concerning diet, physical activity, and social interactions. This way, they actively minimize their risk of potentially chronic diseases, which contributes to their ability to maintain a pleasant and autonomous life, while in turn helping their carers.

In this framework this paper aims to analyze the interaction between the EMPATHIC-VC and the users, mainly focusing on the analysis of the human behavior. Actually, the increasing pervasiveness of the computers in our society requires empirical studies of the human behavior during human–machine interaction that provides guidelines for the design of such interactive machines (Justo et al. 2008; Pedersen et al. 2018). As a consequence, one of the goals of the EMPATHIC project is to ensure an end-user driven design, involving senior users from the beginning and during each phase of the project, by considering their needs, by gathering initial data from them as well as their opinions regarding the technology to be developed, and by allowing them to use the personalized prototype from its first version to the final proof of concept. In order to keep them in the loop, senior users are planned to be involved in several sets of test sessions, the first set being the focus of this paper. In these sessions the seniors carried out an interaction with a Wizard of Oz (WOZ) driven, simulated system (Dahlbäck et al. 1993). That is, they believed they were interacting with an autonomous machine while actually the system was operated by an unseen human being. A coaching strategy based on the GROW model (Whitmore 2010) was used throughout these sessions, so as to guide interactions and engage the elderly with the goals of the EMPATHIC project. This way, the senior users were, on the one hand, given the chance to interact with what they thought was a final system (although the system was still not built) and, on the other hand, able to provide very valuable information as to its potential future developments. In addition, it allowed for the collection of an audiovisual data corpus which is currently used to train the machine learning models underpinning the different modules of the entire EMPATHIC system. Another important aspect to be considered for the design of the virtual coach is its visual aspect, which will have a direct impact on the user reaction. Thus, this paper also reports some studies aimed to design for elders' virtual agent acceptance.

Human behavior during the interaction with technical systems strongly depends on the goals and tasks to be developed by the interacting devices as well as on their ability for adaptation to individual user profiles and skills, preferences and emotional states (Irastorza and Inés Torres 2019; Siegert et al. 2013). Working on the aforementioned interaction framework, both the human and the system behavior were analyzed. The way the wizard implements the GROW coaching strategy is one of the key aspects to be analyzed to characterize the system behavior during the interaction with the seniors. The language used by the virtual agent, which is proposed by the natural language generator, as well as his or her physical aspect are also important cues that define the system behavior and that will be analyzed in this paper.

Regarding the user behavior, the vocal communication provides cues, which can be extracted from the speech and language analysis, that provide information about the speaker's feelings that are closely related to human behavior (Siegert et al. 2013). In the same way, the analysis of the facial expression, gazes and gestures can provide information on the non verbal human communication during the interaction even when the user is not talking. So the main focus of the analysis of the elderly behavior while interacting with the simulated virtual coach relies on their affective state, that might work as an indicator of the success of the Virtual Coach (VC). A set of perception experiments were carried out to identify and annotate the emotional status of the seniors. These experiments focused on the users' speech and also on their facial expressions recorded in the interactions. In addition, in order to engage senior users, their preferences and likes had to be considered. To this end, the effect of the VC on the users was gathered by means of direct questionnaires. These questionnaires were completed after the interactions, once the users had a better understanding of the intended system functionality.

The main contributions of this paper rely on the analysis of the behavior of Spanish[1] elderly people when interacting with a WoZ driven, simulated agent. This analysis is mainly based on the identification of the user emotional status as well as on their direct opinions of the system behavior provided through questionnaires. In addition, the behavior of the system will also be analyzed in terms of language and visual aspect.

[1] EMPATHIC project also runs human–machine interactions in France and Norway so that cross cultural analysis will also be carried out in the near future.

The paper is organized as follows: Sect. 2 describes the building procedure of the virtual coach interacting environment, which includes the WoZ platform, the coaching model and the preliminary studies for the agent acceptance. Section 3 describes the way in which the interaction sessions were designed and carried out and the way in which the end users were recruited. Then, in Sect. 4 the behavior of the wizard is analyzed through the language generated and the aspect of the virtual agent. Sections 5 and 6 provide a whole description of the emotional analysis of the user interactions regarding speech, language and facial expressions. Section 7 closes the work summarizing extracted conclusions and providing some cues for future research directions.

2 Building the virtual coach interacting environment

In order to involve seniors in the definition, development and consequent optimization of the EMPATHIC-VC it was necessary to employ various early stage prototyping methods (e.g. use case descriptions, sketches, scenarios, etc.). One of the used methods, which is particularly popular when building technology based on natural language (Schlögl et al. 2015) or other types of artificial intelligence driven applications (Dahlbäck et al. 1993), was Wizard of Oz (WOZ). The key principle of the WOZ method is that study participants believe they are interacting with an autonomous system while actually the system's actions are controlled by a human (i.e. the 'wizard'). In most cases this wizard is situated in a different room and connected to the study setting through a remote network connection. Consequently, WOZ sessions require a minimum of two researchers, i.e. the wizard controlling the technology and an additional facilitator dealing with all the participant related tasks (i.e. welcoming, informed consent, questionnaires, debriefing, etc.). For the EMPATHIC simulated VC both of these researchers received relevant training to prepare them for their tasks. The facilitator had to follow a strict procedural protocol when receiving participants and administrating questionnaires (cf. Sect. 3). The wizard received dedicated training concerning the used WOZ platform (cf. Sect. 2.1) as well as the dialogue structure which had to be followed.

2.1 The Wizard of Oz platform

Since decisions on the overall architecture of a virtual agent based application, such as the one envisioned by the EMPATHIC-VC, usually require extensive discussions, it was decided to use WebWOZ[2] (Schlögl et al. 2010a) as separate WOZ prototyping platform for early stage investigations. WebWOZ, which has been previously used by a number of research and development initiatives (e.g. Cabral et al. 2012; Milhorat et al. 2013; Sansen et al. 2016), offers an adjustable wizard interface which can be structured according to different dialogue stages (Schlögl et al. 2010b, 2011). For simulating interactions with the EMPATHIC-VC, the WebWOZ wizard interface was further extended by an audio/video transmission and recording function based on the WebRTC standard, a graphical representation of the dialogue to help guide the wizard, and the possibility to upload and consequently integrate text-based utterances. In addition, the WebWOZ client interface was integrated with five different virtual agents, which allowed participants to select their preferred interaction partner (Torres et al. 2019b).

2.2 The coaching scenarios implementing the GROW model

Coaching has been defined as a result-orientated systematic process. It generally uses strong questions in order to provide people the capacity of discovering their own abilities and draw on their own resources. In other words, the role of a coach is to foster change by facilitating a coaches' movement through a self-regulatory cycle (Grant 2003). One of the most common used coaching methodologies is the GROW Model (Whitemore 2009). This model provides a simple methodology and an adaptable structure for coaching sessions. Moreover, efficiency has been demonstrated in some Theoretical Behavior Change Models such as the Trans theoretical Model of Change (TTM) (Passmore 2011, 2012).

A GROW coaching dialogue consists of four phases which give the name to the model: Goals or objectives, Reality, Options and Will or action plan. During the first phase (Goal), the interaction aims at getting the specification of the objective that the user wants to achieve, for example, to reduce the amount of salt in order to diminish the related risk of hypertension. Then, this goal has to be placed within the personal context in which the user lives (Reality), and the potential obstacles which needs to be identified. In the next phase (Options), the agent's goal is to incite the user to analyze his/her options in achieving the objective within his/her reality. Then the final goal of the interaction is the specification of an action plan that the user will carry out in order to advance towards goals (Will). The EMPATHIC-VC is planned to deal with four coaching sub-domains: nutrition (Sayas 2018b), physical activity (Sayas 2018c), leisure (Sayas 2018a) and social and family engagement. A professional coach provided a set of handcrafted coaching sessions for each of these sub-domains. The GROW model uses Goal Set Questions (GSQ—e.g."Welcome Jorge, how can I help you?") to define the objective of the user, Motivational Questions

[2] https://github.com/stephanschloegl/WebWOZ.

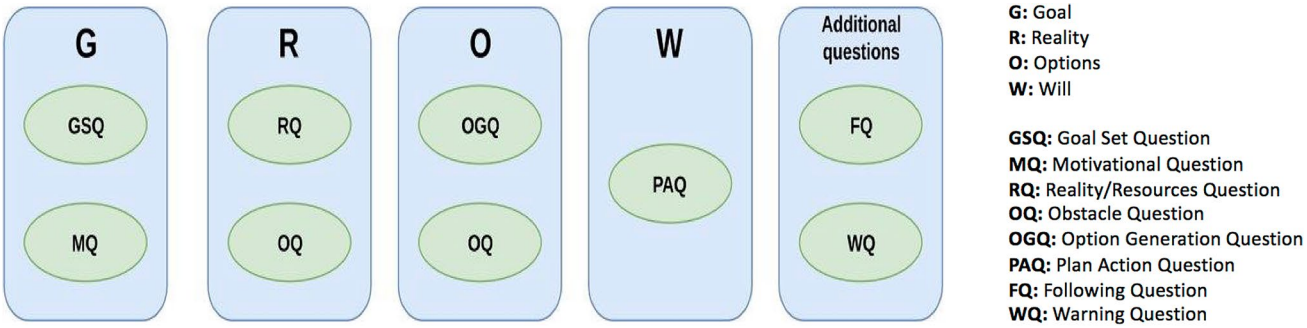

Fig. 1 The structure of the GROW model

Fig. 2 Handmade conversation created by a professional coach

E-To what extent would having regular meal times help you achieve your goal of eating the same or similar amounts of food in each of the main meals? (GSQ)

J- It would bring me much closer to my goal.

E- So what makes you want to have regular meal times? (MQ)

J- I´m clear about that: to manage to eat in a more orderly way, thereby distributing about the overall amount of food across the main meals.

(MQ—e.g."What would you achieve if you changed the way you eat?") to look for some sort of motivation which may help him/her achieve a set goal, Reality/Resources Questions (RQ—e.g."And what happens when you just eat bits?") to analyze the current situation of the user and establish resources, Obstacle Questions (OQ—e.g."And if you are out, what are you going to snack on?") to determine obstacles in the accomplishment of the goal, Option Generation Questions (OGQ—e.g."What small step could you take that would get you closer to your milestone of having meals planned?") to define possible actions a user has to perform in order to achieve the goal, Plan Action Questions (PAQ—e.g."What are you going to do to achieve your goal of adopting a more regular eating pattern?") to establish an action plan, Following Questions (FQ—e.g."How has your plan gone concerning the timing of your meals?") to ask a user about an ongoing plan, and Warning Questions (WQ—e.g."What is your blood pressure like?") to know if the user has any (other) health problems which may need to be considered. The GROW model structure is shown in Fig. 1.

An example of such a handcrafted session is shown in Fig. 2. Then, the wizard strategy was designed according to two different scenarios based on the conversations and indications provided by the professional coach. However, the Wizard had to develop and add new strategies to deal with real user interactions. Thus, a specific wizard profile was created defining a system behavior.

2.3 The user-centered iterative design for elders' virtual agent acceptance

While aiming at implementing a virtual coach devoted to assist the elderly population in their independent living, the goal was to abandon the human–machine interaction techno-centric paradigm and focus on the needs and intentions of the relevant elder end-users, their abilities, aptitudes, preferences, and desires. As for its implementation, the EMPATHIC-VC had as initial requirements accessibility and usability by a wide variety of elderly users, ranging from field experts, practitioners, persons with different knowledge (culture, instruction and occupations), needs (impaired and communicatively disordered individuals) age, and preferences.

To this aims, we have taken a user-centered iterative design, assessing users' interactions in context so that (a) trustworthy human–agent relationships are build, (b) emotional states and negative moods such as depression are reliably detected (Buendia and Devillers 2014; Cavanagh and Millings 2013; DeSteno et al. 2012; Parker and Hawley 2013), and (c) appropriate advice on actions is provided. This was built upon several theoretical experiments, to collect a substantial quantity of data assessing seniors' willingness and interest in initiating and retaining conversations with an agent upon different qualitative agent features (such as gender and voice) in comparison to differently aged populations such as adults and adolescents.

With this research, we have acquired a deeper understanding of how to design emotionally-aware interactive agents that exhibit coherent visual, vocal and gestural affordances, and adapt to the user's underlying intentional and emotional states in a cooperative and ethically sound manner. All the executed experiments were driven by the key idea that any intelligent social ICT interface should be capable of establishing an *empathic relationship*; hence the emphasis of the investigations was on mood enhancement linked to use-cases in e-mental health and support for older/vulnerable people.

The rich repertoire of theoretical results acquired is summarized below, in particular for agent's gender and voice.

A first pilot experiment, focusing on user requirements and expectations with respect to participants' age and familiarity with technological devices (such as smartphones, laptops, and tablets) showed that, as for gender, elders prefer to be assisted by female agents (Esposito et al. 2018b). In this context, an ad-hoc questionnaire was developed to assess senior's preferences, expectations and requirements, in order to customize the consequently developed EMPATHIC-VC to the needs of the targeted end-user population, i.e. elders.

It has to be noted that starting with this pilot, the questionnaire has been gradually modified, in an attempt to incorporate the Theory of Acceptance Model (TAM) proposed by Davis (1989) and the pragmatic and hedonic dimensions proposed by Hassenzahl (2004). The result has been given the name Virtual Agent's Acceptance Questionnaire (VAAQ) and may count as a direct outcome of the Empathic project.

For the above mentioned pilot investigation using an early version of the VAAQ it was further learned that seniors' preference for female agents was significantly higher than for male agents for all the questionnaire dimensions, independently of seniors' genders and technology savviness.

In order to remove the biases introduced by differences in agent's personalities, a second set of experiments was conducted (Esposito et al. 2018a). In these trials, the four proposed agents (two males and two females) were endowed of a "neutral" personality, and their facial expressions were neither smiling, saddening, nor worrying. This test definitively confirmed seniors' preferences to be assisted by female agents which scored significantly better than male agents in all the questionnaire subsections.

In order to assess whether seniors' preferences toward female speaking agents were a specific requirement of the elder population, we defined another set of tests involving adolescents, adults, and seniors for a total of 316 participants split in 7 groups, each composed of approximately 45 subjects, equally balanced for gender (Esposito et al. 2019a). There were two groups of adolescents (mean age = 14.5, SD = ± 0.5 years), two of adults (mean age = 25.1, SD = ± 3.5 years), and two of seniors (mean age = 71.4, SD = ± 6.5 years). It was found that elders' willingness to interact was significantly higher for speaking than mute agents, and, in the speaking context, it was significantly higher for female speaking than male speaking agents. In addition, for elders in the speaking context, female agents were judged significantly more positive than male agents for attractiveness, pragmatic, and hedonic (identity and feeling) qualities. None of these significant differences was observed for adolescents and adults administered with mute and speaking agents and elders administered with mute agents.

When the three elder groups were compared on their enjoyment/acceptance scores for mute, speaking and only voice interfaces, elders' preferences were significantly higher for female speaking agents and only female voice interfaces.

The discussed experiments suggest that the successful incorporation of assistive social technologies in everyday life is strongly depending on the user's perception and acceptance of them (de Graaf et al. 2015). In particular, robots, virtual agents, and generally, interactive assistive user interfaces, need to be specifically tailored to people's needs, and personalized according to their specific requirements and expectations (Seiki et al. 2017),

3 Description of the interaction sessions and user studies

The potential participants for the following interactive study were defined as "healthy seniors" for which the inclusion criteria was: (i) female or male older than 65 years, (ii) living independently (not institutionalized), (iii) being able to read, write and speak fluently in Spanish. For the recruitment of the sample different strategies such as advertising posters, informative notes, mailing and flyers spread in the local areas were used. The consequent study setting employed the previously described WOZ method in order to observe and systematically record both participants' behavior and system operation. In this setting, the first step for participants was to sign an informed consent form before enrolling in the study. Then, the experimental protocol included three steps:

1. The completion of two health questionnaires: Participants were asked to fill in the Geriatric Depression Scale (GDS) and the World Health Organization Quality of Life (WHO-QoL-BREF). The GDS is a dichotomous

Fig. 3 Setup with a participant during a session

("yes" or "no") 30-item (10 negatively worded and 20 positively worded) self-report scale aimed at rating depression (Yesavage et al. 1982). Total scores range from 0 to 30 points, where higher scores mean higher probability of having a depression diagnose. The WHO-QoL-BREF is an abbreviated (26 items) generic quality of life scale developed by the World Health Organization (Who-QoL Group[3]) which assesses four domains: physical health, psychological health, social relationships, and environment. These questionnaires were administered before interacting with the EMPATHIC-VC so as to provide unbiased scores.

2. The interaction with the VC: In this step we used laptops equipped with a webcam, a microphone and a mobile connection (4G/4G+). Participants were logged into a secured session (protected by username and password) with an individual alphanumeric ID code to keep their identity safe. Then, based on their personal preference, they chose one of five available visual representations, i.e. agents, for their VC. Each of these agents (3 female and 2 male) showed different characteristics with respect to their appearance. From that point on, in order to avoid potential impacts the supervisor may have on the dialogue or the actual interaction, participants were left alone with the VC. Two dialogues of 5–10 min each were completed. The first served as an introduction to the system and thus did not focus on any specific issues. The second revolved around a conversation related to nutrition/food (cf. Fig. 3). The structure of this second dialogue was based on the GROW coaching model (Whitmore 2010) presented above. As described earlier, a GROW coaching dialogue consists of four phases; i.e. **G**oals or objectives, **R**eality, **O**ptions and **W**ill or action plan. In the given setting the goal was set on participants' nutritional habits and respective objectives.
3. Finally, after the interaction with the VC, participants were asked to give feedback using a number of user-feedback questionnaires; i.e. the EMPATHIC Virtual Agent Acceptance Questionnaire (VAAQ) (Esposito et al. 2018a), the System Usability Scale (SUS) (Brooke 1996) and the Emotion auto-annotation form. The VAAQ was developed to explore participants' satisfaction in interacting with virtual agents. It contains three sections: (i) the socio-demographic status, (ii) the willingness to be involved in interactions with a Virtual Agent (VA) and (iii) the perceptions of the respective agent features. The SUS contains ten statements regarding a system's usability to which participants respond to on a 5-point Likert scale ranging from "strongly agree" to "strongly disagree". Finally, the emotion auto-annotation form was an ad-hoc questionnaire that asked participants about their two most intense emotions experienced during the contacts with the VC.

A total of 156 WOZ user studies (78 Spanish individuals in 2 sessions) were conducted. The following insights are based on the collected demographic data, the feedback provided by wizards who simulated the EMPATHIC-VC, facilitators and study participants.

3.1 Study set up

Experience has shown that at least two people were required to realistically conduct a WOZ user study, one who acts as a human simulator, i.e. the wizard and one who acts as a facilitator, greeting study participants, introducing them to the study purpose, administrating questionnaires, and helping the participants in case of confusion or technical issues. From a procedural point of view, we further found it imperative that, once the interaction with the VC started, the facilitator had to leave the room. Otherwise the participant tended to look at and talk to the facilitator instead of conversing with the actual agent. This behavior may be explained by a participant's lack of reassurance when interacting with a novel technology.

3.2 Study participants

With our sample size of 78 individuals, we found a higher concentration of users from the first age cohort. That is, 60% of participants were from the age group 65-70, with 69.23% identified as female; and 67.14% had higher education (HE).

In general, we found that the concept of a virtual agent seemed rather frightening to many people of the targeted age group (i.e. aged 65 or older). While we did use face-to-face meetings to overcome this fear as much as possible, it should be noted that for this type of technology anxiety poses a significant challenge, particularly when it comes to

[3] https://doi.org/10.1016/0277-9536(95)00112-K.

the recruitment of study participants. Consequently, recruitment via flyers/posters was difficult (even when conducted in senior centers or elderly homes). However, we found that recommendations coming from other participants who had already taken part and enjoyed the study, helped mitigate the problem. Still, a lot of personal coaching was usually required to make people feel comfortable. Here, our experience has shown that participants needed approx. Ten minutes interaction with the VC to 'lose their fear' regarding the technology—in particular, when studies took place somewhere away from peoples' homes or familiar living environments. With regard to the study inclusion criteria, the studies have shown that elderly people are rather pessimistic when evaluating their personal health status. That is, while initially we were searching for 'healthy' participants aged 65 or older, we had to realize that most representatives of this group would not include themselves due to minor health issues they perceived (e.g. minor hearing problems, minor vision impairments). As for the interaction, it seemed important that participants thought they would interact with a prototypical system. This helped keep the expectations regarding speed and accuracy low. In this context, the speed with which a simulated system responds may be seen a particular challenge. Especially in cases where the wizard could not use a pre-defined utterance and had to type a response. An additional challenge with this generation of on-the-fly utterances concerns the great potential for typos and other mistakes, which are forwarded to the text-to-speech module and, consequently, spoken out loud to a study participant. However, being aware of the prototypical status of the system, study participants were rather tolerant toward these types of issues.

3.3 The dialogues

The participants were usually pre-informed about some of the content to be addressed by the coach so that they could think about relevant topics in advance (e.g., they were told to think about certain goals they would like to achieve before starting the conversation). Such was necessary to keep the interaction going and reduce the number of "yes/no" answers. Still, in particular with respect to the nutrition scenario, it was difficult to keep the conversation flowing, as the scenario was looking for personal goals, yet people were often satisfied with their status-quo and, thus, did not find much to talk about.

Changing the conversational focus due to missing participant goals also caused some side effects. Finally, from a conversational point of view, we found that different types of back-channeling (i.e., approving a participant's input) had a significant influence on the 'smoothness' of the conversation. That is, while rather basic approval utterances such as "interesting" or "good" seemed to distort the conversation,

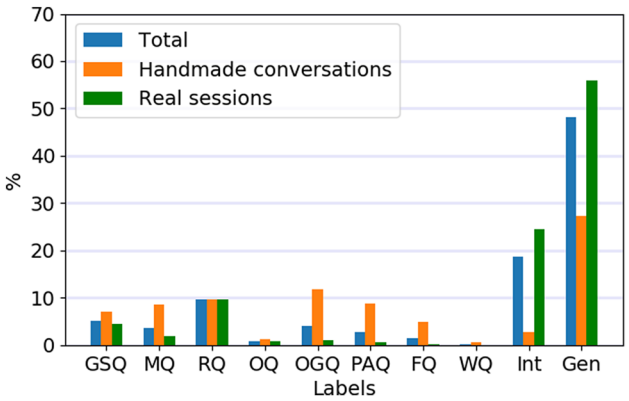

Fig. 4 Distribution of Dialogue Acts

other strategies which re-used participants' words or sentence structures (e.g., Participant: "I like to walk 2 hours every day"; Agent: "You walk 2 hours every day?") helped in keeping participants engaged and consequently the conversation flowing.

4 Analysis of Wizard behavior

The analysis of the Wizard behavior was carried out over the system turns used in the WOZ sessions. These turns were initially established to define each scenario. However, the Wizard had to create new turns to go ahead with the GROW model, to get more developed answers from the user or even to resolve situations caused by the behavior of real users, as analyzed in the previous section. System turns were labeled according to the needs of the Natural Language Generation (NLG) module. They were annotated in terms of Dialogue Acts, Polarity and Entities. These annotations allow for a structured analysis of the VC behavior when addressing participants.

In addition to labeling the WOZ sessions, annotations were also assigned to the set of (handmade) coaching sessions proposed by the professional coach (see Fig. 1). A comparison between the interactions during the conversations with the simulated EMPATHIC-VC and the ones created by the professional coach is a way to test the behavior of the Wizard strategy (although it has to be noted that these VC interactions pose a higher level of artificiality and they might be far from real interactions conducted with a human coach).

4.1 Annotation in terms of Dialogue Acts

The set of Dialogue Acts (DAs) defining the turns of the coach consist of the eight questions used in the GROW mode, i.e. (GSQ), (MQ), (RQ), (OQ), (OGQ), (PAQ),

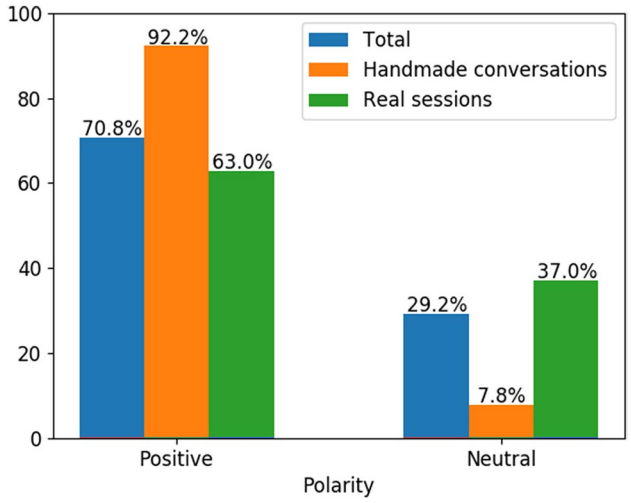

Fig. 5 Distribution of the polarity

(FQ), (WQ) extended by the Introduction label (Int) defining a typical sentence uttered by the coach during the first session with the user, and the General label (Gen) used for all the other interventions the coach performed during the conversations (greetings, agreements, etc.). This labeling structure allows to evaluate the wizard's alignment with the GROW model (cf. Fig. 2).

Figure 4 shows the distribution of DAs for three different sets: the interactions coming from the conversations created by the professional coach, the interactions of the WOZ, and a new set joining the previous two. In all the cases, the most used utterances were general sentences (Gen), which were related to the most common expressions in a conversation. This shows that the wizard was indeed following the instructions of the professional coach considering that a conversation should not be a succession of GROW questions, but it should rather follow a more natural dialogue structure resembling a bidirectional conversation.

The second most frequent label in the wizard data was related to the introduction session (Int). Looking at the overall data distribution it can be observed that there were few sentences in the handmade conversations annotated with this label. This is, however, due to the fact that the professional coach did not follow an entire introduction procedure but rather gave us some sample instructions on how such an introduction session should be mapped out. The wizard data shows that those instructions were followed.

Generally, we can see that the sessions with the human coach exhibit a quite balanced distribution of GROW labels, whereas in the sessions with the wizard a significantly higher number of GSQ, MQ and RQ labels appear, all of which are situated in the initial two phases of the GROW model. This fact suggests that the wizard managed to connect with participants but seemed to have difficulties advancing deeper into the coaching dialogue.

4.2 Annotation of polarity

A key quality defining a successful coach is not to show mood characteristics which may be perceived as negative. For the EMPATHIC-VC it was thus established that the system should express a positive attitude whenever the user's mood is perceived positive or neutral attitude whenever the user's mood is perceived negative.

Keeping a positive attitude was also a guideline expressed by the professional coach. In fact, less than one tenth of the utterances recorded in the handmade sessions were annotated with neutral mood (cf. Fig. 5). In the wizard sessions, there were less differences in polarity values. That is, even though the predominant mood was also positive, a significant number of mood stages were labeled as neutral. Such could be caused by the characteristics of the conversations. As was mentioned before, the sessions dealt mostly with the G and R phases of the GROW model, which focus on exposing the problems that users found in their life. In this context, users often express negative moods and the wizard, consequently, neutral behavior.

4.3 Annotation of entities

One way to make users feel that the machine understands them is to use the same or similar linguistic elements that they have used in their turns. Respective linguistic elements include named entities such as names, places, food, etc. These entities, which have to be identified in the user turns, can be added in the responses of the VA. Thus, in this phase, the annotation focused on identifying all the linguistic elements which can be interpreted as an entity and assigning them to the corresponding category. The following types of entities were defined: *Actions*, *Dates*, *Food*, *Frequency*, *Hobbies*, *Places*, *Quantities*, *Topic*, *User Name* and *Others*. The distribution of these entities in the different types of conversations are shown in Fig. 6.

As Fig. 6 shows, *Action* is the type of entity which was most identified in the sessions with the professional coach. These entities are related to objectives, obstacles, options and action plans of the user. The percentage is lower for the wizard session given that the wizard was mainly working in the first two phases of the GROW model, so some of the entities related to *Action* did never appear.

Another type of entity the professional coach tried to introduce in the conversations is the user's name. This is considered a way of increasing the perceived friendliness. Indeed, the professional coach did frequently use the user's name during a coaching session. The wizard, however, did not seem to include the name so frequently.

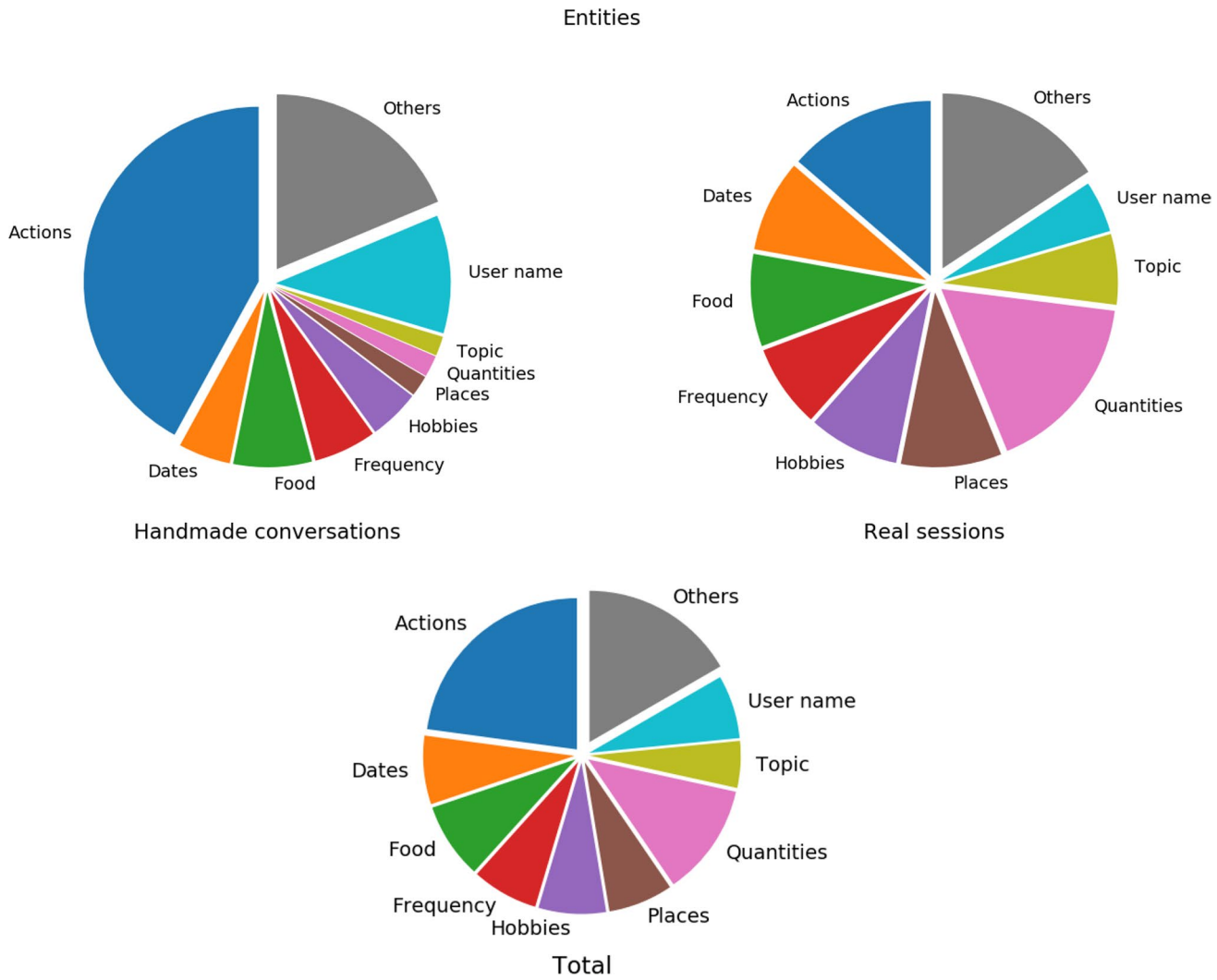

Fig. 6 Distribution of entities for the handmade conversations, for the real wizard sessions and for all the conversations together

As for the other labeled entities, hobbies, music and travel were the main elements users talked about during the first sessions. Thus, we can see that here the entities *Places* and *Hobbies* were among the most frequent ones. Similarly, *Quantities* and *Food* were often identified in the second sessions, related to nutrition.

4.4 Behavior profile of the Wizards

Based on these comparisons carried out between real and handmade sessions, we can thus conclude that the wizard behavior was very similar to the behavior described by the professional coach. Some differences have been found, but they seem to be more related to the progress of the conversation than to the wizard's strategy.

To sum up, we have found that the wizard tried to mix the use of GROW questions or introduction sentences with more general expressions so as to maintain a fluent and natural conversation, while focusing on the actual topic of each session. Furthermore, the wizard kept a positive mood when possible and the neutral mood was employed otherwise. Finally, an attempt was made to let the users lead the conversation without forcing them into achieving the final stages of the GROW model.

4.5 Agent preference

For their interactions, participants had to select one out of five different agents, shown in Fig. 7. The analysis of the agent preference shows that 66.7% of the participants selected a female agent, and that Natalie was the most popular one selected by 49.7% of the participants (remaining 17.9% Lena, 10.3% Alice, 15.4% Christian and 1.3% Adam).

Comparing female and male participants regarding agent gender preference, 79% of male and 70% of female participants

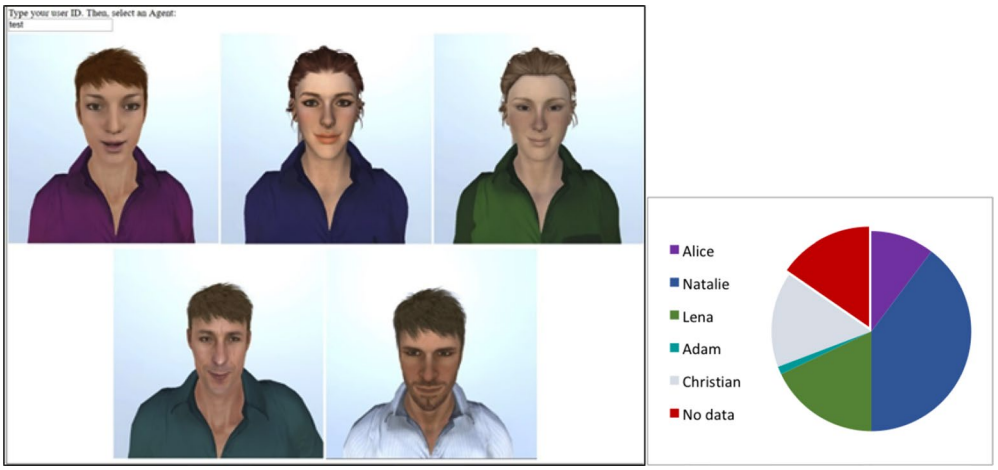

Fig. 7 The five agents that could be chosen by the users and a diagram representing the their selection using different colors

Fig. 8 Annotation in terms of categorical and VAD model of two generic audio segments

preferred a female agent and 21% of male and 8.5% of female participants preferred a male agent.

Concerning the agent's age, the majority of the participants (i.e. 69.2%) preferred an agent looking 29-48 years old; 25.6% preferred an age between 29–38 years, and 43.6% and age between 39–48 years. When asked to guess the age of the agent, participants perceived the agent to be on average 34.4 years old (SD of 5.57). This is a good indication that the 'looks' of the agent did correlate with the participants' preferred age.

Additional comments given by participants concerned the VC's general physical appearance and its latency with respect to responses and movements.

5 Emotional status from speech and language

In order to analyze the emotional status of participants, the conversations between the participants and the wizard were recorded and annotated in terms of emotions. The speech signal was manually labeled from scratch by three Spanish native annotators. Since emotion perception is gender dependent (Vidrascu 2007b), two men and one woman were selected for the annotation task.

The annotators determined manually the emotional state limits (i.e., segment) and also the emotional label associated to that segment. No particular instructions were given to them, except that they should annotate all the signal (no segment without annotation) and that a high agreement between annotators was desirable.

The annotation was made in terms of both a categorical and a dimensional model. The dimensional VAD model is a psychological model that characterizes affect states in terms of two or three dimensions, namely valence, arousal and dominance (VAD) (Gunes and Pantic 2010; Valstar et al. 2014). Thus, the annotators assigned four labels to each audio segment: one label related to a specific category and 3 additional labels, one related to the valence, another one to the arousal and a final one to the dominance as shown in Fig. 8.

The annotation procedure was organized in three steps. First, a set of files was chosen to be annotated by each annotator separately. Then, the inter-annotator agreement was computed. If the agreement was less than a predefined threshold, the annotators discussed and re-annotate the files as shown in Fig. 9. Following this procedure, we managed to reach an agreement level for the categorical model annotations that was greater than 90% for all emotions and even 100% for *sad* and *tense*.

5.1 Analysis of dimensional annotation

The labels assigned to the dimensional VAD model were:

- Valence: positive, neither positive nor negative, negative
- Arousal: excited, slightly excited, neutral

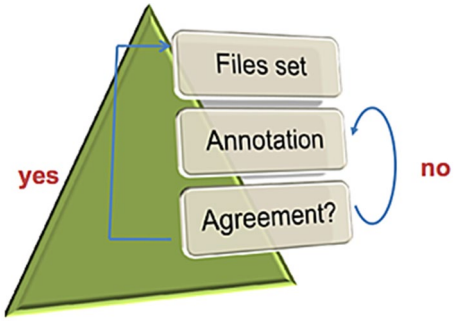

Fig. 9 Annotation procedure

- Dominance: dominant, neither dominant nor intimidated, defensive

The three labels assigned to each segment were converted to a real point in a 3D space where the axes correspond to valence, arousal and dominance. To this end, a discrete value was assigned to each level assuming that all levels are equidistant. For instance, the assigned values to the different levels of arousal are Excited: 1, Slightly excited: 0.5, Neutral: 0. Then, the average value of the annotations provided by the three annotators was computed to represent each annotated segment in the 3D space.

Figure 10 shows the probability density function of each variable (valence, arousal, dominance) estimated by using a Gaussian kernel density estimator. The results show that, in most of the cases, low values of Arousal along with positive values of Valence and neutral values of Dominance were obtained. This indicates that in the interaction with the wizard users did not achieve high excitation levels, which corresponds with our expectations, and means that the interaction with the system did not unsettle people. The dominance values were also quite neutral, neither dominant nor intimidated. This means that the behavior of the virtual coach was appropriate and did not make people feel intimidated.

Table 1 Number of segment annotated with each category label

Annotation	Calm	Sad	Amused	Puzzled	Tense
First	7017	17	260	347	12
Second	7794	19	292	297	24
Third	7655	21	244	360	20
Agreement	3368	12	100	90	13

Finally, the valence results show that users have positive feelings with regard to the interaction with the system.

5.2 Categorical model

The categorical labels assigned to each audio segment were: *calm/tired/bored*, *sad*, *amused/satisfied*, *puzzled* and *tense*. Let us note that some categories were intentionally combined into one label, (e.g. calm, tired and bored) because we saw in previous experiments that they were frequently mixed up in this task. Moreover, keeping a long list of categories would have increased the difficulty of the annotators' task, providing lower agreement values. From now on, the mixed categories will be referred to by their first label, that is *calm* for *calm/tired/bored* and *amused* for *amused/satisfied*.

The annotation of the database by the three annotators led to the categories shown in Table 1. Additionally, we decided to consider for our work only those segments where all the three annotators agreed on the given label. Since the segment limits were also defined by the annotators there could be a mismatch among them, thus, we selected the segment intersection with the same label from all the annotations, even if this led to a fragmentation of some segments into two or more different ones. This happened, for instance, with *tense*, where there were 12 segments from the first annotator and a higher number of segments (13) when agreement was required.

According to the obtained results, it can be concluded that the most frequent label was *calm*, suggesting that the

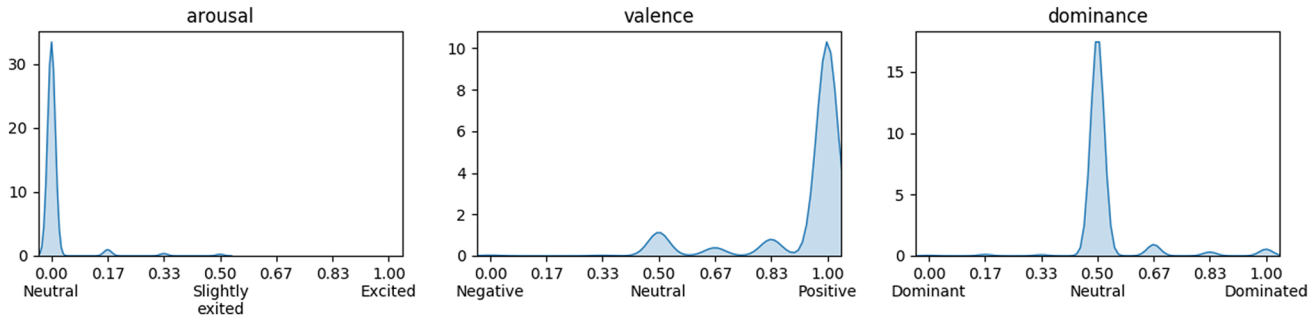

Fig. 10 The probability density function of Valence, Arousal and Dominance according to the data

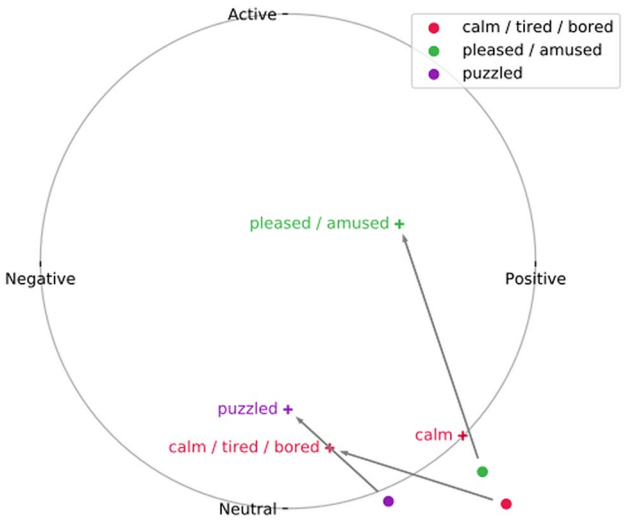

Fig. 11 Comparison between the theoretical and experimental values of arousal and valence associated to each emotion. Experimental data (circles) were achieved by computing the average of valence and arousal values of all the audio segments labeled with a specific emotion. The theoretical points (crosses) were extracted from the diagram given in Cambria et al. (2012)

dialogue system seems to have been perceived as friendly by the users. The labels *sad* and *tense* were quasi absent. In fact, when agreement between the annotators was required they almost disappeared, which means that the virtual coach did not provoke such negative feelings in users while they were interacting.

In addition, we compared the theoretical and experimental values of arousal and valence associated to each emotion. To do so, the average of valence and arousal values of all the audio segments labeled with a specific emotion was computed and the corresponding point represented in Fig. 11 using a circle. Then, a cross was used to represent the position of the same emotions in the ⟨ arousal, valence⟩ space according to the diagram given in Cambria et al. (2012). For instance, the purple circle in Fig. 11 was achieved by computing the average of valence and arousal values for all the samples labeled as *puzzled* in our database. The purple cross, instead, was extracted from the diagram given in Cambria et al. (2012) that represents where the values of valence and arousal should theoretically be for *puzzled*. Note that *Tense* and *sad* emotions were removed from Fig. 11 because there were not enough samples to represent them confidently. In order to theoretically represent the mixed classes, *calm/tired/bored* and *amused/pleased*, the mid point of the different classes was computed. That is, for *amused/pleased* the theoretical points of *amused* and *pleased* were considered and the mid point between them computed. According to this comparison, it can be concluded that in our experiments there were some differences regarding theoretical values. The differences might be due to the specific task, where real and not acted emotions were involved to some extent, as it also happened in deVelasco et al. (2019) where a different task was considered. Specifically, the real arousal values of our experiments were lower than the expected ones for all emotions. This could be due to the fact that real emotions seem to be more subtle than the acted ones and not so extreme. However, real values of valence were higher than the expected ones for all categories in our experiments. This might also due to the task. Participants that accept to take part in such a trial are usually curious and show a positive predisposition with regard to the situation. Furthermore, they interact with a system in a controlled environment with other people in the surrounding, not alone, so they usually do not allow negative feelings to rule their behavior. Looking at the vectors illustrating the differences between real and theoretical values, the one related to *calm* appears to be a bit different from the others. We may explain this by showing the theoretical value of calm. It seems that most annotators that used *calm/tired/bored* actually were labeling with calm and not with the other two emotions.

5.3 Analysis from text

Besides the acoustic information, there are other sources that can provide information about the emotional status of the users. The semantic meaning involved in a user utterance, for instance, can provide complementary information in some scenarios (Justo et al. 2018). Thus, we analyzed the users behavior focusing on the text obtained from the transcriptions of the user utterances when they were interacting with the system. Specifically, we consider the polarity of the text associated to the utterances.

Firstly, the transcriptions of the audio recordings were manually extracted, in terms of user turns, by professional annotators. Then, each transcription was manually labeled by Spanish native annotators. Although up to nine different annotators were involved in the process, only one annotator labeled each transcript. They were asked to consider each user turn and divide it into segments according to the topic they were dealing with and then to assign a polarity value to the corresponding segment. The possible polarity values were: *negative*, *neutral* and *positive*.

The histogram of the segments labeled with the different polarity values is shown in Fig. 12, where it can be observed that more than 60% of the segments were neutral and around 28% positive. Negative segments were almost absent, accounting for only 5% of the total segments. These results can be compared to the valence values obtained from the annotation of acoustic segments. In both cases the negative values are not significant, meaning that the users do not show negative feelings when regarding the interactions

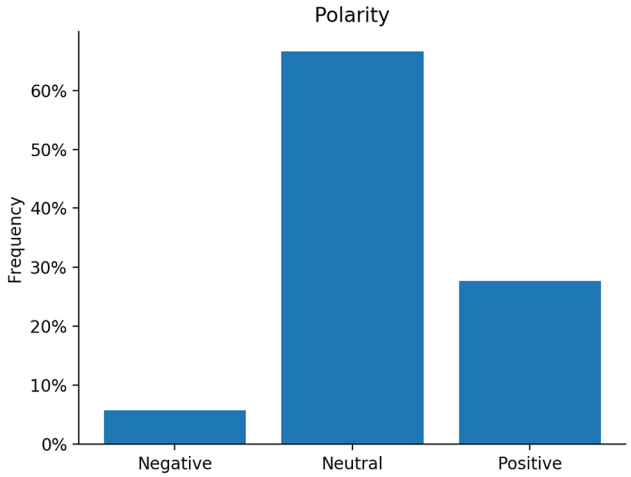

Fig. 12 Histogram with the text segments labeled with each polarity value

with the system, as mentioned above. However, in this case there are more segments labeled as neutral and less labeled as positive. This might be due to the bias associated to the specific annotators and also due to the information itself, because there might be utterances were a positive feeling can be perceived from the acoustic information but the semantic meaning of the message does not imply anything that denotes positiveness. For instance, the text *'I usually eat varied fruits and vegetables'* does not show a positive polarity but, depending on the way it is pronounced, it might be associated to an acoustic segment labeled with a positive valence.

6 What the participant facial expressions says

Following the analysis of Sect. 5, the visual modality was also manually labeled from scratch by two Spanish annotators to analyze facial expressions of emotion in user-wizard interaction videos. To guarantee that only visual information was taken into account, videos were muted throughout the annotation procedure.

6.1 Annotation protocol

As for speech, the annotators determined the emotional state limits and the emotional label associated to that segment using a categorical model. This time, however, the annotators were instructed with particular guidelines to follow so as to ensure a common annotation protocol. First, only facial expressions and head movements had to be taken into account to annotate emotions. Out-of-face information, such as body and hand movements, were out of the scope of the annotation. Second, some participants have a specific neutral expression according to their physiognomy. For instance, some people have facial features that can be perceived as happy, even though they are not trying to communicate a state of happiness at that time. To learn this baseline neutral face, annotators were requested to watch the whole video once before starting the annotation procedure.

The annotation procedure was similar to that of emotions from speech (see Fig. 9). First, a subset of 4 videos was selected to train the annotators. Once annotated, the inter-annotator agreement was computed. Then, annotators were requested to discuss the minimum level an expression must be perceived to label it with a specific category, to reach a consensus and re-annotate the files with their updated protocol. This process continued until all the videos reached a valid inter-annotator agreement.

The categorical labels assigned to each segment were: *sad, annoyed/angry, surprised, happy/amused, pensive* and *other*. As for speech, some categories were combined in one label due to the outcome of previous experiments. The first four are included in Ekman's universal expressions of emotion (Ekman and Keltner 1997). *Pensive* is not an emotion per se; however, it is included in our model as it has shown to be a frequent facial expression present in conversation and it is informative of our internal and cognitive states (El Kaliouby and Robinson 2005; Rozin and Cohen 2003). Annotators were instructed to annotate as one of the first 5 categories those segments in which it was clear for them that the expression was present. *Other* was used to denote either those segments in which one expression was taking place but which was not included in our expression list, or when more than one expression from the list was present. Finally, all non-labeled instances were considered to be a *neutral* expression, denoting the baseline face as well as calmed, quiet, or very subtle emotions which do not exceed the consensual expression thresholds.

6.2 Analysis of categorical annotations

The final inter-rater agreement level for our selected categorical annotations was high, above 80% on average. For the reminder of the section, we consider as gold standard those segments where both annotators agreed on a given label. Following the speech analysis, we also selected the segment intersection with the same label for both annotators. To do so, we included *neutral* as another label, even though it was not manually labeled. This intersection procedure caused some small segments to have no assigned label, which usually happens at the annotated segment limits when the onset/offset of a facial expression takes place.

Table 2 reports the number of segments for each category and annotator, as well as the gold standard (*Agreement*),

Table 2 Number of segments annotated with each category label

Annotation	Sad	Angry	Surprised	Happy	Pensive	Other	Neutral	Total
First	0	0	12	234	2033	0	2250	4529
Second	0	1	44	151	2060	3	2245	4504
Agreement	0	0	5	141	1825	0	2382	4353

Table 3 Percentage of each category label with respect to the total amount of annotated time

Annotation	Sad	Angry	Surprised	Happy	Pensive	Other	Neutral
Agreement	0	0	0.01	0.63	11.95	0	87.41

for all annotated videos. Table 3 shows the frequency of each category for the gold standard with respect to the total amount of annotated time. As we can observe, *pensive* is the most frequent manually-labeled expression, appearing 12% of the time, followed by *happy/amused*, present in 6% of the total annotated time. The lack of perceived negative emotions is in line with Sect. 5's results. This suggests that the interaction with the wizard was positive and that the users were engaged in the conversation. Despite such findings, the absence of facial expressions (*neutral*) clearly dominates over all categories. While it is on par with *pensive* with respect to number of segments, participants spent most of the interaction (around 87% of the time) showing no apparent emotion. This is expected, as users had to wait for the system responses for most part of the interaction. However, this is a good sign, as participants could have started to feel angry or sad due to such waiting times, but instead they tended to remain calm.

It is worth noting that, even though the speech and video results follow the same trend, the total annotated time for video is much higher than for speech, due to the fact that speech instances constitute just a fraction of the whole recorded interaction. Therefore, there is more emotional information from video than from audio in a user-wizard interaction. There are indeed some expressions of emotion that can be better perceived from video than from speech. *Pensive*, for instance, which appears frequently right before or while speaking, can only be inferred from the visual modality. However, what we say and how we say it are also informative of our emotional status. Hence, information from speech, language semantics and facial expressions should be combined in a multi-modal manner in order to better understand the emotional status of the user at a given time.

7 Conclusion and future work

This paper analyzed the sessions that a selected set of Spanish seniors carried out to interact with a Wizard of Oz driven, simulated system. A coaching strategy based on the GROW model was used throughout these sessions. In this interaction framework, both the human and the system behavior were analyzed. Regarding the system behavior, the analysis concluded that the wizard was not intended to implement a succession of GROW questions, but rather follow a more natural dialogue structure resembling a bidirectional conversation. Moreover, the sessions with the wizard show that the conversation stayed mainly at the initial two phases of the GROW model, that is, the wizard managed to connect with participants but seemed to have difficulties advancing deeper into the coaching dialogue.

On the other hand, the wizard mood was frequently labelled as positive (60%) when analyzing the language generated to be pronounced by the virtual agent, and neutral the remaining times. A higher percentage of positive content will probably be achieved when the sessions go beyond the G and R phases of the GROW model.

Regarding the user behavior, the probability density function of each dimension (valence, arousal, dominance) of the VAD model show low values of Arousal along with positive values of Valence and neutral values of Dominance, when analyzing the emotional labels extracted from speech. This indicates that users did not achieve high excitation levels when interacting with the Wizard. In fact, the behavior of the virtual coach seems to be appropriate and did not make people feel intimidated. Finally, the valence results show that users have positive feelings with regard to the interaction with the system. In terms of categories, the most frequent label was *calm*, suggesting also that the dialogue system seems to been perceived user friendly.

Regarding the analysis of the emotional labels associated to video segments the most frequent label was *pensive* followed by *happy/amused*. The lack of perceived negative emotions is in line with results obtained from the speech analysis. This also suggests that the interaction with the wizard was positive and that the users were engaged in the conversation. Despite such findings, participants spent most of the interaction (around 87% of the time) showing no apparent emotion, because users had to wait for the system responses for most part of the interaction.

It is worth noting that, even though the speech and video results follow the same trend, the total annotated time for video is much higher than for speech, due to the mute part of videos where users were waiting for the system's interaction. Some expressions of emotion can only be inferred from the visual modality since they are mainly associated to silent participant. Hence, information from speech, language semantics and facial expressions should be combined in a multi-modal manner in order to better understand the emotional status of the user at a given time.

Future work will include the analysis of on going interaction sessions carried out in Norway and France, which will allow a cross-cultural analyses of the user behavior. Moreover, cross-model analysis of the emotional analysis will be carried out in depth by also including the results of the questionnaires in the analysis.

Acknowledgements The research presented in this paper is conducted as part of the project EMPATHIC that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no 769872.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Anna E, Terry A, Marialucia C, M EA, Alda T, Inées TM, Stephan S, Gennaro C (2019) Seniors' acceptance of virtual humanoid agents. In: Leone A, Caroppo A, Rescio G, Diraco G, Siciliano P (eds) Ambient assisted living. Springer International Publishing, Cham, pp 429–443

Brinkschulte L, Mariacher N, Schlögl S, Torres MI, Justo R, Olaso JM, Esposito A, Cordasco G, Chollet G, Glackin C et al (2018) The empathic project: building an expressive, advanced virtual coach to improve independent healthy-life-years of the elderly. In: SMARTER LIVES 2018: digitalisation and quality of life in the ageing society. Universität Inssbrück, pp 36–52

Brooke J (1996) SUS-A quick and dirty usability scale. Usability evaluation in industry. CRC Press, Boca Raton ISBN: 9780748404605

Buendia A, Devillers L (2014) From informative cooperative dialogues to long-term social relation with a robot. Natural interaction with robots, knowbots and smartphones. https://doi.org/10.1007/978-1-4614-8280-2_13

Cabral JP, Kane M, Ahmed Z, Abou-Zleikha M, Székely E, Zahra A, Ogbureke KU, Cahill P, Carson-Berndsen J, Schlögl S (2012) Rapidly testing the interaction model of a pronunciation training system via wizard-of-oz. In: Proceedings of the LREC international conference on language resources and evaluation, Istanbul

Cambria E, Livingstone A, Hussain A (2012) The hourglass of emotions. In: Esposito A, Esposito AM, Vinciarelli A, Hoffmann R, Müller VC (eds) Cognitive behavioural systems. Springer, Berlin

Cavanagh K, Millings A (2013) (Inter)personal computing: the role of the therapeutic relationship in e-mental health. J Contemp Psychother 43(4):197–206. https://doi.org/10.1007/s10879-013-9242-z

Cordasco G, Esposito M, Masucci F, Riviello MT, Esposito A, Chollet G, Schlogl S, Milhorat P, Pelosi G (2014) Assessing voice user interfaces: the vassist system prototype. In: 2014 5th IEEE conference on cognitive infocommunications (CogInfoCom), pp 91–96

Dahlbäck N, Jönsson A, Ahrenberg L (1993) Wizard of oz studies—why and how. Knowl Based Syst 6(4):258–266

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 13(3):319–340. http://www.jstor.org/stable/249008

de Graaf M, Ben Allouch S, Klamer T (2015) Sharing a life with harvey: exploring the acceptance of and relationship-building with a social robot. Comput Hum Behav 43:1–14. https://doi.org/10.1016/j.chb.2014.10.030

DeSteno D, Breazeal C, Frank RH, Pizarro D, Baumann J, Dickens L, Lee JJ (2012) Detecting the trustworthiness of novel partners in economic exchange. Psychol Sci 23(12):1549–56

deVelasco M, Justo R, López-Zorrilla A, Torres MI (2019) Can spontaneous emotions be detected from speech on tv political debates? In: Proceedings of 10th IEEE international conference on cognitive infocommunications **(in press)**

Ekman P, Keltner D (1997) Universal facial expressions of emotion. In: Segerstrale U, Molnar P (eds) Nonverbal communication: Where nature meets culture, pp 27–46

El Kaliouby R, Robinson P (2005) Real-time inference of complex mental states from facial expressions and head gestures. In: Real-time vision for human–computer interaction. Springer, Berlin, pp 181–200

Esposito A, Amorese T, Cuciniello M, Esposito AM, Troncone A, Torres MI, Schlögl S, Cordasco G (2018a) Seniors' acceptance of virtual humanoid agents. In: Italian forum of ambient assisted living. Springer, Berlin, pp 429–443

Esposito A, Schlögl S, Amorese T, Esposito A, Torres MI, Masucci F, Cordasco G (2018b) Seniors' sensing of agents' personality from facial expressions. In: Miesenberger K, Kouroupetroglou G (eds) Computers helping people with special needs. Springer International Publishing, Cham, pp 438–442

Esposito A, Amorese T, Cuciniello M, Riviello MT, Esposito AM, Troncone A, Cordasco G (2019a) The dependability of voice on elders' acceptance of humanoid agents. In: Proc. Interspeech 2019, pp 31–35. https://doi.org/10.21437/Interspeech.2019-1734

Esposito A, Amorese T, Cuciniello M, Riviello MT, Esposito AM, Troncone A, Torres MI, Schlögl S, Cordasco G (2019b) Elder user's attitude toward assistive virtual agents: the role of voice and gender. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-019-01423-x

Grant AM (2003) The impact of life coaching on goal attainment, metacognition and mental health. Soc Behav Pers 31(3):253–263

Gunes H, Pantic M (2010) Automatic, dimensional and continuous emotion recognition. Int J Synth Emot 1(1):68–99. https://doi.org/10.4018/jse.2010101605

Hassenzahl M (2004) The interplay of beauty, goodness, and usability in interactive products. Hum Comput Interact 19(4):319–349

Hassenzahl M (2008) The interplay of beauty, goodness, and usability in interactive products. Hum Comput Interact 19(4):319–349. https://doi.org/10.1207/s15327051hci1904_2

Irastorza J, Inés Torres M (2019) Tracking the expression of annoyance in call centers. Springer International Publishing, Cham, pp 131–151. https://doi.org/10.1007/978-3-319-95996-2_7,

Justo R, Saz O, Guijarrubia V, Miguel A, Torres MI, Lleida E (2008) Improving dialogue systems in a home automation environment. In: Proceedings of the 1st international conference on ambient media and systems. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Ambi-Sys '08, pp 2:1–2:6. http://dl.acm.org/citation.cfm?id=1363163.1363165

Justo R, Manso JI, Pérez S, Torres MI (2018) Bi-modal annoyance level detection from speech and text. Procesamiento del Lenguaje Natural 61:83–89. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5647

Kirkwood TB (2005) Understanding the odd science of aging. Cell 120(4):437–447

Milhorat P, Schlögl S, Chollet G, Boudy J (2013) What if everyone could do it?: A framework for easier spoken dialog system design. In: Proceedings of the 5th ACM SIGCHI symposium on engineering interactive computing systems. ACM, New York, EICS '13, pp 217–222. https://doi.org/10.1145/2494603.2480325

Montenegro C, López Zorrilla A, Mikel Olaso J, Santana R, Justo R, Lozano JA, Torres MI (2019) A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. Multimodal Technol Interact 3(3):52. https://doi.org/10.3390/mti3030052

Parker SG, Hawley MS (2013) Telecare for an ageing population? Age Ageing 42(4):424–425. https://doi.org/10.1093/ageing/aft056. http://oup.prod.sis.lan/ageing/article-pdf/42/4/424/28849/aft056.pdf

Passmore J (2011) Motivational interviewing—a model for coaching psychology practice. Coach Psychol 7(1):35–39

Passmore J (2012) An integrated model of goal-focused coaching: an evidence-based framework for teaching and practice. Int Coach Psychol Rev 7(2):146–165

Pedersen T, Johansen C, Jøsang A (2018) Behavioural computer science: an agenda for combining modelling of human and system behaviours. Hum Centric Comput Inf Sci 8(1):7. https://doi.org/10.1186/s13673-018-0130-0

Rozin P, Cohen AB (2003) High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of americans. Emotion 3(1):68

Sansen H, Chollet G, Glackin C, Badii A, Torres MI, Petrovska-Delacrétaz D, Schlögl S, Boudy J (2016) The Roberta IRONSIDE Project: a humanoid personal assistant in a wheelchair for dependent persons. In: Proceedings of the ATSIP international conference on advanced technologies for signal and image processing, Monastir. https://doi.org/10.1109/ATSIP.2016.7523110

Sayas S (2018a) Dialogues on leisure and free time. Tech. Rep. DP3, Empathic project

Sayas S (2018b) Dialogues on nutrition. Tech. Rep. DP1, Empathic project

Sayas S (2018c) Dialogues on physical exercise. Tech. Rep. DP2, Empathic project

Schlögl S, Doherty G, Karamanis N, Luz S (2010a) Webwoz: a wizard of oz prototyping framework. In: Proceedings of the 2nd ACM SIGCHI symposium on engineering interactive computing systems. ACM, New York, EICS '10, pp 109–114. https://doi.org/10.1145/1822018.1822035

Schlögl S, Doherty G, Karamanis N, Scheider A, Luz S (2010b) Observing the wizard: in search of a generic interface for wizard of oz studies. In: Proceedings of the Irish HCI conference, Dublin, Ireland, pp 43–50

Schlögl S, Schneider A, Luz S, Doherty G (2011) Supporting the wizard: interface improvements in wizard of oz studies. In: Proceedings of the BCS HCI conference on human–computer interaction, Newcastle

Schlögl S, Doherty G, Luz S (2015) Wizard of oz experimentation for language technology applications: challenges and tools. Interact Comput 27(6):592–615. https://doi.org/10.1093/iwc/iwu016

Seiki ST, Tamamizu K, Saiki S, Nakamura M, Yasuda K (2017) Virtualcaregiver: personalized smart elderly care. Int J Softw Innov (IJSI) 5:1–14

Siegert I, Hartmann K, Philippou-Hübner D, Wendemuth A (2013) Human behaviour in hci: Complex emotion detection through sparse speech features. In: Proceedings of 4th international workshop on human behavior understanding, vol 8212. Springer, New York, pp 246–257. https://doi.org/10.1007/978-3-319-02714-2_21

Torres MI, Olaso JM, Glackin N, Justo R, Chollet G (2019a) A spoken dialogue system for the empathic virtual coach. In: D'Haro LF, Banchs RE, Li H (eds) 9th International workshop on spoken dialogue system technology. Springer Singapore, Singapore, pp 259–265

Torres MI, Olaso JM, Montenegro C, Santana R, Vázquez A, Justo R, Lozano JA, Schlögl S, Chollet G, Dugan N, Irvine M, Glackin N, Pickard C, Esposito A, Cordasco G, Troncone A, Petrovska-Delacretaz D, Mtibaa A, Hmani MA, Korsnes MS, Martinussen LJ, Escalera S, Cantariño CP, Deroo O, Gordeeva O, Tenorio-Laranga J, Gonzalez-Fraile E, Fernandez-Ruanova B, Gonzalez-Pinto A (2019b) The empathic project: mid-term achievements. In: Proceedings of the 12th ACM international conference on pervasive technologies related to assistive environments, ACM, New York, PETRA '19, pp 629–638. https://doi.org/10.1145/3316782.3322764

Valstar M, Schuller B, Smith K, Almaev T, Eyben F, Krajewski J, Cowie R, Pantic M (2014) Avec 2014: 3d dimensional affect and depression recognition challenge. In: Proceedings of the 4th international workshop on audio/visual emotion challenge. ACM, New York, AVEC '14, pp 3–10

Vidrascu L (2007a) Analyse et détection des émotions verbales dans les interactions orales. Ph.D. thesis, Paris11 University

Vidrascu L (2007b) Analysis and detection of emotions in real-life spontaneous speech. Theses, Université Paris Sud-Paris XI. https://tel.archives-ouvertes.fr/tel-00624085

Whitemore J (2009) Coaching for performance: growing human potential and purpose: the principles and practice of coaching and leadership. Nicholas Brealey Publishing, London

Whitmore J (2010) Coaching for performance: growing human potential and purpose: the principles and practice of coaching and leadership. People skills for professionals. Nicholas Brealey Publishing. https://books.google.es/books?id=eTZiP_8dqIYC

Willcox DC, Scapagnini G, Willcox BJ (2014) Healthy aging diets other than the Mediterranean: a focus on the okinawan diet. Mech Ageing Dev 136–137:148–162

Yesavage JA, Brink T, Rose TL, Lum O, Huang V, Adey M, Leirer VO (1982) Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res 17(1):37–49

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.