Design of Multimodal Interaction with Mobile Devices
Challenges for Visually Impaired and Elderly Users

Michela Ferron[1], Nadia Mana[1], Ornella Mich[1] and Christopher Reeves[2]
[1]*Fondazione Bruno Kessler, Via Sommarive 18, Trento, Italy*
[2]*Streetlab, 17 Rue Moreau, 75012 Paris, France*

Keywords: Visually Impaired, Elderly People, mid-Air Gestures, Multimodal, Smartphone, Tablet PC.

Abstract: This paper presents two early studies aimed at investigating issues concerning the design of multimodal interaction - based on voice commands and mid-air gestures - with mobile technology specifically designed for visually impaired and elderly users. These studies have been carried out on a new device allowing enhanced speech recognition (interpreting lip movements) and mid-air gesture interaction on Android devices (smartphone and tablet PC). The initial findings and challenges raised by these novel interaction modalities are discussed. These mainly centre on issues of feedback and feedforward, the avoidance of false positives and point of reference or orientation issues regarding the device and the mid-air gestures.

1 INTRODUCTION

Accessing ICT (Information and Communication Technologies) is essential to participate in a modern, interconnected society that relies on technology for handling everyday tasks. The visually impaired and elderly people are very different user groups in terms of needs, desires and requirements, but both face significant barriers in accessing ICT because of their fragilities and/or disabilities (e.g., perception, cognition and movement control). Voice commands and non-touchscreen-based hand and finger gestures might make human-computer interaction easier and more natural for these user groups (De Carvalho Correia et al., 2013).

The visually impaired population is already used to assistive technology to help them in their daily living activities, with one of the drawbacks being that this technology is often specially manufactured and so tends to be expensive. Recently, however, accessibility features such as speech output, speech recognition and customisable screens (large text, personalisable colours and contrast) have been built into mainstream portable products, especially smartphones, which has helped access for these users. This covers, for example, the Android (TalkBack, Google Now), Apple iOS (VoiceOver, Siri) and Windows (Narrator, Cortana) platforms.

Older adults are frequently portrayed as generally resistant to technology (Ryan et al., 1992), but a substantial amount of studies have showed that they do not reject technology more than other age groups. On the contrary, they are willing to use novel technologies when these meet their needs and expectations (Fisk et al., 2009; Lindsay et al., 2012). Vocal and gestural interaction can potentially increase the accessibility of elderly users to technology, because they can allow users to overcome the difficulties related to motor disabilities (e.g., when fine movements are required to select small icons on touch interfaces). As well as this, gestural interfaces are considered an effective way to reduce the learning curve (Grandhi et al., 2011) and should present advantages over other interaction paradigms as people already express themselves through gestures in their everyday social interactions. For these reasons, vocal and gestural interfaces could foster technology adoption in those user groups, such as older adults, who find traditional technology difficult to use.

To help build on these interaction modalities, the ECOMODE project was set up, funded by the European Commission under the Horizon 2020 Programme (see Section 2). ECOMODE makes use of a new 'event driven' camera (Clady et al., 2017) to enhance existing interaction modalities, such as speech recognition (by combining speech recognition with lip movement analysis), and to introduce more novel interaction techniques, such as mid-air gesture recognition. This paper reports on

Ferron, M., Mana, N., Mich, O. and Reeves, C.
Design of Multimodal Interaction with Mobile Devices - Challenges for Visually Impaired and Elderly Users.
DOI: 10.5220/0006638901400146
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 2: HUCAPP*, pages 140-146
ISBN: 978-989-758-288-2

two data collection studies, aimed at building a dataset of mid-air gestures and voice commands to be used for training the recognition algorithms, and discusses the challenges that we faced during these studies regarding the design of multimodal interaction for visually impaired and elderly users.

2 ECOMODE TECHNOLOGY

The ECOMODE project aims at exploiting the recently matured biologically-inspired technique of event-driven compressive sensing (EDC) of audio-visual information, for enabling more natural and efficient ways of human interaction with ICT.

One of the main challenges of the project is to integrate different EDC technology hardware components, and to combine them into battery-powered mobile devices, such as tablet computers and smartphones. While traditional techniques are slow and computationally expensive because they use cameras to process sequences of images frame-by-frame, EDC technology is frame-free. This technology exploits a mid-air gesture control set processing for hand and finger gesture recognition, and a vision-assisted speech recognition set that combines auditory input with visual information from lip and chin motion, to gain robustness and background noise immunity in the recognition of spoken commands and speech-to-text input. These characteristics allow EDC technology to work efficiently in challenging conditions, such as poor lighting and high background noise conditions.

Figure 1: ECOMODE prototype running on smartphone.

The first step of the project was to integrate different hardware components to build a camera prototype that was attached to the top of an Android smartphone and tablet (see Figure 1), in order to make possible the first data collection to train the recognition algorithms. A further step will be to evaluate the prototype with visually impaired and elderly users to refine the design specifications of the technology.

3 USER STUDY

While carrying out data collection aimed at building a dataset of mid-air gestures and voice commands, the recording sessions with the target users have been exploited to investigate users' needs, preferences and requirements, as well as how they performed and remembered the multimodal commands.

3.1 Participants

Whilst the visually impaired population tends to be mainly over 60 years of age, the population group is actually varied both in terms of their age range and the types of visual deficiency that affects them (World Health Organization, 2012). For our studies, we use the World Health Organisation (WHO, 2004) categorisation with a split between categories 1 to 3 (partially sighted) and categories 4 and 5 (blind). For the data collection with visually impaired people, 17 adults (7 females; M=51 years-old; SD=12) were recruited from the database set up and run by Streetlab. Thirteen of them fell into WHO categories 1 to 3 (partially sighted) and 4 fell into WHO categories 4 and 5 (blind).

For the older adults group, 20 participants (10 females; M=70.63 years-old; SD=8.61) were recruited among volunteers (relatives and acquaintances of colleagues and friends) and members of a local senior association. According to the categorisation by age proposed by Fisk et al. (2009), 13 were "young older adults", i.e. 60-75 years old (M=65.92, SD=4.97), and 7 were "old older adults", i.e. over 75 (M=80.14, SD=4.91). According to the categorisation based on phycho-physical conditions (Gregor et al., 2002), 13 were "fit older adults" (able to live independently, with no main disabilities), 6 were "frail older adults" (with one or more disabilities, or a general reduction of their functionalities), and 1 was a "disabled older adult" (with long-term disabilities).

3.2 Material

The prototype used for our data collection (see Figure 1) consisted of a functioning camera attached

to an Android smartphone (for visually impaired participants) and to a tablet PC (for elderly participants). The choice of using different devices for the two user groups was informed by a) the need to develop and test EDC technology on different portable devices, and b) users' preferences that emerged during a set of explorative interviews that were conducted in a previous phase of the project.

The mobile device was running an application that showed video descriptions of the multimodal gestures to be performed. As an alternative, the visually impaired participants could choose an audio description of the gestures, if their visual deficiency was very severe or they simply preferred audio to video.

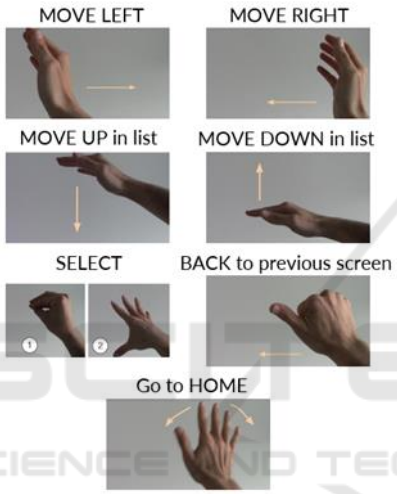

Figure 2: Basic interaction commands designed within the ECOMODE project.

Seven basic interaction commands (see Figure 2) designed within the ECOMODE project were collected.

A computer running the Mobizen mirroring application (https://www.mobizen.com/) was used to control the participant's device from the experimenter's notebook PC.

3.3 Procedure

On arrival, the experimenter explained to the participant the purpose and duration of the task (~60 minutes for the visually impaired participants and ~30 minutes for the elderly participants – but in addition, the latter ones collected 100 examples of speech only commands, for a total duration of about 1 hour and a half). The participant signed a consent form and then the device equipped with the ECOMODE camera was presented.

The participant was instructed about the distance to hold the device from the lips (about 30 cm), about the camera orientation, and about the distance of the gesture from the camera, in order to favour an effective capture of gesture and speech.

The experimenter then controlled the device to allow the participant to either watch or hear the description of the first gesture, as many times as they liked. When the specific mid-air gesture and voice command to perform were clear to the participant, the experimenter remotely controlled the recording application running on the participant's device to start and end the recording. This process was repeated until each of the seven multimodal gestures was recorded ten times for the visually impaired participants and four times for the elderly ones.

Figure 3: Two participants of the user study.

The recordings were performed in a relatively controlled environment: all were performed against a blank wall to reduce visual noise. In order to have a certain variability, useful for the automatic recognition purposes, the visually impaired participants were asked to accomplish the task sitting on a stool (Figure 3, left), while the elderly participants were standing (Figure 3, right), unless they had physical problems (e.g., back or leg pain).

In order to investigate memorability issues, at the end of all the recording sessions, which lasted about one hour and a half, the elderly participants were asked to recall the mid-air gestures performed at the beginning. Half of the group were asked to recall the mid-air gesture from the voice command (Test 1), whereas inversely the other half were asked to recall the voice command, given the corresponding mid-air gesture (Test 2).

For the visually impaired participants, the memorability test was done a week later by sending them an email containing a list naming the seven gestures performed during the data collection and asking them to reply with a description of each gesture.

4 OBSERVATIONS AND PRELIMINARY FINDINGS

During the data collection, the experimenters took note of important interaction issues and of participants' comments. After the recording sessions, the experimenters' observations on users' engagement with the device were discussed and categorised into themes as reported in the two sections below.

4.1 Visually Impaired Users

- *Distance of the device from lips and hand detection issues*. The majority of participants (76%) would naturally hold their phone outside of the current ideal distance for gesture and lip movement detection (30 cm). Some tended to hold it closer (better for lip movement detection) and some further away (better for mid-air gesture identification). Two participants did not always hold their phone but placed it on a surface or in their pocket. This has implications for the optics to be used with the camera, and advocates for the need to design and implement appropriate feedback and feedforward to guide the user towards the optimal detection zone for speech and gesture recognition.
- *Point of reference/orientation issues*. These were observed either due to a poor orientation of the hand during the mid-air gesture, not properly centring the gesture in the camera's field of view, not being at the right distance from the camera or the camera not being oriented directly towards themselves (pointing to the side).
- *Compound gestures*. Gestures that require several consecutive movements (e.g., closing fingers of the hand then moving hand to the left - see 'Back' in Figure 2) are considered *complex* by users, and require dynamic continual feedforward (such as the one used by Bau and Mackay, 2008). Hand specific gestures (e.g., thumb pointing to the left) should also be avoided.
- *Vertical VS horizontal swipe preference variability*. The preference for vertical or horizontal swipes for navigation (Up/Down and Left/Right) is quite variable and even the direction of navigation for each swipe can be interpreted differently (inversed especially for left handers).
- *Hand VS finger gesture*s. As opposed to the results of the previous explorative interviews, hand mid-air gestures (65%) were generally preferred over finger gestures (29%), with one subject indifferent. Even though they need to be limited in amplitude, hand gestures could be easier to perform than finger gestures, which could pose more inter-finger and intra-finger constraints (Kortum, 2008). In addition, users felts that hand gestures would be more robust and easier to recognise than finger gestures, and so produce less errors. However, this preference should be re-tested in the future in more 'realistic' contexts.
- *Personalisation of the gesture set*. During the interaction with the device, users expressed the desire to personalise the mid-air gestures (although this would require a system able to record and learn mid-air gestures).
- *Recall issues.* Even though most participants felt the gestures would be easy to remember (88%), only two of the seven gestures (Home and Select) were actually reasonably correctly remembered after one week (see Figure 4). Metaphoric gestures, which are meaningful to the user because they exploit primary metaphors to connect gestures to abstract interactive content, should be preferred over semaphoric (symbolic) gestures (Hurtienne et al., 2010; Saffer, 2008).
- *Feedback on the system status*. As this was a data collection aimed at building the dataset to be used for training the recognition algorithms, no feedback was built into the system. Despite this, observations and participants' comments highlighted the need for timely and useful feedback of mid-air gesture recognition (currently the feedback was given by the experimenter).
- *General usefulness for visually impaired users*. Doubts were raised about the usefulness of mid-air gesture interaction for visually impaired people. For instance, participants said that mid-air gestures are "more adapted to the elderly than visually impaired", and "visually impaired people are used to touching things, so mid-air gestures are not natural", "more useful for tablets or the television". A further investigation on the actual use of mobile technology among visually impaired people could help to identify to what extent and in what contexts mid-air gesture interaction could improve accessibility for these target users.

These preliminary data and the small number of participants do not allow us to investigate in greater

detail the subtlest differences between the partially sighted and blind groups, but after an initial analysis of the data it can be noted that:

- Handedness could have an impact for both groups, both in terms of mid-air gesture direction and system interpretation.
- No immediate difference is evident between visually impaired and blind participants in terms of the perceived complexity or memorability of the mid-air gestures. However, it can be noted that all the blind participants stated a preference for using hand rather than finger mid-air gestures.

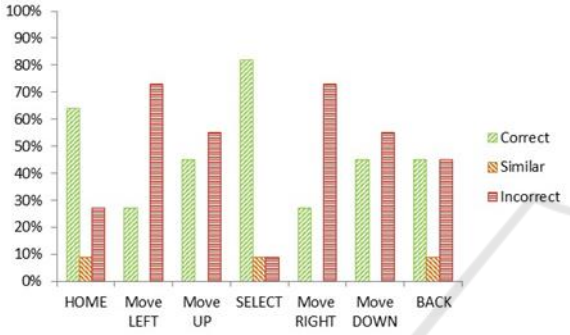

Figure 4: Visually impaired users. Interaction gesture recall (from voice command to mid-air gesture), function by function, after one week.

4.2 Elderly Users

- *Distance of the device from lips and hand detection issues*. These issues are similar to those found with visually impaired users. Although the elderly participants were instructed to hold the tablet PC at about 30 cm from the face, most of them (13 out of 17 – three participants carried out the task sitting on a chair with the tablet PC placed on a table, due to physical problems) tended to hold it farther away (about 40-45 cm) to have space for performing the hand gesture. Moreover, the majority of participants (65%) performed the gestures too close to the camera to be appropriately recorded.
- *Point of reference/orientation issues*. About 60% of the elderly participants often performed the gestures partially out of the camera field of view. Again, this issue and the previous one should guide the choice of the optics, and highlight the need to include appropriate feedback and feedforward.
- *Variability of gesture performance*. No gesture was felt complex to be performed by the participants, but a certain variability (in particular different tablet orientation and wider gesture amplitude) has been observed between subjects.
- *Co-occurrence of gesture and speech command*. A certain number of participants (25%) showed difficulties in performing gestures concurrently with voice commands. Indeed, most of them tended to perform gestures before the speech commands.
- *Grip issues*. Some users complained about the difficulty of holding the tablet without touching the screen, or being afraid of dropping it. A grip on the side of the tablet, or a belt on the back to insert their left hand, were suggested.

From the results of the memorability test, it is evident that for the elderly users it is overall easier to recall a voice command given the mid-air gesture, rather that recalling a gesture from the associated voice command (Figure 5). In particular, three of the seven gestures (Move left, Move up and Move right) were correctly remembered more often than the others, even though this percentage was quite low, i.e. around 30% (Figure 6).

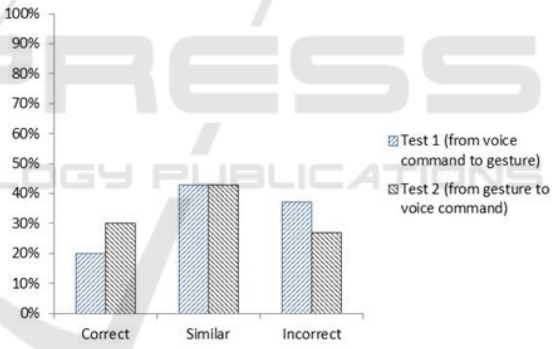

Figure 5: Elderly users. Results of Memorability Test 1 (recall a gesture from the voice command) and Test 2 (recall a voice command from the gesture).

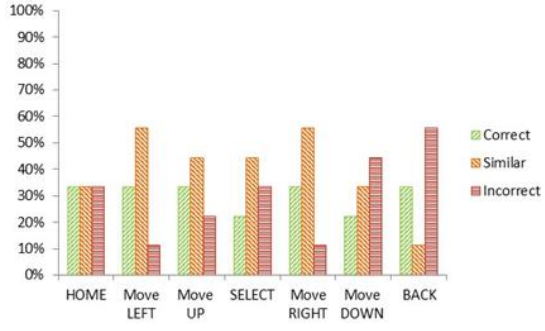

Figure 6: Elderly users. Results of Memorability Test 2 (recall a voice command from the gesture), function by function, after one hour and a half but having performed one hundred other speech only commands.

5 INTERACTION DESIGN ISSUES AND CHALLENGES

Although these initial studies were performed on a limited interactive prototype, various issues and challenges have been brought to the fore.

First, feedback needs to be provided to users to allow them to know that the mid-air gesture has been correctly recognised. Feedback on proper orientation of the device, or positioning of their hand in front of it, would also be useful. Several means exist for providing this feedback either individually or simultaneously in the tactile, visual or auditory modalities, directly on the smartphone or tablet PC (e.g., Oh et al., 2015; Wensveen et al., 2004).

Participants should also be supported in performing correctly the interactive command by means of a suitable feedforward. Unfortunately, few research studies have reported on how to create this. Whilst, using the classification provided by Bau and Mackay (2008), our current feedforward can be classified with a level of detail as whole gesture and an update rate of once prior to execution, there is an implication that continuous feedforward would be useful during the entire interaction. On top of this, there are currently no examples of feedforward in the aural modality and only one in the tactile modality (Vermeulen et al., 2013). This means that, specifically for blind users, some novel feedforward mechanisms will need to be identified and defined since the visual modality cannot be used. The modalities used will also have to be relevant and accessible to our end user groups and potentially take note of restrictions in human information processing resources in terms of conflicts or complementarity as outlined by, for instance, Wickens' Human Information Processing Model (Wickens et al., 2015).

Another issue is that concerning the reduction of false positives in the mid-air gesture recognition process. This issue needs to be addressed, either through robust detection algorithms or the use of physical actions to start and stop detection (clutch) (Chen et al., 2014; Wigdor and Wixon, 2011).

Finally, reference points are needed, especially for visually impaired users, for remaining inside the camera's field of view and ensuring the correct distance and orientation. This implies an accurate choice of the camera optics, which should ensure a wide enough field of view to capture the mid-air gesture whilst having enough detail to identify lip movements. It also implies providing timely and useful feedback to the end users to help them perform the gesture in the camera's field of view.

6 CONCLUSIONS

This paper reported on two early user studies aimed at investigating issues concerning the design of multimodal interaction - based on speech commands and mid-air gestures - with mobile technology specifically designed for visually impaired and elderly people.

This type of multimodal interaction, which enables a natural and efficient way of human-computer interaction with mobile devices, is investigated in the ECOMODE project, which aims at exploiting the biological-inspired technique of EDC of audio-visual information.

Seventeen visually impaired people and twenty older adults were involved in our study, which consisted of collecting and analysing a set of audio-visual data, which were recordings of our users when performing a sequence of seven multimodal commands.

The data was collected by means of a special camera, a prototype developed inside the ECOMODE project. The camera was externally attached to the used mobile device - a smartphone for visually impaired people and a tablet PC for older adults. The mobile devices were running an application that showed the video or audio descriptions of the multimodal (speech and mid-air) commands to be performed. At the end of the recording session, users were also invited to perform a memorability test.

Several design issues emerged from the analysis of experimenters' observations and users' comments, concerning for example how visually impaired people and older adults hold the mobile device, which was mostly held too far away, or concerning the fact that mid-air gestures were often performed too close to the camera.

The main design challenge concerns the necessity of providing effective feedback and feedforward to the users, to allow them to know if their commands have been correctly recognised. Users should also be informed if they are correctly holding the mobile device, which must be held so as to favour the gesture being performed in the camera's field of view. The reduction of false positives in the mid-air gesture recognition process also needs to be addressed.

All the emerging issues and challenges will be addressed by the ECOMODE project in the near future, so as to arrive at a fully interactive mobile device, incorporating automatic mid-air gesture and speech recognition and using an application with some basic functionality (camera, messaging,

contacts, etc.) that can be tested in more realistic environments and settings.

ACKNOWLEDGEMENTS

This work was supported by the Horizon 2020 Project ECOMODE "Event-Driven Compressive Vision for Multimodal Interaction with Mobile Devices" (Grant Agreement No. 644096).

REFERENCES

Bau, O., & Mackay, W. E. (2008). OctoPocus: a dynamic guide for learning gesture-based command sets. In Proceedings of the 21st annual ACM symposium on User interface software and technology (pp. 37-46). ACM.

Chen, X. A., Schwarz, J., Harrison, C., Mankoff, & J., Hudson, S. E. (2014). Air+ touch: interweaving touch & in-air gestures. In Proceedings of the 27th annual ACM symposium on User interface software and technology (pp. 519-525). ACM.Clady, X., Maro, J. M., Barré, S., & Benosman, R. B. (2017). A motion-based feature for event-based pattern recognition. Frontiers in neuroscience, 10, 594.

De Carvalho Correia, A. C., Cunha de Miranda, L. and Hornung, H. (2013). Gesture-Based Interaction in Domotic Environments: State of the Art and HCI Framework Inspired by the Diversity. In Human-Computer Interaction – INTERACT 2013 (Lecture Notes in Computer Science), 300–317.

Fisk, A. D., Rogers, W. A., Charness, N., Czaja, S. J., & Sharit, J. (2009). Designing for older adults: Principles and creative human factors approaches: CRC press.

Grandhi, S. A., Joue, G., & Mittelberg, I. (2011). Understanding naturalness and intuitiveness in gesture production: insights for touchless gestural interfaces. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Gregor, P., Newell, A. F., & Zajicek, M. (2002). Designing for dynamic diversity: interfaces for older people. Paper presented at the Proceedings of the fifth international ACM conference on Assistive technologies.

Hurtienne, J., Stößel, C., Sturm, C., Maus, A., Rötting, M., Langdon, P., & Clarkson, J. (2010). Physical gestures for abstract concepts: Inclusive design with primary metaphors. *Interacting with Computers, 22*(6), 475-484.

Kortum, P. (2008). HCI beyond the GUI: Design for haptic, speech, olfactory, and other nontraditional interfaces: Morgan Kaufmann.

Lindsay, S., Jackson, D., Schofield, G., & Olivier, P. (2012). Engaging older people using participatory design. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Oh, U., Branham, S., Findlater, L., & Kane, S. K. (2015). Audio-Based Feedback Techniques for Teaching Touchscreen Gestures. ACM Transactions on Accessible Computing (TACCESS), 7(3), 9.

Ryan, E. B., Szechtman, B., & Bodkin, J. (1992). Attitudes toward younger and older adults learning to use computers. Journal of gerontology, 47(2), P96-P101.

Saffer, D. (2008). Designing gestural interfaces: Touchscreens and interactive devices. O'Reilly Media, Inc.

Vermeulen, J., Luyten, K., van den Hoven, E., & Coninx, K. (2013). Crossing the bridge over Norman's Gulf of Execution: revealing feedforward's true identity. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1931-1940). ACM.

Wensveen, S. A., Djajadiningrat, J. P., & Overbeeke, C. J. (2004)). Interaction frogger: a design framework to couple action and function through feedback and feedforward. In Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques (pp. 177-184), ACM.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). Engineering psychology & human performance. Psychology Press.

Wigdor, D., & Wixon, D. (2011). Touch versus in-air gestures. In Brave NUI World, chapter 15 (pp. 97-103).

World Health Organization. (2004). ICD-10: international statistical classification of diseases and related health problems: tenth revision.

World Health Organization. (2012). Global data on visual impairments 2010. Geneva: World Health Organization.