Designing Mobile Multimodal Interaction for Visually Impaired and Older Adults: Challenges and Possible Solutions

Michela Ferron*†, Nadia Mana*, Ornella Mich*, Christopher Reeves‡, Gianluca Schiavo*

*Fondazione Bruno Kessler
Via Sommarive 18, Trento, Italy
email: {mana,mich,schiavo}@fbk.eu
www: fbk.eu

‡Streetlab
Rue Moreau 17, Paris, France
email: christopher.reeves@streetlab-vision.com
www: streetlab-vision.com

†now at Dedagroup Public Services
Via di Spini 50, Trento, Italy
email: michela.ferron@dedagroup.it
www: dedagroup.it

Abstract

This paper presents two early studies aimed at investigating issues concerning the design of multimodal interaction - based on voice commands and one-hand mid-air gestures - with mobile technology specifically designed for visually impaired and elderly users. These studies were carried out on a new device allowing enhanced speech recognition (including lip movement analysis) and mid-air gesture interaction on Android operating system (smartphone and tablet PC). We discuss the initial findings and challenges raised by these novel interaction modalities, and in particular the issues regarding the design of feedback and feedforward, the problem of false positives, and the correct orientation and distance of the hand and the device during the interaction. Finally, we present a set of feedback and feedforward solutions designed to overcome the main issues highlighted.

Digital Peer Publishing Licence

Any party may pass on this Work by electronic means and make it available for download under the terms and conditions of the current version of the Digital Peer Publishing Licence (DPPL). The text of the licence may be accessed and retrieved via Internet at http://www.dipp.nrw.de/.

First presented at the 3rd International Conference on Human-Computer Interaction Theory and Applications (HUCAPP) 2018, extended and revised for JVRB

Keywords: Visually Impaired, Older Adults, Mid-Air Gestures, Multimodal, Smartphone, Tablet PC, Feedback, Feedforward.

1 Introduction

Accessing Information and Communication Technologies (ICT) is essential to participate in a modern, interconnected society that relies on technology for handling everyday tasks. Although very different user groups in terms of needs, desires and requirements, both the visually impaired and elderly people face significant barriers in accessing ICT with traditional interaction modalities (mouse + keyboard), because of their fragilities and/or disabilities (e.g. perception, cognition and movement control).

Older adults face several unique age-related constraints, such as sensory, motor and cognitive impairments and other chronological diseases. This age-related functional decline poses obstacles in the pro-

cess of interacting with computer-based technology. On the one hand, motor and dexterity impairments can make it difficult to interact using traditional mouse, as studies demonstrate: older adults are slower in pointing and dragging tasks when compared to younger adults [CBF+99] and have more difficulties clicking and double-clicking [SSC99]. On the other hand, the decline in computer-related sensing, cognitive and responding abilities, which have been long studied (see for example [Sal16, Wel81]), can make older adults feel confused and frustrated by the unfriendliness and poor readability of interface design. Older adults are frequently portrayed as generally resistant to technology [RSB92, MBF+10, HMM+13], but a substantial amount of studies have shown 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 [CL17, CRF+09, LJSO12, Sel04, SMFM17, VCG17].

According to the World Health Organization (WHO), the worldwide visually impaired population has been estimated at 285 million people of whom 39 million are blind and 246 million have low vision [PM12]. This group is very diverse but common major factors affecting visual impairment include age, gender and socioeconomic status. For this reason, it is important to highlight that, since the large majority of visually impaired people are over 50 years old, they are exposed to a certain extent to the same issues that were illustrated above regarding the elderly. Nonetheless, 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.

Voice commands and non-touchscreen-based hand and finger gestures might make human-computer interaction easier and more natural for these user groups [CTCG17, dCCdMH13]. 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 [GJM11] 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 [MMF17] 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 two early 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. Moreover, in analysing the interaction we focus on feedback and feedforward mechanisms, two important concepts in interaction design [VLvdHC13]. Feedback is the interaction mechanism that communicates the results of the interaction, making it visible and understandable to the user. The concept of feedforward refers to any interaction mechanism that tells users what the result of their action will be before the action is fully performed. In the context of gestural interfaces, through feedback users receive information about the effectiveness of their gesture, while feedforward helps users in performing the correct gesture by suggesting to them the range of possible gestures and what will happen when a certain gesture is invoked. The notion of feedforward differs indeed from that of feedback since the latter occurs during or after the user's action and provides information on the result of the action itself. The results of our studies contribute to the wider understanding of feedback and feedforward mechanism for multimodal interaction.

2 The ECOMODE technology

The ECOMODE project[1] aimed at exploiting the recently matured biologically inspired technique of event driven compressive sensing (EDC) of audio-visual information [CMZRLB+12], for enabling more natural and efficient ways of human interaction with ICT [CTCG17].

While traditional techniques are based on frame-by-frame processes, and therefore generally slow, computationally expensive and include latencies of sensing, transmitting and processing, EDC technology is frame-free and therefore faster and more precise (see [BCL+13, DLBCP10, PCZS+13] for more technical details).

Furthermore, such technology is able to support device operation under uncontrolled conditions, where devices based on traditional techniques typically fail. In particular, when EDC technology is applied to gesture and speech recognition, use contexts with bad lighting (poor or excessive) and high background noise no longer present obstacles. In the case of traditional technology gesture recognition cannot cope with the high intra-scene dynamic range given by the high levels of infrared background from natural (sun) light, and conventional speech recognition becomes unfeasible due to high levels of background noise.

A further advantage of EDC technology lies in the sparse nature of its information encoding that makes EDC excel over conventional approaches in terms of energy efficiency, yielding an ideal solution for mobile, battery-powered devices. Furthermore, event-based cameras rely on a new principle that naturally allows all of the information contained in a standard video stream of several megabytes to be compressed in an event stream of a few kilobytes.

If compared to frame-drives approaches, one of the main advantages of EDC technology is real-time recognition that uses only the computational power of a mobile phone, whereas frame-based image analysis requires off-chip resources. Dynamic background suppression can also be introduced to help achieve high recognition rates in walking situations [MB18]. Finally, this technology overcomes the issues related to glare and motion blur that can occur in frame-based techniques.

One of the main challenges of the ECOMODE project was to integrate different EDC technology hardware components, and to combine them into battery-powered mobile devices, such as tablet computers and smartphones, to provide an application based on a touchless multimodal interaction, i.e. combining gestures and voice commands.

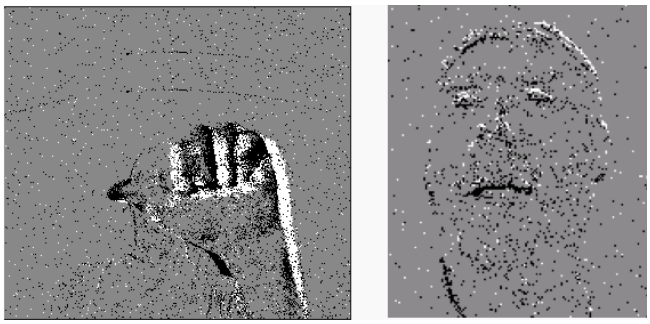

Figure 1: *Back* command gesture and visual speech, captured by the ECOMODE camera

In particular, the ECOMODE technology exploits a mid-air gesture control set processing for hand and finger gesture recognition [CMBB17], and a vision-assisted speech recognition set that combines auditory input [BCFM16] with visual information from lip and chin motion [GB17, STH+18], in order to gain robustness and background noise immunity in the recognition of spoken commands and speech-to-text input (Figure 1).

Figure 2: ECOMODE prototype running on smartphone and tablet

These characteristics promise to make the ECOMODE prototype (Figure 2) work more efficiently in real world contexts and in more challenging environments.

[1] http://www.ecomode-project.eu/

3 User studies

Data collection aiming at building a dataset of mid-air gestures and voice commands to be used to train and test the recognition algorithms was organized within the ECOMODE project by involving a panel of target users. While carrying out the data collection, the experimental sessions with the visually impaired and elderly 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 [Org11]. For our studies, we referred to the WHO categorisation [Org04] 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 [CRF+09], 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 psycho-physical conditions [GNZ02], 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 2) 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

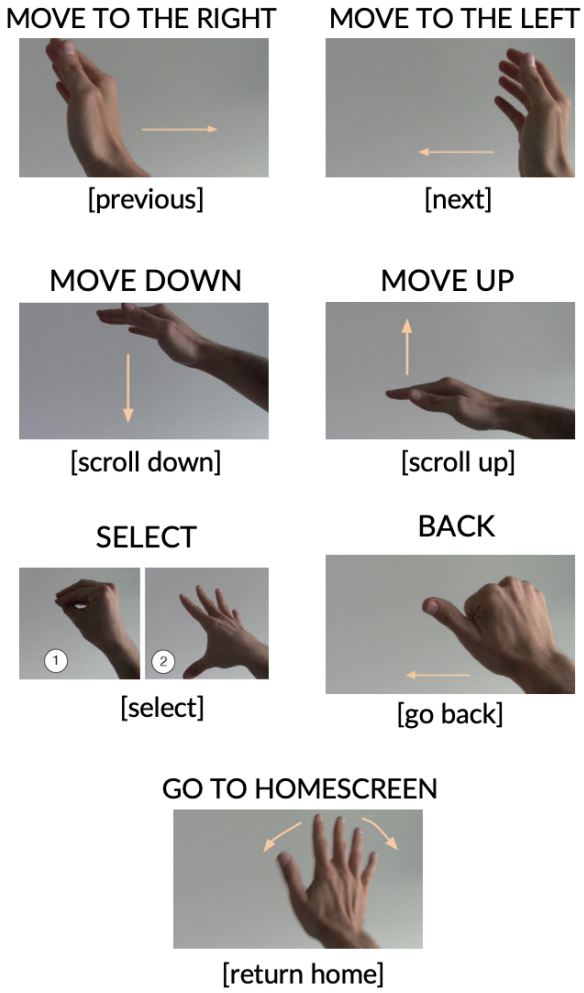

Figure 3: Basic multimodal interaction commands - mid-air gesture and voice input [in squared brackets]

EDC technology on different portable devices, and b) users' preferences that emerged from the exploratory user research that was conducted in a previous phase of the project. For older adults, we interviewed two experts with experience working with older adults associations, and conducted an exploratory study observing six older users taking pictures with a smartphone or a tablet device [FMM15, FMM19]. Experts highlighted that older adults, although not necessarily familiar with tablet technology, were eager to learn to use it and attended with interest the courses held by the associations, regarding the tablet as a means of entertainment and fun. On the other hand, older adults appreciated the larger screen of the tablet, which offered higher visibility with respect to

the smaller screen of mobile phones. As for visually impaired users, to identify current technology use we conducted an initial survey with 169 respondents. Of these, 82% had used a smartphone and 60% had used a tablet. When asked which of the two they used most often, 64% replied that they used their smartphone more, with 15% using their smartphone and tablet equally and only 8% using their tablet more often (the remaining percentages were made up of those who used a computer or other device more often). These results led to the decision to use smartphone devices for visually impaired users.

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 de ciency was very severe or they simply preferred audio to video.

Seven basic multimodal interaction commands (Move to the left, Move to the right, Move down, Move up, Select, Back and Go to homescreen - see Figure 3) designed within the ECOMODE project were collected. The mid-air gestures and the corresponding voice commands (which are reported in squared brackets in Figure 3) were intended to be performed simultaneously. A computer running the Mobizen[2] mirroring application was used to control the participant's device from the experimenter's notebook PC.

3.3 Task description

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 one 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 30cm), 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 rst gesture, as many times as they liked. When the speci c 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. The different number of repetitions in the two user groups is justi ed by the need to collect an additional dataset of voice commands performed by older adults, large enough to train robust machine learning algorithms. Indeed, machine learning algorithms for speech recognition are commonly trained on datasets made of younger adult voices. However, ageing brings signi cant changes to the human voice [Kel06], which can include hoarseness and differences in articulation patterns, resulting in a rougher and breathier vocal quality, and making common speech recognition algorithm ineffective. For this reason, only for older adults we collected 100 instances of speech commands for each participant (for a total of 2000 examples). Since a longer data collection task would have been physically strenuous for seniors, we reduced the number of repetitions of each multimodal gesture from ten to four.

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 4, left), while the elderly participants were standing (Figure 4, 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

Figure 4: Two participants during the user studies

[2] https://www.mobizen.com

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, who were of a younger age and less likely to have cognitive issues, to make the task slightly harder 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 nd-ings

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 identi cation). 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 eld 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 ngers of the hand then moving hand to the left - see 'Back' in Figure 3) are considered complex by users, and require dynamic continual feedforward (such as the one used by [BM]). Hand speci c 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 nger gestures. As opposed to the results of the previous explorative interviews, hand mid-air gestures (65%) were generally preferred over nger gestures (29%), with one subject indifferent. Even though they need to be limited in amplitude, hand gestures could be easier to perform than nger gestures, which could pose more inter- nger and intra- nger constraints [Kor08]. In addition, users felts that hand gestures would be more robust and easier to recognise than nger 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 (Go to homescreen and Select) were actually reasonably correctly remembered after one week (Figure 5). 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 [HSS+10, Saf08].

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,

Figure 5: Visually impaired users. Interaction gesture recall (from voice command to mid-air gesture), function by function, after one week

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', 'they could be 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 nger mid-air gestures.

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 30cm 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 eld 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 dif culties 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 dif culty 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 6). In particular, three of the seven gestures (Go to Home, Move left and Move up) were correctly remembered more often than the others, even