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.

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 process 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.

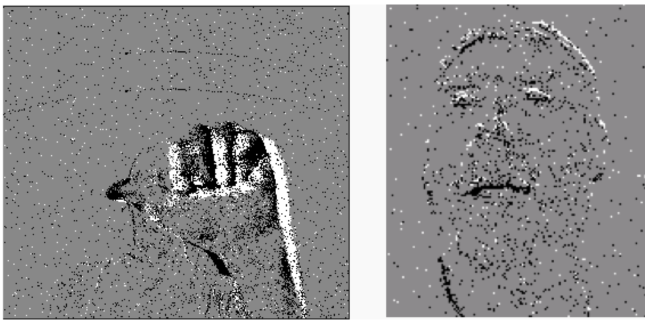


Figure 1: *Back* command gesture and visual speech, captured by the ECOMODE camera

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).

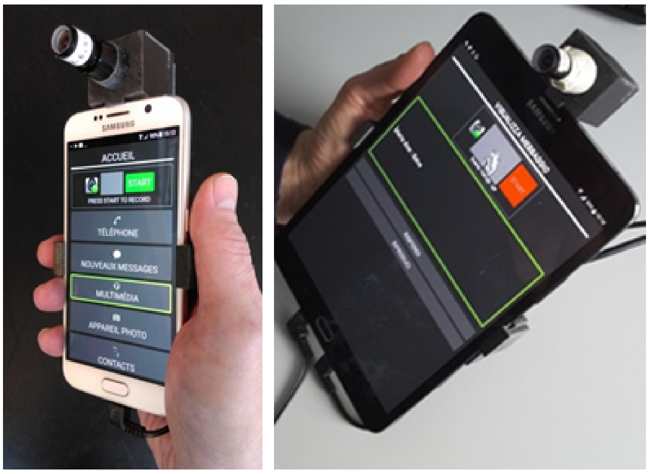


Figure 2: ECOMODE prototype running on smartphone and tablet

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 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

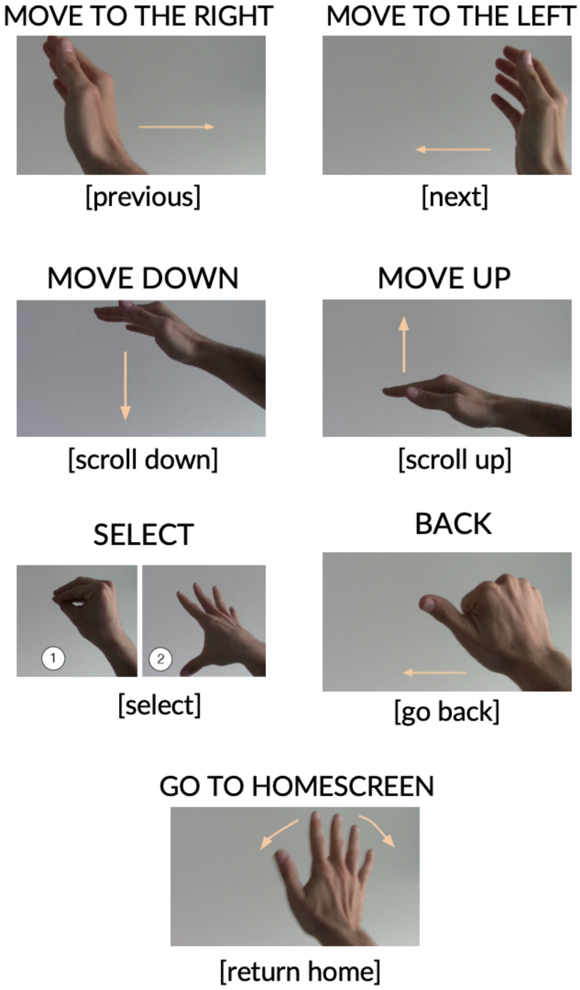


Figure 3: Basic multimodal interaction commands - mid-air gesture and voice input [in squared brackets]

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 deficiency 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 ($\sim$ 60 minutes for the visually impaired participants and $\sim$ 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 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. The different number of repetitions in the two user groups is justified 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 significant 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

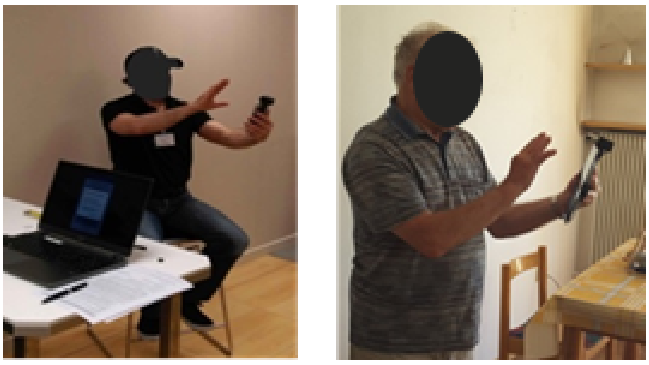


Figure 4: Two participants during the user studies

[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 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 3) are considered complex by users, and require dynamic continual feedforward (such as the one used by [BM]). 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 gestures.* 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 [Kor08]. 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 (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,

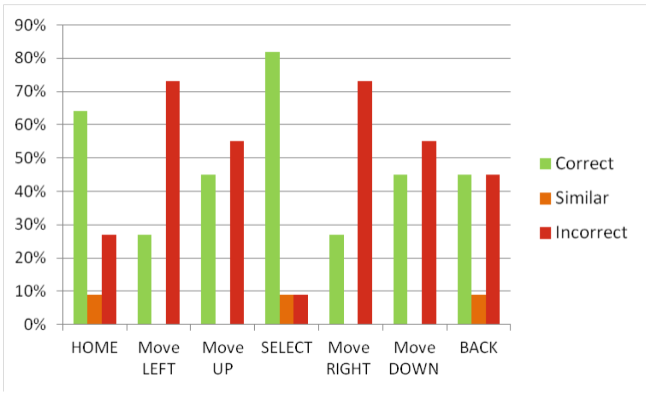


Figure 5: Visually impaired users. Interaction gesture recall (from voice command to mid-air gesture), function by function, after one week

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 finger 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 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 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

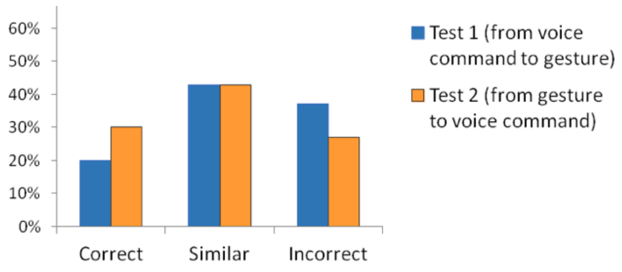


Figure 6: Elderly users. Results of Memorability Test 1 (recall a gesture from the voice command) and Test 2 (recall a voice command from the gesture)

though this percentage was quite low, i.e. around 30% (Figure 7). We suggest that these gestures were better recalled since they refer to meaningful metaphors of how people interact with the world [HSS+10], with respect to other commands.

Finally, comparing the performance of visually impaired and elderly users in the memorability test, we observed that on average, and despite performing the test a week after the data collection, visually impaired users could remember the gesture with higher accuracy (50% of correctly recalled commands, Figure 5) compared to elderly participants (20% of correct answers, Figure 7). This suggests that, when designing multimodal interaction for older adults, implementation efforts should be directed towards (a) more robust recognition algorithms that are sensitive to significant variability in gesture performance, (b) a clear and visible feedforward mechanism and (c) easily accessible training sessions to help users to familiarize with the interaction.

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., [OBFK15, WDO04]).

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 [BM], 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 [VLvdHC13]. 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. Within the time frame and scope of this work, complete aural or tactile feedforward mechanisms could not be developed, but ideas for feedback and feedforward for the visually impaired are provided in the section below. 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 [WHBP15].

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) ([CSH+14, WW11].

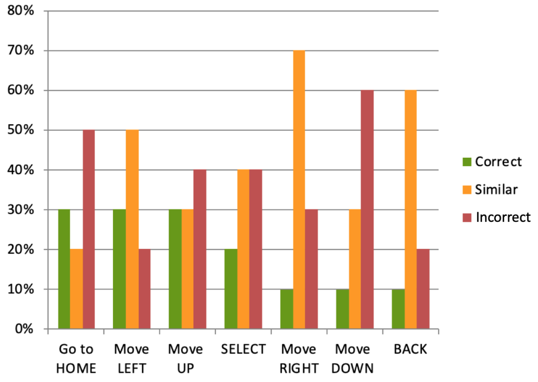


Figure 7: Elderly users. Results of Memorability Test 1 (recall a gesture from the voice command), function by function, after one hour and a half but having performed one hundred other speech only commands in the meantime

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 Designing feedback and feedforward

Starting from the challenges identified in the analysis of users' interaction with the system during the data collection and discussed in Section 5, we organized two brainstorming sessions to address the design of feedback and feedforward mechanisms.

The brainstorming sessions were organized at FBK and Streetlab with expert users and UX researchers, with the aim of collecting ideas and proposals for feedforward and feedback mechanisms for the ECOMODE application. The brainstorming sessions included three phases (Figure 8):

Team Brainstorming (15 minutes): the team of participants was divided into pairs to generate as many ideas as possible within a given period of time (10 minutes). When the time was up, each pair presented their ideas through a post-it summary.

Post-it vote (5 minutes): each participant made a mark on their favourite idea represented on a post-it (3 votes per person)

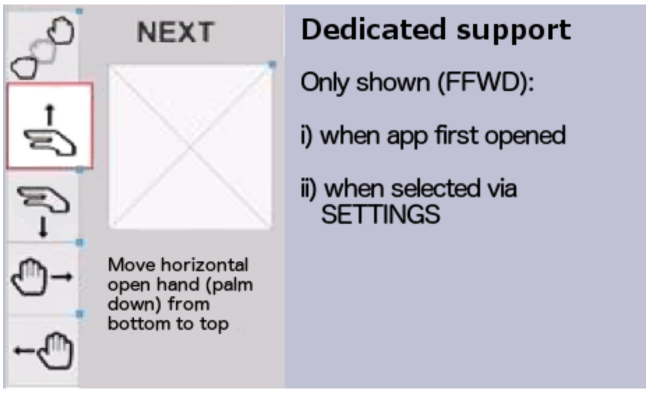
Brainwriting (15 minutes): the ideas with the most votes were taken and given one to each participant. Over the next five minutes, each person read the original suggestion and generated three additional ideas based on the original suggestion. Once completed, the enhanced ideas were passed to another colleague a further 3 to 5 times. After completing the cycle, each participant briefly presented the extra thoughts on the post-it idea in front of them.

At FBK, 8 experts in HCI, UX and Computer Science participated in the team brainstorming, generating a total of 22 ideas, 6 of which were further explored during the brainwriting session. Similarly, at Streetlab 8 experts in Ergonomics and Accessibility participated in the team brainstorming, generating a total of 31 ideas, 6 of which were further explored during the brainwriting session.

During a post analysis stage, each idea was included in one of seven categories based on their function (Onscreen support, Modality, Tutorial/assistant, Trigger, Recognition, Accuracy and Augmentation). Similarly the type of modality used was noted (Visual, Audio or Tactile / Vibration). Post-its with similar concepts and design ideas were merged in a single proposal.

Figure 8: Different moments of the brainstorming session, and close-up of a voted idea

Following the brainstorming results, three main types of support were explored which make use of feedback, feedforward or both:



Dedicated support

Only shown (FFWD):

i) when app first opened

ii) when selected via SETTINGS

Move horizontal open hand (palm down) from bottom to top

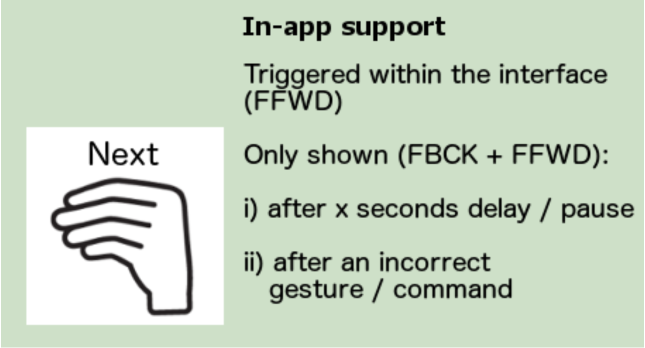
Figure 9: Example of dedicated support

1. Dedicated support is provided through an information panel/drawer that summarises indications on how to perform the mid-air gesture and vocal commands (Figure 9). This panel, providing feedforward, is accessible the first time the application is started and can also be selected via the app's settings. The panel displays information on all the commands in an extensive way, including a video and a description of how to perform the gesture in audio format for the visually impaired. This type of dedicated support does however raises two questions: (a) How to navigate the list if the user is unsure of the relevant air gesture / voice command? (b) How to open and close the support, e.g., through an on screen icon, a swipe gesture, a menu?

2. In-app support, provided directly in the application, includes a briefer version of the command panel, giving indication on how to perform the gesture when the interaction is unsuccessful or the command is not recognized by the system (Figure 10). For this kind of support it is indeed crucial to create an intuitive and familiar representation of the different gestures in order to support the user in remembering how to correctly perform the command. The use of detailed icons symbolizing the different gestures can make it easier for users of all ages, but especially for older adults, to learn what the icons mean. Instead of using abstract or arbitrary icons, it would be preferable to use pictograms that clearly depict the hand and the gesture.

   The command panel can also be displayed when the system detects user inactivity within a certain number of seconds. After the user performs a command, the system should implement the corresponding action immediately and return a visual or audio feedback within the interface.

   A further advantage of this type of in-app support is that, unlike the dedicated support, it could only show the possible or useful air-gestures and voice commands available at the current position rather than the full range of all that is available. This would make the content more relevant and targeted to the specific context.



In-app support

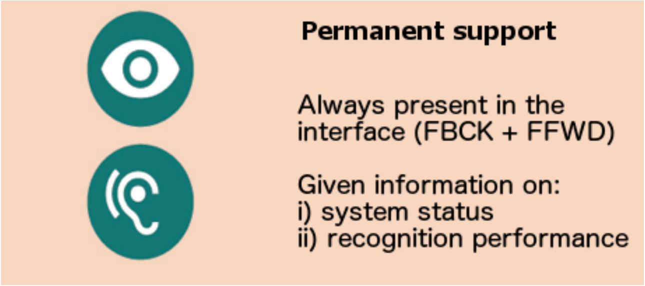
Triggered within the interface (FFWD)

Only shown (FBCK + FFWD):

i) after x seconds delay / pause

ii) after an incorrect gesture / command

Figure 10: Example of in-app support



Permanent support

Always present in the interface (FBCK + FFWD)

Given information on:
i) system status
ii) recognition performance

Figure 11: Example of permanent support

3. Permanent support is provided with graphical elements in the interface that give indications on the system status (whether the camera or the microphone are working properly) and information on the accuracy recognition. Different ways for communicating this information to the user, e.g. based on abstract representations or metaphoric and non-metaphoric animated icons, may be used.

   Permanent support should provide information in the event that one of the two modality channels is not available, due to technical issues or environmental conditions (Figure 11). The graphical elements should also provide an indication that the system is able to detect gestural and vocal inputs. This would also provide the users an immediate visual cue that the system is listening for commands that can be performed either using the gestures or the voice. As well as the visual cue, an audio cue has been proposed to indicate that detection is active. In addition, the gyroscope in smartphones can be exploited to provide audio feedforward to help realign the camera for ideal vision-assisted speech recognition, should visually impaired users inadvertently turn the camera too far away from themselves (audio to move the device left, right, up or down).

7 Discussion and conclusion

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, enabling a natural and efficient way of human-computer interaction with mobile devices, is investigated in the ECOMODE project. The technical goal of ECOMODE was to exploit the biological-inspired EDC technique to quickly and accurately process audio-visual information. In parallel, the project investigated multimodal interaction for the visually impaired and elderly users.
This paper reports on two user studies that involved seventeen visually impaired participants and twenty older adults for collecting and analysing a set of audio-visual data, i.e. recordings of our users while performing a sequence of seven multimodal commands. The data was collected by means of a event-based camera externally attached to a mobile device: a smartphone for visually impaired people and a tablet device for older adults. Participants were asked to perform several repetitions of a set of multimodal commands, while type of interactions and their memorability were analysed.

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

Among the emerged design issues, one of the main challenges concerns the necessity of providing effective feedback and feedforward to the users. Indeed, users should understand if their interaction commands have been correctly recognised. Moreover, users should also be informed about the system status and whether they are correctly holding the mobile device, which must be held so as to favour the gesture being performed in the cameras' field of view. Furthermore, users should also be guided in performing the correct mid-air gestures in the correct way.

In this work, we addressed the challenge of designing feedback and feedforward support by organizing two brainstorming sessions with UX researchers and expert users. The main ideas developed after the brainstorming resulted in three different types of support:

> (1) dedicated support, which can be accessed when the application runs for the first time or can be selected via the app settings, and provides the user extensive feedforward with a description, audio explanation and video of all the commands available;
> (2) in-app support, which is provided directly within the application in a persistent area and provides the user essential feedforward on how to perform the commands and feedback about the result of the interaction;
> (3) permanent support, which is implemented with graphical elements to inform about the system status (i.e., the correct functioning of the camera and the microphone) and information about the accuracy of the recognition. Additional audio elements provide help to visually impaired users so they know when the system is active and the camera properly orientated for optimal use.

The different types of support described above were designed to serve different purposes. While dedicated support could be particularly helpful for novice users to learn which commands are available, and disregarded by expert users, in-app support could provide a more unobtrusive guide, which can also be context-aware and provide the user selected information with respect to the commands available. Finally, permanent support is designed to inform the user about the status of the system.

To conclude, this paper presents the initial findings and challenges raised by multimodal interaction for visually impaired and elderly users. In particular, the user studies investigated the variability and memorability of mid-air gestures and speech multimodal commands. Moreover, the paper highlights the challenges in the design of feedback and feedforward mechanisms and proposes a set of solutions designed to overcome the main highlighted issues.

8 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

[BCFM16] Leonardo Badino, Claudia Canevari, Luciano Fadiga, and Giorgio Metta, *Integrating articulatory data in deep neural network-based acoustic modeling*, Computer Speech & Language **36** (2016), 173–195, DOI 10.1016/j.csl.2015.05.005.

[BCL+13] Ryad Benosman, Charles Clercq, Xavier Lagorce, Sio-Hoi Ieng, and Chiara Bartolozzi, *Event-based visual flow*, IEEE transactions on neural networks and learning systems **25** (2013), no. 2, 407–417, DOI 10.1109/TNNLS.2013.2273537.

[BM] Olivier Bau and Wendy E. Mackay, *OctoPocus: a dynamic guide for learning gesture-based command sets*, Proceedings of the 21st annual ACM symposium on User interface software and technology, pp. 37–46.

[CBF+99] Alex Chaparro, Michael Bohan, Jeffrey Fernandez, Sang D. Choi, and Bheem Kattel, *The impact of age on computer input device use:: Psychophysical and physiological measures*, International Journal of Industrial Ergonomics **24** (1999), no. 5, 503–513, DOI 10.1016/S0169-8141(98)00077-8.

[CL17] Ching-Ju Chiu and Chia-Wen Liu, *Understanding older adult's technology adoption and withdrawal for elderly care and education: mixed method analysis from national survey*, Journal of medical Internet research **19** (2017), no. 11, 374, DOI 10.2196/jmir.7401.

[CMBB17] Xavier Clady, Jean-Matthieu Maro, Sébastien Barré, and Ryad B. Benosman, *A Motion-Based Feature for Event-Based Pattern Recognition*, Frontiers in neuroscience **10** (2017), 594, DOI 10.3389/fnins.2016.00594.

[CMZRLB+12] Luis Camunas-Mesa, Carlos Zamarreno-Ramos, Alejandro Linares-Barranco, Antonio J. Acosta-Jimenez, Teresa Serrano-Gotarredona, and Bernab Bernabé Linares-Barranco, *An event-driven multi-kernel convolution processor module for event-driven vision sensors*, IEEE Journal of Solid-State Circuits **47** (2012), no. 2, 504–517, DOI 10.1109/JSSC.2011.2167409.

[CRF+09] Sara J. Czaja, Wendy A. Rogers, Arthur D. Fisk, Neil Charness, and Joseph Sharit, *Designing for older adults: Principles and creative human factors approaches*, CRC press, 2009, ISBN 9781138053663.

[CSH+14] Xiang Anthony Chen, Julia Schwarz, Chris Harrison, Jennifer Mankoff, and Scott E. Hudson, *Air+ touch: interweaving touch & in-air gestures*, Proceedings of the 27th annual ACM symposium on User interface software and technology, ACM, 2014, DOI 10.1145/2642918.2647392, pp. 519–525.

[CTCG17] Micael Carreira, Karine Lan Hing Ting, Petra Csobanka, and Daniel Gonçalves, *Evaluation of in-air hand gestures interaction for older people*, Universal Access in the Information Society **16** (2017), no. 3, 561–580, DOI 10.1007/s10209-016-0483-y.

[dCCdMH13] Ana Carla de Carvalho Correia, Leonardo Cunha de Miranda, and Heiko Hornung, *Gesture-based interaction in domotic environments: State of the art and HCI framework inspired by the diversity*, IFIP Conference on Human-Computer Interaction, Springer, 2013, DOI 10.1007/978-3-642-40480-1_19, pp. 300–317.

[DLBCP10] Tobi Delbrück, Bernabe Linares-Barranco, Eugenio Culurciello, and Christoph Posch, *Activity-driven, event-based vision sensors*, Proceedings of 2010 IEEE International

Symposium on Circuits and Systems, IEEE, 2010, DOI 10.1109/ISCAS.2010.5537149, pp. 2426–2429.

[FMM15] Michela Ferron, Nadia Mana, and Ornella Mich, *Mobile for older adults: towards designing multimodal interaction*, Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia, ACM, 2015, DOI 10.1145/2836041.2841211, pp. 373–378.

[FMM19] Michela Ferron, Ornella Mich, and Nadia Mana, *Designing Mid-Air Gesture Interaction with Mobile Devices for Older Adults*, Perspectives on Human-Computer Interaction Research with Older People (Sergio Sayago, ed.), Springer, 2019, DOI 10.1007/978-3-030-06076-3_6, pp. 81–100.

[GB17] Arren Glover and Chiara Bartolozzi, *Robust visual tracking with a freely-moving event camera*, Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on, IEEE, 2017, DOI 10.1109/IROS.2017.8206226, pp. 3769–3776.

[GJM11] Sukeshini A. Grandhi, Gina Joue, and Irene Mittelberg, *Understanding naturalness and intuitiveness in gesture production: Insights for touchless gestural interfaces*, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2011, DOI 10.1145/1978942.1979061, pp. 821–824.

[GNZ02] Peter Gregor, Alan F. Newell, and Maja Zajicek, *Designing for dynamic diversity: Interfaces for older people*, Proceedings of the fifth international ACM conference on Assistive technologies, ACM, 2002, DOI 10.1145/638249.638277, pp. 151–156.

[HMM+13] Melinda Heinz, Peter Martin, Jennifer A. Margrett, Mary Yearns, Warren Franke, Hen I. Yang, Johnny Wong, and Carl K. Chang, *Perceptions of technology among older adults*, Journal of Gerontological Nursing **39** (2013), no. 1, 42–51, DOI 10.3928/00989134-20121204-04.

[HSS+10] Jörn Hurtienne, Christian Stößel, Christine Sturm, Alexander Maus, Matthias Rötting, Patrick Langdon, and John Clarkson, *Physical gestures for abstract concepts: Inclusive design with primary metaphors*, Interacting with Computers **22** (2010), no. 6, 475–484, DOI 10.1016/j.intcom.2010.08.009.

[Kel06] Brigitte Zellner Keller, *Ageing and speech prosody*, Speech Prosody, 2006, pp. 696–700.

[Kor08] Philip Kortum, *HCI beyond the GUI: Design for haptic, speech, olfactory, and other nontraditional interfaces*, Elsevier, 2008, ISBN 9780123740175.

[LJSO12] Stephen Lindsay, Daniel Jackson, Guy Schofield, and Patrick Olivier, *Engaging older people using participatory design*, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2012, DOI 10.1145/2207676.2208570, pp. 1199–1208.

[MB18] Jean-Matthieu Maro and Ryad Benosman, *Event-based Gesture Recognition with Dynamic Background Suppression using Smartphone Computational Capabilities*, arXiv preprint (2018).

[MBF+10] Tracy L. Mitzner, Julie B. Boron, Cara Bailey Fausset, Anne E. Adams, Neil Charness, Sara J. Czaja, Katinka Dijkstra, Arthur D. Fisk, Wendy A. Rogers, and Joseph

Sharit, *Older adults talk technology: Technology usage and attitudes*, Computers in human behavior **26** (2010), no. 6, 1710–1721, DOI 10.1016/j.chb.2010.06.020.

[MMF17] Nadia Mana, Ornella Mich, and Michela Ferron, *How to increase older adults' accessibility to mobile technology? The new ECO-MODE camera*, ForItAAL-Forum Italiano Ambient Assisted Living, Springer, 2017, DOI 10.1007/978-3-030-04672-9_6.

[OBFK15] Uran Oh, Stacy Branham, Leah Findlater, and Shaun K. Kane, *Audio-based feedback techniques for teaching touchscreen gestures*, ACM Transactions on Accessible Computing (TACCESS) **7** (2015), no. 3, 9, DOI 10.1145/2764917.

[Org04] World Health Organization, *ICD-10: international statistical classification of diseases and related health problems: tenth revision*, 2004.

[Org11] ______, *Global data on visual impairments 2010*, 2011.

[PCZS+13] José Antonio Pérez-Carrasco, Bo Zhao, Carmen Serrano, Begona Acha, Teresa Serrano-Gotarredona, Shouchun Chen, and Bernabé Linares-Barranc, *Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing–application to feedforward ConvNets*, IEEE transactions on pattern analysis and machine intelligence **35** (2013), no. 11, 2706–2719, DOI 10.1109/TPAMI.2013.71.

[PM12] Donatella Pascolini and Silvio Paolo Mariotti, *Global estimates of visual impairment: 2010*, British Journal of Ophthalmology **96** (2012), no. 5, 614–618, DOI 10.1136/bjophthalmol-2011-300539.

[RSB92] Ellen Bouchard Ryan, Barbara Szechtman, and Jay Bodkin, *Attitudes toward younger and older adults learning to use computers*, Journal of Gerontology **47** (1992), no. 2, P96–P101, DOI 10.1093/geronj/47.2.p96.

[Saf08] Dan Saffer, *Designing gestural interfaces: touchscreens and interactive devices*, O'Reilly Media, Inc., 2008, ISBN 9780596518394.

[Sal16] Timothy A. Salthouse, *Theoretical perspectives on cognitive aging*, Psychology Press, 2016.

[Sel04] Neil Selwyn, *The information aged: A qualitative study of older adults' use of information and communications technology*, Journal of Aging studies **18** (2004), no. 4, 369–384, DOI 10.1016/j.jaging.2004.06.008.

[SMFM17] Gianluca Schiavo, Ornella Mich, Michela Ferron, and Nadia Mana, *Mobile multimodal interaction for older and younger users: exploring differences and similarities*, Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia, ACM, 2017, DOI 10.1145/3152832.3156615, pp. 407–414.

[SSC99] Michael W Smith, Joseph Sharit, and Sara J Czaja, *Aging, motor control, and the performance of computer mouse tasks*, Human factors **41** (1999), no. 3, 389–396, DOI 10.1518/001872099779611102.

[STH+18] Arman Savran, Raffaele Tavarone, Bertrand Higy, Leonardo Badino, and Chiara Bartolozzi, *Energy and Computation Efficient Audio-visual Voice Activity Detection Driven by Event-cameras*, 13th IEEE Conference on Automatic Face and Gesture Recognition, 2018, DOI 10.1109/FG.2018.00055.

[VCG17] Eleftheria Vaportzis, Maria Giatsi Clausen, and Alan J Gow, *Older adults perceptions of technology and barriers to interacting with tablet computers: A focus group study*, Frontiers in Psychology **8** (2017), 1687, DOI 10.3389/fpsyg.2017.01687.

[VLvdHC13] Jo Vermeulen, Kris Luyten, Elise van den Hoven, and Karin Coninx, *Crossing the bridge over Norman's Gulf of Execution: Revealing feedforward's true identity*, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2013, DOI 10.1145/2470654.2466255, pp. 1931–1940.

[WDO04] Stephan AG Wensveen, Johan Partomo Djajadiningrat, and CJ Overbeeke, *Interaction frogger: a design framework to couple action and function through feedback and feedforward*, Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques, ACM, 2004, DOI 10.1145/1013115.1013140, pp. 177–184.

[Wel81] Alan T. Welford, *Signal, noise, performance, and age*, Human Factors **23** (1981), no. 1, 97–109, DOI 10.1177/001872088102300109.

[WHBP15] Christopher D. Wickens, Justin G. Hollands, Simon Banbury, and Raja Parasuraman, *Engineering psychology & human performance*, Psychology Press, 2015, ISBN 978-0205021987.

[WW11] Daniel Wigdor and Dennis Wixon, *Touch versus In-Air Gestures*, Brave NUI World (2011), 97–103, DOI 10.1016/B978-0-12-382231-4.00015-0.

Citation

Michela Ferron, Nadia Mana, Ornella Mich, Christopher Reeves, Gianluca Schiavo, *Designing Mobile Multimodal Interaction for Visually Impaired and Older Adults: Challenges and Possible Solutions*
Journal of Virtual Reality and Broadcasting, 16(2019), no. 2, December 2019, urn:nbn:de: 0009-6-348, ISSN 1860-2037.