Multimodal Interaction and Proactive Computing

Stephen A Brewster

Glasgow Interactive Systems Group
Department of Computing Science
University of Glasgow, Glasgow, G12 8QQ, UK
E-mail: stephen@dcs.gla.ac.uk Web: www.dcs.gla.ac.uk/~stephen

Abstract. One important issue for proactive computing is how users control and interact with the systems they will carry and have access to when they are out in the field. One solution is to use multimodal interaction (interaction using different combinations of sensory modalities) to allow people to interact in a range of different ways. This paper discusses gestural interaction as an alternative for input. This is advantageous as it does not require users to look at a display. For output non-speech audio and tactile displays are presented as alternatives to visual displays. The advantages with these types of displays are that they can be unobtrusive and do not require a user's visual attention. The combination of these underutilised senses has much potential to create effective interfaces for proactive systems.

1. Introduction

As more and more devices incorporate some form of computation people will soon carry and be connected to a large number of systems and services all of the time. Users going about their everyday lives need effective ways of managing them otherwise the effort of controlling them will be too great and they will not be used. To avoid such problems we need flexible, efficient ways to interact with and monitor the systems and services. In a proactive computing world these devices will also be making decisions for users, who will need to be kept informed of status and outcomes without unnecessary disruption [20]. Designing user interfaces to support such activities is not well understood (nor is how we realistically evaluate their effectiveness).

Ljungstrand *et al.* [14] suggest some important questions that need to be answered to develop the area of human-computer interaction (HCI) within proactive computing, amongst these are:

- What are the best means for controlling proactive computers and agents?
- What kind of manipulation and feedback mechanisms do users need, at what levels, how often, and how should feedback be manifested?
- How can we design user interfaces that take advantage of all the human senses, as well as our inherent skills in moving about in the real world and manipulating real things?

This paper will begin to deal with some of these issues, but much more work will need to be done to really understand how to design good proactive computing interactions. A starting point in thinking about how interactions might be designed is to look at how people currently cope with complex situations. In the real world we deal with large amounts of data all the time. We do this using a range of different senses and the combination of senses avoids any one becoming overloaded. Multimodal human-computer interaction studies the use of multiple different sensory modalities to enable users to interact effectively with computers.

The interface designs of most mobile and wearable computers are based heavily on those of desktop graphical user interfaces. These were originally designed for users sitting at a computer to which they could give their full (visual) attention. Users of proactive or mobile systems are often in motion and performing other tasks when they use their devices. If they are interacting whilst walking, running or driving, they cannot easily devote all of their visual attention to the interface; it must remain with the main task for safety. It can be hard to design visual interfaces that work well under these circumstances.

Much of the interface work on wearable computers tends to focus on visual displays, often presented through head-mounted graphical displays [1]. These can be obtrusive and hard to use in bright daylight, plus they occupy the users' visual attention [11] when it may be needed elsewhere. Other solutions utilise nearby resources that your 'personal server' might connect to for display or input [21]. These again may be difficult to use when on the move. One of the foci of work at Glasgow is on how far we can push non-visual interaction so that we do not tie users to visual displays and conventional input devices.

Both input and output need to be considered when designing proactive interactions. This paper will discuss some of the possibilities of the different senses and give some examples of how they might be used to create an effective proactive interface.

2. Input Techniques

Making input when in the kinds of scenarios envisaged by proactive computing is problematic; users will be out in the real world doing tasks that may be supported by computers. They may be mobile or engaged in an activity that needs the focus of their attention so cannot give it all to the computer they are carrying.

Current mobile and wearable computers typically use a touch screen and stylus, or small keyboard. These are effective when stationary but can be difficult to use when mobile. Buttons and widgets on touch screens tend to be small due to the small screens required to make the devices portable. This makes the targets hard to hit and input error prone, because the device and stylus are both moving as the user moves around the environment, making accurate pointing difficult. Similar problems affect stylus input of characters when on the move. Brewster [4] showed that when a stylus based device was used whilst walking performance dropped by over 30% compared to sitting. Small keyboards tend to have similar difficulties as the keys must be small enough to allow the keyboard to be easily carried and so become hard to press.

In all of these cases much visual attention is required to make input. Users must look closely to see the small targets and the feedback to indicate they have been used correctly. Visual attention is, however, needed for navigating the environment around the user. If too much is required for the interface then users may have to stop what they are doing to interact with the system, which is undesirable.

Many of these techniques also require two hands, which can be problematic if the user is engaged in other activities. The 'Twiddler' [1], a small chord keyboard, requires only one hand but it can be hard to use and requires learning of the chords.

Speech recognition is often suggested as a future alternative input technique. This has great potential but at present is not good in fully mobile environments due to high processor and memory requirements and highly variable background noise levels. There are also issues of error recovery without visual displays. If great care is not taken, error recovery can become very time consuming.

2.1 Gestural interaction

One alternative technique gaining interest is gestural interaction. Gestures can be done with fingers on touch screens, or using head, hands or arms (or other body parts) with appropriate sensors attached. They can also be attached to devices such as handheld computers or mobile phones to allow them to be able to generate gestures for input. Harrison *et al.* [13] showed that simple, natural gestures can be successfully used for input in a range of different mobile situations.

Gestures are a good method for making input because they do not require visual attention; you can do a gesture with your hand, for example, without looking at it because of your powerful kinaesthetic sense – you know the positions, orientations and movements of your body parts because you sense them through your muscles, tendons and joints. This means that input can be made without the need for visual attention.

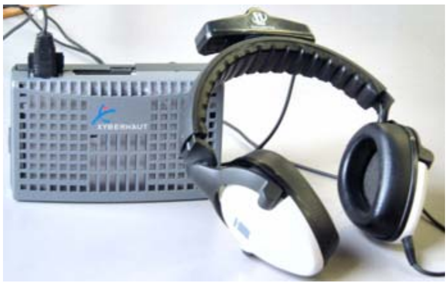

Figure 1: A simple wearable computer system comprising a Xybernaut MAV wearable computer, a pair of standard headphones and an Intersense orientation tracker for detecting head movements (on top of the headphones) [7].

The use of hands or arms may be problematic if users are carrying equipment, but there are still possibilities for input via the head. We have looked at using head nods for making selections whilst on the move [7]. Head pointing is more common for desktop users with physical disabilities [15], but has advantages for all users, as head movements are very expressive. There are many situations where hands are busy but the head is still free to be used for input. There are still important issues of gesture recognition to be dealt with as users nod and shake their heads as part of normal life and we need to be able to distinguish these nods, or nods that people might do when listening to music, from nods to control the interface.

Figure 1 shows an example of a simple audio-based wearable computer that used head gestures for input [7]. The sensor we used was an off-the-shelf model which could easily be made much smaller and integrated into the headphones. Figure 2 shows a Compaq iPAQ with an accelerometer attached (devices such as mobile phones are now also incorporating accelerometers). This can be used to detect movement and orientation of the device. We have also used this to allow tilting for input. In the simplest case this might be tilting to scroll (although this can be difficult as the more you tilt the harder the screen is to see) or more sophisticated interactions may use tilting for text entry. Gesturing with the whole device is also possible, for example to allow users to point at objects or draw simple characters in space in front of them.

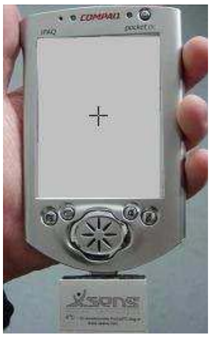

Figure 2: A Compaq iPAQ handheld computer with an Xsens 3-axis accelerometer for detecting device movements (www.xsens.com).

To assess the use of fingers on touch screens for input we developed a gesture driven mobile music player on a Compaq iPAQ [17]. Centred on the functions of the music player – such as play/stop, previous/next track – we designed a simple set of gestures that people could perform whilst walking. Users generated the gestures by dragging a finger across the whole of the touch screen of the device (which was attached to a belt around their waist) and received non-speech audio feedback upon completion of each gesture. Users did not need to look at the display of the player to be able use it. An experiment showed that the audio/gestural interface was significantly better than the standard, graphically based, media player on the iPAQ when users were operating the device whilst walking. One reason for this was that they could use their eyes to watch where they were going and their hands and ears to control the music player.

These kinds of interactions have many benefits for proactive systems. Users can make input with parts of their bodies that are not being used for the primary task in which they are involved. Certain types of input can be made without the need for a screen, or even a surface, which makes them very flexible and suitable for the wide range of interaction scenarios in which proactive computer users might find themselves.

2.2 Sensing additional information from accelerometers

One extra advantage of devices equipped with accelerometers (such as Figure 2) and other motion sensors is that other useful information can be gained about the context of the interaction in addition to data for gesture recognition. This is important for proactive systems as they must communicate with their users in subtle but effective ways and knowing something about the user's context will help this. There is much existing work in the area of context-aware computing which is beyond the scope of this paper, but data from accelerometers gives some other useful information that has not been considered so far. With instrumented devices we can collect information to provide input to allow a system to make decisions about how and when to present information to a user, and when to expect input.

Alongside gesture recognition, we can use the accelerometers to gather information about the user's movement. When users are walking, for example, we can extract gait information from the data stream. Real-time gait analysis allows the display to be changed to reduce its complexity if the user is walking or running, as the users attention will be elsewhere, or to compensate for input biases and errors that occur because of the movement. For example, we have found that users are significantly more accurate when tapping targets during particular parts of the gait cycle. So, any system we create must allow the user to interact appropriately when on the move or we may end up with a system that is unusable, or alternatively forces the user to stop what he/she is doing to operate the interface. The accelerometers also give us information about tremor from muscle movements that we can use to infer device location and use. We have used this, for example, to allow the user to squeeze the device to make selections; the tremor frequency changes when the user is squeezing and we can easily detect this change and use it as an input signal.

3. Output

Current mobile and wearable devices use small screens for displaying information and this makes interaction difficult. Screen size is limited as the devices must be small enough to be easily carried. As mentioned above, the user interfaces of many current mobile and wearable computers use interaction and display techniques based on desktop computer interfaces (for example, windows, icons, pull-down menus). This is not necessarily the best solution as users will not be devoting their full attention to the systems and devices they are carrying; they will need to keep some of their attention on the tasks they are performing and the environment through which they are moving. Head-mounted augmented-reality displays overcome some of these problems by allowing the user to see the world around them as well as the output from their wearable systems. However, there will always be problems with the competing demands on visual attention (and also the obtrusive technologies that users currently have to wear). Humans have other senses which are useful alternatives to the visual for information display, but they are often not considered. One aim of the research done at Glasgow is to create systems that use as little of the users' visual attention as possible by taking advantage of the other senses.

3.1 Non-speech audio display

There is much work in the area of speech output for interactive systems, but less on non-speech sounds. These sounds include music, sound from our everyday environment and sound effects. These are often neglected but can communicate much useful information to a listener. With non-speech sounds the messages can be shorter than speech and therefore more rapidly heard (although the user might have to learn the meaning of the non-speech sound whereas the meaning is contained within the speech – just like the visual case of icons and text). The combination of these two types of sounds makes it easy for a proactive system both to present status information on continuously monitored tasks in the background and to capture a user's attention with an important message.

There are two basic types of non-speech sounds commonly used: Earcons [2] and Auditory Icons [10] (for a full discussion of the topic see [3]). Earcons are highly structured sounds based around principles from music, encoding information using variations in timbre, rhythm and melody. Auditory icons use natural, everyday sounds that have an intuitive link to the thing they represent in the computer. The key advantages of non-speech sounds is that they are good for giving status information, trends, for representing simple hierarchical structures and grabbing the user's attention. This means that information that may normally be presented visually could be presented in sound, thus allowing users to keep visual attention on the world around them.

Sound can significantly improve interaction in mobile situations. Brewster [4] showed that the addition of simple non-speech sounds to aid targeting and selection in a stylus/touch screen interface significantly reduced subjective workload, increased tapping performance by 25% and allowed users to walk significantly further. This was because the user interface required less of the users' visual attention, which they could then use for navigating the environment. This suggests that information delivered in this way could be very beneficial for proactive computing environments.

Sawhney and Schmandt's Nomadic Radio [18] combined speech and auditory icons. The system used a context-based notification strategy that dynamically selected the appropriate notification method based on the user's attentional focus. Seven levels of auditory presentation were used from silent to full speech rendering. If the user was engaged in a task then the system was silent and no notification of an incoming call or message would be given (so as not to cause an interruption). The next level used 'ambient' cues (based on Auditory Icons) with sounds like running water indicating that the system was operational. These cues were designed to be easily habituated but to let the user know that the system was working. Other levels used speech, expanding from a simple message summary up to the full text of a voicemail message. The system attempted to work out the appropriate level to deliver the notifications by listening to the background audio level in the vicinity of the user (using the built-in microphone) and if the user was speaking or not. For example, if the user was speaking the system might use an ambient cue so as not to interrupt the conversation

One extension of basic sound design is to present sounds in three-dimensions (3D) around the listener. This gives an increased display space, avoiding the overload that can occur when only point source or stereo sounds are used. Humans are very good at detecting the direction of a sound source and we can use this to partition the audio space around the listener into a series of 'audio windows' [9]. To increase the accuracy of perception most 3D auditory interfaces just use a plane around the users head at the height of the ears. Audio sources can then be played in different segments of the circle around the head. The use of a head-tracker (see Figure 1) means that we can update the sound scene dynamically, allowing egocentric or exocentric sound sources.

Brewster *et al.* [7] used a 3D auditory display to create an 'eyes-free' interaction for use on the move. As mentioned above, this interface used head nods to allow users to interact: a nod in the appropriate directed selected a source. The idea behind the system was that a user might have a range of different sound sources around his/her head playing in the background but when required a nod would bring a source to the centre of attention. An evaluation of this interaction was undertaken whilst users were walking. A wide range of usability measures was taken, from time and error rates to subjective workload, percentage preferred walking speed and comfort. These showed that such an interaction was effective and users could easily make selections of auditory objects when on the move. It also showed that egocentric positioning of sound sources allowed faster interactions but with higher error rates than exocentric positioning. This shows that a proactive system that used sound (and gestures) in this way could be used whilst the user was mobile. Work is progressing on the development of more sophisticated interactions in a 3D audio space [16].

3.2 Vibrotactile displays

Vibrotactile displays are another possibility for non-visual output. They have been very effective in mobile telephones and personal digital assistants (PDAs), but their displays are crude, giving little more than an alert that someone is calling. The sense of touch can do much more. As Tan [19] says "In the general area of human-computer interfaces … the tactual sense is still underutilised compared with vision and audition". Our cutaneous (skin-based) sense is very powerful, but has been little studied in terms useful for proactive computing. This has begun to change as more sophisticated devices are now easily available that can be used on mobile devices (see Figure 3). Tactile displays have an advantage over audio ones in that they are private, so others around you cannot hear the information being presented,

Recent work has started to investigate the design of tactile icons, or Tactons. These are structured vibrotactile messages that can be used alongside audio or visual displays to extend the communication possibilities [5, 8]. The key parameters of touch that can be used to encode information are: waveform, rhythm and body location. Brown *et al.* [8] have shown that information can be encoded into Tactons in the same way as in Earcons, with the same levels of recognition. Brewster and King [6] have shown that Tactons can successfully provide information about the progress of tasks.

This is important as it means that progress and status information can be delivered using this modality without requiring the visual attention of the user.

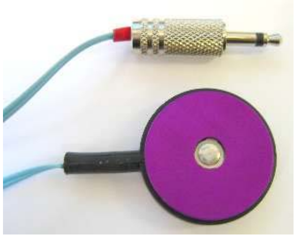

Figure 3: An Engineering Acoustics Inc. C2 tactile display (www.eaiinfo.com).

Tactile displays can be combined with audio and visual ones to create fully multimodal displays. There are interesting questions about what type of information in the interface should be presented to which sense. Tactons are similar to Braille in the same way that visual icons are similar to text, or Earcons are similar to synthetic speech. For example, visual icons can convey complex information in a very small amount of screen space, much smaller than for a textual description. Earcons convey information in a small amount of time as compared to synthetic speech. Tactons can convey information in a smaller amount of space and time than Braille. Research will show which form of iconic display is most suitable for which type of information. Crudely, visual icons are good for spatial information, Earcons for temporal. One property of Tactons is that they operate both spatially and temporally so they can complement both icons and Earcons. Further research is needed to understand fully how these different types of feedback work together.

4. Users with a range of abilities

Proactive computing using multimodal interaction offers many new possibilities for people with disabilities. These may be physical disabilities or disabilities caused by the environment or working conditions. For example, the multimodal displays described above are valuable for visually-impaired people as they do not use visual presentation. They can also be effective for older adults; Goodman *et al.* [12] showed that older users could perform as well as younger ones in a mobile navigation task when multimodal displays were used on a handheld computer. Another advantage of multimodal displays is that information can be switched between senses. So someone with hearing loss could use a tactile and visual display, whilst someone with poor eyesight could use a tactile and audio one to access the same systems and services. These advantages also apply to physically able users who are restricted by environment (for example, bright sun makes visual displays hard to use, loud background noise makes audio input and output impossible) or clothing (jobs requiring gloves or goggles make it hard to use keyboards or screens). Information can be switched to a different modality as appropriate to allow users to interact effectively.

5. Discussion and Conclusions

This paper has presented a range of input and output techniques using different sensory modalities. One of the key issues for interaction with proactive computer systems is that computing takes place away from the office and out in the field [20]. This causes problems for standard interaction techniques as they are not effective when users are on the move. Using different senses for input and output can avoid some of these problems. Our different senses are all capable of different things and interaction designers can take advantage of this to create suitable interactions. This is also dynamic as users out in the field will be subject to changing environments and tasks. Good proactive interface design will allow interaction to move between different techniques and senses as situations change.

Evaluating interfaces to proactive systems has had little attention. New techniques will need to be developed to allow us to test the sophisticated interactions we need to develop in realistic usage scenarios. At Glasgow we have begun to develop a battery of tests to allow us to evaluate mobile and wearable devices in mobile but controlled conditions so that we can discover if our new interaction designs are successful or not [4, 7, 17].

As Ljungstrand *et al.* suggest, there are many questions to be answered before we can construct effective user interfaces to proactive computing systems and much research is still needed. However, we can see that multimodal displays are a key part of these interactions. Using gestures, for example, is a good way to allow flexible, dynamic input whilst the user is involved in other tasks. Gestures do not need a visual display and many different parts of the body can be used to make gestures so they can be effective even if the hands are busy. Feedback through audio or tactile displays offer solutions when visual displays are not possible. The combination of all three types of display can be very powerful. We have also seen that when we deliver feedback and expect input can have significant effects on users in terms of selection accuracy and movement. If we force them to attend to information and make input when it is not suitable then there may be consequences for the primary task in which they are involved.

Acknowledgements

This work was funded by EPSRC Advanced Research Fellowship GR/S53244.

References

1. Barfield, W. and Caudell, T. (eds.). *Fundamentals of wearable computers and augmented reality*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 2001.
2. Blattner, M., Sumikawa, D. and Greenberg, R. Earcons and icons: Their structure and common design principles. *Human Computer Interaction*, *4* (1). 11-44.
3. Brewster, S.A. Chapter 12: Non-speech auditory output. In Jacko, J. and Sears, A. eds. *The Human Computer Interaction Handbook*, Lawrence Erlbaum Associates, 2002, 220-239.

4. Brewster, S.A. Overcoming the Lack of Screen Space on Mobile Computers. *Personal and Ubiquitous Computing*, *6* (3). 188-205.
5. Brewster, S.A. and Brown, L.M., Tactons: Structured Tactile Messages for Non-Visual Information Display. In *Proceedings of Australasian User Interface Conference 2004*, (Dunedin, New Zealand, 2004), Austalian Computer Society, 15-23.
6. Brewster, S.A. and King, A.J., The Design and Evaluation of a Vibrotactile Progress Bar. In *Proceedings of WorldHaptics 2005*, (Pisa, Italy, 2005), IEEE Press.
7. Brewster, S.A., Lumsden, J., Bell, M., Hall, M. and Tasker, S., Multimodal 'Eyes-Free' Interaction Techniques for Wearable Devices. In *Proceedngs of ACM CHI 2003*, (Fort Lauderdale, FL, USA, 2003), ACM Press, Addison-Wesley, 463-480.
8. Brown, L., Brewster, S.A. and Purchase, H., A First Investigation into the Effectiveness of Tactons. In *To appear in Proceedings of World Haptics 2005*, (Pisa, Italy, 2005), IEEE Press.
9. Cohen, M. and Ludwig, L.F. Multidimensional audio window management. *International Journal of Man-Machine Studies*, *34*. 319-336.
10. Gaver, W. The SonicFinder: An interface that uses auditory icons. *Human Computer Interaction*, *4* (1). 67-94.
11. Geelhoed, E., Falahee, M. and Latham, K. Safety and comfort of eyeglass displays. In Thomas, P. and Gellersen, H.W. eds. *Handheld and Ubiquitous Computing*, Springer, Berlin, 2000, 236-247.
12. Goodman, J., Brewster, S.A. and Gray, P.D. How can we best use landmarks to support older people in navigation? *Behaviour and Information Technology*, *24* (1). 3-20.
13. Harrison, B.L., Fishkin, K.P., Gujar, A., Mochon, C. and Want, R., Squeeze me, hold me, tilt me! An exploration of manipulative user interfaces. In *Proceedings of ACM CHI'98*, (Los Angeles, CA, 1998), ACM Press Addison-Wesley, 17-24.
14. Ljungstrand, P., Oulasvirta, A. and Salovaara, A., Workshop Forward. In *Workshop 6: HCI Issues in Proactive Computing (Workshop at NordiCHI 2004)*, (Tampere, Finland, 2004), iv-v.
15. Malkewitz, R., Head pointing and speech control as a hands-free interface to desktop computing. In *Proceedings of ACM ASSETS 98*, (Marina del Rey, CA, 1998), ACM Press, 182-188.
16. Marentakis, G. and Brewster, S.A., A Study on Gestural Interaction with a 3D Audio Display. In *Proceedings of MobileHCI 2004*, (Glasgow, UK, 2004), Springer LNCS, 180-191.
17. Pirhonen, A., Brewster, S.A. and Holguin, C., Gestural and Audio Metaphors as a Means of Control for Mobile Devices. In *Proceedings of ACM CHI 2002*, (Minneapolis, MN, 2002), ACM Press, 291-298.
18. Sawhney, N. and Schmandt, C. Nomadic Radio: speech and audio interaction for contextual messaging in nomadic environments. *ACM Transactions on Human-Computer Interaction*, *7* (3). 353-383.
19. Tan, H.Z. and Pentland, A., Tactual Displays for Wearable Computing. In *Proceedings of the First International Symposium on Wearable Computers*, (1997), IEEE.
20. Tennenhouse, D. Proactive Computing. *Communications of the ACM*, *43* (5). 43-50.
21. Want, R., Pering, T. and Tennenhouse, D. Comparing autonomic computing and proactive computing. *IBM Systems Journal*, *42* (1). 129-135.