In *Proceedings of the ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments,* Kloster Irsee, Germany, 17–19 June 2002. (http://www.sigmedia.org/ids02/)

Usability Issues and Methods for Mobile Multimodal Systems

Anthony Jameson

DFKI, German Research Center for Artificial Intelligence; and
International University in Germany

Abstract. Mobile multimodal systems raise some novel usability challenges. Some of these are due to the combination of two characteristics of mobile systems and multimodal systems, respectively: the competition between the system and the environment for the user's attention; and the availability of multiple modalities for user input and system output. This paper first presents a theoretical argument that the set of problems raised by the combination of these two characteristics is more than just the union of the two sets raised by each characteristic itself. It then discusses one relatively new method—mobile eye tracking—that can help with the empirical study of these problems. Finally, it considers the question of how automatic system adaptation to a user's current resource limitations might ultimately enhance the usability of mobile multimodal systems.

1 Expanding the Scope of Usability Research

Combining the properties of mobility and multimodality raises many technical challenges, many of which are discussed in papers presented at this workshop—concerning issues ranging from how to deal with severe computational resource constraints to how to achieve accurate and robust interpretation of users' input. But even if all technical problems had already been solved satisfactorily, there would be enough issues of another sort left to fill an entire workshop: issues concerning the usability of mobile multimodal systems.

This type of usability issue has been studied most intensively within the subarea of in-car systems for drivers (see, e.g., Wierwille, 1993; Bernsen & Dybkjær, 2001), which overlaps with the area of mobile multimodal systems. With this type of system, it is especially obvious that it is not enough for a system to be usable in isolation: The driver's interaction with an in-car system must fit seamlessly into a complex context. For example, even a slight tendency of the system to distract the driver unnecessarily, which might go unnoticed in any other type of system, can have disastrous consequences. Motivated by the high safety and financial stakes involved, research in this area has produced a substantial and rapidly growing set of empirical research methods, theoretical and computational models, and empirical results.

Usability issues have also been studied intensively in the overlapping subarea of spoken dialog systems (see, e.g., Dybkjær & Bernsen, 2000). Here as well, the number of systems developed and tested so far—and their commercial significance—have been sufficient to give rise to a useful collection of methods, guidelines, and specific results.

With other types of mobile multimodal system, there has been some attention to usability issues, but it has been less systematic and extensive than in the two overlapping subareas just mentioned. One goal of the present contribution is to encourage increased adoption of the sort of usability research that is already found in these two subareas. Consider, for example, a mobile multimodal tourist guide worn by a user who is sightseeing on foot in an unfamiliar town: The consequences of excessive demands on the user's attention will be less drastic than they would be if she were driving fast in complex traffic. If she becomes confused about what she can say and do with the system, she will not abandon it as quickly as a caller who experiences similar problems with an inadequately tested spoken telephone dialog system. Still, usability deficits of these and other types can significantly diminish the quality of the user's experience, perhaps leading to the ultimate rejection of the system—regardless of any innovative technical solutions that the system may embody.

Extending the scope of usability research from these two overlapping subareas is not just a matter of transferring methods and results. Many mobile multimodal systems have quite different usage scenarios than in-car systems; and most of what is known about the usability of spoken dialog systems concerns speech-only systems. The present paper will discuss examples of usability issues and methods that deserve increased attention.[1]

2 System-Environment Competition for User Resources

One family of usability issues is due to a combination of two characteristics of mobile multimodal systems:

1. *System-environment competition:* There exists largely continual competition between the system ("$\mathcal{S}$") and the environment for various perceptual, motor, and cognitive *resources* of the user ("$\mathcal{U}$").
   This fact is due to the mobility of such systems: Users of stationary systems can usually largely ignore their environment (though the extent to which a system's usability is robust to distractions such as phone calls and visitors is sometimes an issue).
   $\mathcal{U}$'s attention to the environment may be due simply to distracting stimuli in the environment (as when $\mathcal{U}$ is being driven in a taxicab while using $\mathcal{S}$); but often $\mathcal{U}$ will be attending actively to the environment while performing actions in it (e.g., handling objects or navigating through the environment).
2. *Flexibility afforded by multimodality:* $\mathcal{U}$ typically has considerable flexibility in deciding what resources to employ when generating input to $\mathcal{S}$ and when processing $\mathcal{S}$'s output.
   This fact is due to multimodality. At any given time, $\mathcal{U}$ may be able to choose between providing her input via speech, writing, gesture, or some combination of these modalities. And $\mathcal{S}$'s output may often include some redundancy due to the use of more than one modality. For example, $\mathcal{U}$ may be able to choose whether to listen to an instruction generated with synthetic speech; to read the instruction

[1] Further examples and illustrative material can be found in the slides for the workshop address, which are available from http://dfki.de/~jameson/abs/Jameson02IDS.html.

Speed−Manual	*Speed−Voice*
Recall speed number	Move hand to phone (*)
Move hand to phone (*)	Attend to phone
Attend to phone	Press *Power* (*)
Press *Power* (*)	Move hand to wheel (*)
Attend to phone	Say name (*)
Press speed number	Listen for name (*)
Press *Send* (*)	Listen for *Connecting* (*)
Move hand to wheel (*)	

Fig. 1. *Summaries of two cognitive models of two methods of dialing a phone number in a car: manually, using a single-digit shortcut (left); and with speech, using a name as a shortcut (right). Each asterisk (*) indicates a point at which $\mathcal{U}$ is expected to return attention to the driving task. Adapted from Table 2 of Salvucci (2001).*

on the system's screen; to use both methods; or to alternate between the two of them.

Each of these characteristics is quite familiar in connection with the type of system in question (mobile systems and multimodal systems, respectively). The simultaneous presence of these characteristics in mobile multimodal systems does not mean merely that designers of such systems have to deal with both of them; the characteristics also interact to produce an overall challenge which is greater than the sum of the two challenges individually. To see why, we can look at each characteristic more closely.

2.1 Competition Between System and Environment

One way of analyzing the competition between system and environment is to construct some sort of cognitive model that explicitly describes $\mathcal{U}$'s cognitive processing, physical actions, and perception. For example, Salvucci (2001) used the ACT-R cognitive architecture to model several methods of dialing a car phone (see Figure 1 for simple summaries of two of the methods). With each of these methods, $\mathcal{U}$ has no choice among modalities for input or output. Hence, each model is able to state fairly definitely what $\mathcal{U}$ needs to do in order to execute the method in question. Adequate modeling still raises some subtle issues, such as: At what points in the procedure will $\mathcal{S}$ shift her attention back to the driving task (cf. the asterisks in the figure)? And predicting the exact result of the competition between the system and the environment requires detailed consideration of each step in the model.[2]

The use of such models becomes more problematic when users can choose freely among modalities. To take a simple example, suppose that a dialing system offers not

[2] Salvucci (2001) implemented the dialing models and a driving model in ACT-R, using the models to derive quantitative predictions, which fit fairly well with empirical data on dialing and driving performance. But even just a careful pencil-and-paper consideration of models like this can help to focus thinking about the resource demands of different methods.

Speed–Voice with Visual Feedback
Move hand to phone (*)
Attend to phone
Press *Power* (*)
Move hand to wheel (*)
Say name (*)
[Select:
  Listen for name (*)
  Look for appearance of name on display (*)]
[Select:
  Listen for *Connecting* (*)
  Look for confirmation of connection on display (*)]

Fig. 2. *Extension of the second model of Table 1 to describe a voice dialing system that offers redundant visual feedback.*

only spoken feedback concerning the status of the dialing process but also redundant visual feedback on the phone's display (cf. the discussion of a related design issue by Bernsen & Dybkjær, 2001). The model of the dialing process then needs to be extended as in Figure 2. This apparently innocuous extension can actually drastically change the nature of the task analysis. Since the earliest days of cognitive modeling in the field of human-computer interaction (see, e.g., Card, Moran, & Newell, 1983), it has been recognized that where a task analysis shows that users have some choice among alternative methods, their behavior is relatively hard to predict. If there is only one way of successfully performing a task, it is reasonable to assume that users will learn that method sooner or later and henceforth execute it in a more or less predictable fashion. Where free choices are available, various factors can determine which alternative a given user will select in a given situation, as will be discussed below.

For now, we should note that the frequency with which users choose to attend to the visual feedback can make a great difference to the evaluation of the voice dialing method shown in Figure 2. If users ignore this feedback, the favorable results reported by Salvucci (2001) apply: relatively short dialing times and relatively little impairment of driving performance. But if (for whatever reason) some users tend to look at the display as they await visual feedback, their driving performance may be seriously impaired.

2.2 Drawbacks of Flexibility

A principal motivation for giving users more than one way to perform a given task is to allow them to choose, in each individual case, the method that seems most appropriate given the demands of the task and their own capabilities. For example, Oviatt (1999) notes that one factor that tends to enhance the recognition accuracy of multimodal systems is people's "natural intelligence about when and how to deploy input modes effectively" (p. 79). It is not obvious, however, that this "natural intelligence" will serve

users as well when environmental demands enter the picture. In a stationary multimodal system, in order to learn that a spatial reference can be accomplished more effectively through drawing than through speech, a user may need only to try each method a few times. But what about a choice between drawing and speech when the current environment needs to be taken into account? In addition to the inherent suitability of the two methods for the task at hand, qualitatively different factors may have to be considered by the user (e.g., the drawbacks of speech in terms of disturbing other persons or revealing sensitive information; the extent to which drawing would interfere with the user's ability to navigate through the environment). Since specific combinations of task states and environmental situations may arise that the user has never encountered before, previously formed habits may be of limited use, or even detrimental. And users typically do not have enough resources available to make carefully reasoned decisions—even if they had enough knowledge of the relevant factors to do so.

Moreover, users' choices among methods are not always rational and optimal even in simpler situations. As Card et al. (1983) found in their early studies, users may apply a subset of the available methods out of habit or because they are not familiar with all of the possibilities.

And finally, choosing among alternative methods is often in itself a cognitive process that consumes resources. In some situations, the potential benefits of being able to choose are outweighed by the overhead involved in choosing (see, e.g., Olson & Nilsen, 1987/1988).

In sum, the combination of system-environment competition for resources and the flexibility of multimodal interaction makes life more difficult for designers and usability analysts on the one hand and for users on the other hand. The designers and usability analysts have an especially hard time predicting important aspects of users' behavior on the basis of theoretical considerations. And the users are continually being faced with choices that can have significant consequences but that they are in a poor position to make.

3 Methods for Tracking the Allocation of Attention

Because of the considerations discussed in the previous section, it is important to have available effective methods for investigating empirically how users allocate their resources to the system and to the environment. Collecting data on this question—and on other usability questions as well—is made more difficult by both mobility and multimodality. Mobility introduces computational constraints for data capture tools while at the same time introducing a need to capture a great deal of information that is in general not required in studies of a stationary system, namely information about the user's current environment and activities in that environment. Multimodality makes data collection relatively cumbersome because of the different data streams involved and the fact that some types of data (e.g., gesture), are harder to record and analyse than data generated with more traditional systems.

One ambitious data collection infrastructure was presented by Oviatt (2000): With equipment worn by a roaming user, recordings were made of the user's speech and gestural input and of the system's output. Somewhat more recently, Lyons and Starner

Fig. 3. *Image from a conventional video of a user interacting with a handheld computer.*

(2001) introduced into a similar data collection system a way of capturing the user's view of the current environment: an eyeglasses-mounted video camera that shows the user's field of view. (Other cameras can be added, for example, to record the movements of the user's hands.)

While these methods can yield useful data for analysis of problems such as those discussed above, what is still lacking is detailed information on what the user is looking at. For this purpose, an eye tracker is a valuable addition to the usability analyst's repertoire. Eye tracking has a long tradition in studies of driving and the use of in-car systems for drivers (see, e.g., Sodhi et al., 2002), but handheld and wearable computers raise somewhat different methodological challenges. Figure 3 shows a conventional video image taken of a user interacting with a handheld computer while walking around a room.[3] This image does not reveal whether the user is looking at the handheld computer's screen or at the sheet of paper that she has positioned behind and above it. Figure 4 shows images captured by the head-mounted camera of an ASL 501 Mobile eye tracker. Although the video images show the scene from the user's currrent perspective (like the head-mounted camera of Lyons & Starner, 2001), they do not in themselves reveal the focus of her attention. But this information is given with quite usable accuracy by the black cross generated by the eye tracking hardware and software: It is possible to see not only when the user is looking at the screen but also roughly where on the screen she is looking.

[3] This video was made in the DFKI Evaluation Center for Language Technology by Boris Brandherm, Kerstin Klöckner, and Marie Norlien.

Fig. 4. *Images from the video stream produced by the mobile eye tracker worn by the user shown in Figure 3.*

The address at the workshop includes further examples illustrating some of the potential and limitations of mobile eye tracking for the study of mobile multimodal systems.[4]

To be sure, the analysis of data generated in the way shown in Figure 4 is more challenging than the analysis of more familiar types of eye tracking data. On the positive side, we have found that even brief, informal testing of this sort, without quantitative analysis, can reveal important usability problems with design implications. For example, informal investigation of a small sample of users may reveal a strong tendency to look at some of the redundant visual feedback on the screen, even when it is in principle unnecessary—and potentially dangerous—to do so. Such results can constitute adequate motivation to redesign some aspects of the system—for example, by providing more adequate nonvisual feedback and/or by eliminating the redundant visual feedback.

4 Recognizing and Adapting to Environmental Demands

Given the importance of system/environment competition in determining the success of interaction in mobile multimodal systems, it is natural to consider the possibility of automatic adaptation by the system to the user's current environmentally determined resource limitations. Figure 5 illustrates the type of adaptation that has been investigated in the research project READY at Saarland University (cf. Jameson, Schäfer, Weis, Berthold, & Weyrath, 1999). The first user of the airport assistance system shown in the figure is thinking mainly about the task of moving swiftly through the airport terminal (and about the prospect that his plane will leave without him). He is therefore probably best served with instructions that require minimal attention and processing. By contrast, the second user is in a position to benefit from system behavior that calls for a much greater investment of time and attention. In principle, the desired system behavior could be chosen explicitly by the user; but as was discussed in 2.2, giving the

[4] Workshop participants are also invited to bring forward their own systems for testing with the mobile eye tracker during the days of the workshop.

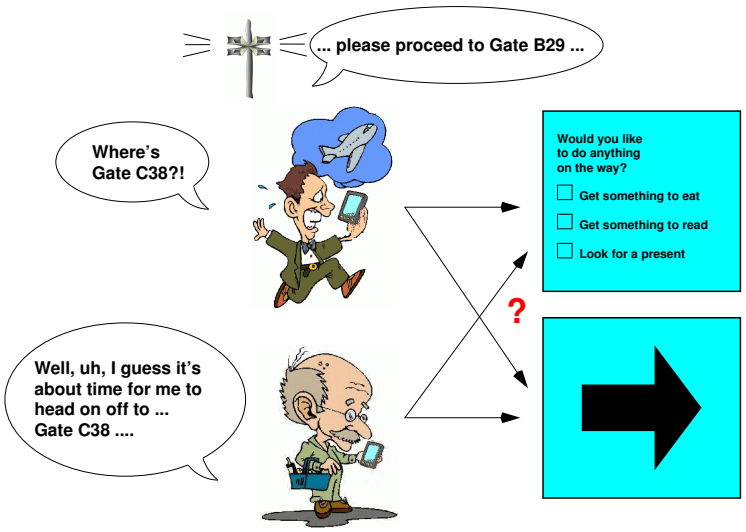

Fig. 5. *Example of how a user's current resource limitations can call for different system responses.*

user more choices can have serious drawbacks with mobile multimodal systems. So the questions arise:

- How can a mobile multimodal system assess on the basis of indirect evidence the extent to which the user is currently being distracted by the environment?
- How can the system appropriately adapt its behavior to take these assessments into account?

4.1 Recognizing Resource Limitations

Regarding the first question, one potentially valuable source of evidence is the user's speech input. Building upon many studies by previous researchers, Müller, Großmann-Hutter, Jameson, Rummer, and Wittig (2001) investigated in an experimental environment the ways in which a user ($\mathcal{U}$)'s speech is affected by two orthogonal manipulations: (a) whether or not $\mathcal{U}$ is required to navigate through a two-dimensional simulation of an airport while speaking; and (b) whether $\mathcal{U}$ is subjected to time pressure in the speaking task. A recent replication (Kiefer, 2002) added a third dimension: whether $\mathcal{U}$ is being distracted by irrelevant speech in the form of the loudspeaker announcements that are typical of the airport environment. Figure 6 gives an overview of the eight conditions in which users generated speech to a hypothetical assistance system. Conventional data analyses showed significant effects of these manipulations on a number of aspects of speech, including various types of pauses, the length of utterances, articulation rate, and the presence of disfluencies such as false starts. The effects of the two types of distraction (navigation and acoustic distraction) on speech are on the whole similar; and as one might expect, time pressure has a rather different pattern of effects.

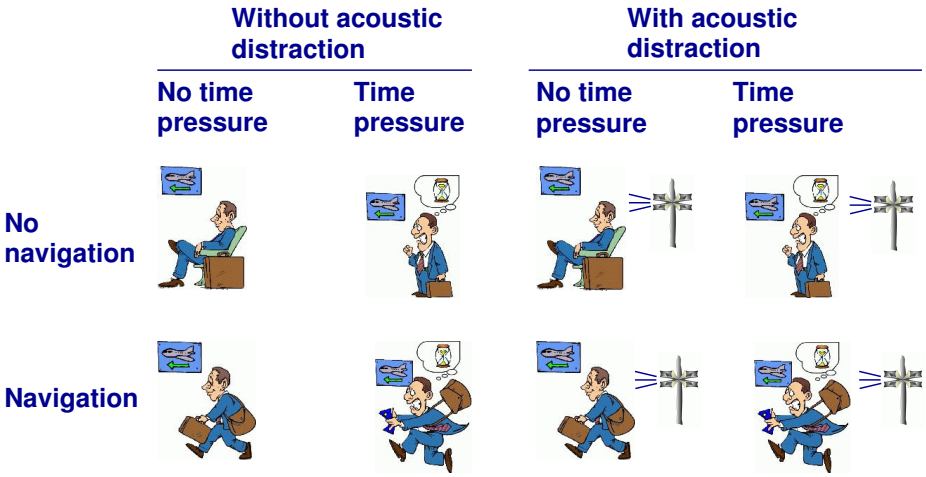

Fig. 6. *Visualization of the eight conditions realized in two experiments on the recognition of resource limitations on the basis of speech.*

The key question with regard to the possibility of automatic adaptation is the extent to which a system, given some samples of a user's speech in one of the eight conditions shown in Figure 6, can recognize which of the conditions the user was in when he produced that speech. Analyses of the experimental data with regard to this question showed that on the whole a system can recognize each of the three factors with considerably above-chance accuracy (e.g., assigning a probability of about 0.65 to the correct hypothesis after 3 utterances); but that the system's ability to make such discriminations varies considerably from one condition to another. For example, it seems to be easier to recognize whether a user is navigating if he is *not* under time pressure (his utterances then being longer, more complex and more revealing) and if he *is* distracted by loudspeaker announcements (perhaps because then the navigation task is more likely to exceed the available resources).

Although the analysis of the combined data from these two experiments is still in progress at the time of this writing, it seems clear that evidence from the user's speech can by itself yield only a probabilistic assessment of the user's currently available resources. We are working on ways of integrating other types of evidence as well:

- Previous theory and experimental results suggest that certain features of $\mathcal{U}$'s motor behavior (e.g., tapping especially hard on the touch screen, or tapping on the wrong icon) ought to occur more frequently under conditions of cognitive load and/or time pressure (Lindmark, 2000).
- Many eye tracking studies have shown that the diameter of a user's pupils changes as a function of cognitive load (see, e.g., Granholm, Asarnow, Sarkin, & Dykes, 1996). In situations where it is feasible to employ an eye tracker during normal system use, data concerning pupil diameter should be usable as evidence.
- The actions that $\mathcal{U}$ is currently performing with the system $\mathcal{S}$ (e.g., giving commands to a navigation system vs. scrolling at normal reading speed through the

text of a novel) may suggest the extent to which $\mathcal{U}$ is or is not currently attending to the environment as well as to the system.

4.2 Adapting the System's Behavior to Perceived Resource Limitations

Although it seems plausible that different system behaviors may be appropriate given different situationally determined resource limitations, determining suitable adaptations is in general a tricky and multifaceted problem. $\mathcal{S}$'s assessment of $\mathcal{U}$'s resource limitations will almost always be error-prone, and conflicting and situation-dependent goals may need to be taken into account (e.g., speed of task execution vs. avoidance of errors). Jameson et al. (2001) discuss ways of dealing with these and other problems within a decision-theoretic framework. On a more general level, it can be important for the system's behavior to be predictable, understandable, and controllable (cf. Jameson, 2002; Jameson & Schwarzkopf, 2002), and these goals can seriously constrain the types of adaptation that are desirable.

5 Conclusions

This essay has raised more questions than it has answered; but this fact appears to be symptomatic of the current state of our understanding of the usability issues raised by mobile multimodal systems. The competition between system and environment for the user's resources, combined with the flexibility afforded by multimodality, raises new challenges for designers who wish to predict how users will interact with their systems; for evaluators who want to study such interaction empirically; and for those who aim to enhance the usability of systems by realizing automatic adaptation. The goal of the present paper has been mainly to encourage researchers in this area to allocate adequate resources to issues such as those discussed here, despite the many other demands on their attention that they experience when developing and studying mobile multimodal systems.

Acknowledgements

This paper was prepared in the context of the Evaluation Center for Language Technology Systems of the DFKI project COLLATE, which is funded by the German Ministry of Education, Research, Science, and Technology (BMB+F). The research summarized in Section 4 was conducted in the project READY of Saarland University's Collaborative Research Center on Resource-Adaptive Cognitive Processes, which is funded by the German Science Foundation (DFG).

References

Bernsen, N. O., & Dybkjær, L. (2001). Exploring natural interaction in the car. In *Proceedings of the CLASS Workshop on Natural Interactivity and Intelligent Interactive Information Representation.* Verona, Italy.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction.* Hillsdale, NJ: Erlbaum.

Dybkjær, L., & Bernsen, N. O. (2000). Usability issues in spoken dialogue systems. *Natural Language Engineering*, *6*(3–4), 243–271.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, *33*(4), 457–461.

Jameson, A. (2002). Adaptive interfaces and agents. In J. A. Jacko & A. Sears (Eds.), *Human-computer interaction handbook.* Mahwah, NJ: Erlbaum.

Jameson, A., Großmann-Hutter, B., March, L., Rummer, R., Bohnenberger, T., & Wittig, F. (2001). When actions have consequences: Empirically based decision making for intelligent user interfaces. *Knowledge-Based Systems*, *14*, 75–92.

Jameson, A., Schäfer, R., Weis, T., Berthold, A., & Weyrath, T. (1999). Making systems sensitive to the user's changing resource limitations. *Knowledge-Based Systems*, *12*, 413–425.

Jameson, A., & Schwarzkopf, E. (2002). Pros and cons of controllability: An empirical study. In P. De Bra (Ed.), *Adaptive hypermedia and adaptive web-based systems: Proceedings of AH 2002.* Berlin: Springer.

Kiefer, J. (2002). *Auswirkungen von Ablenkung durch gehörte Sprache und eigene Handlungen auf die Sprachproduktion [Effects on speech production of distraction through overheard speech and one's own actions].* Unpublished master's thesis, Department of Psychology, Saarland University.

Lindmark, K. (2000). *Interpreting symptoms of cognitive load and time pressure in manual input.* Unpublished master's thesis, Department of Computer Science, Saarland University, Germany.

Lyons, K., & Starner, T. (2001). Mobile capture for wearable computer usability testing. In T. Martin & V. Bahl (Eds.), *Proceedings of the Fifth International Symposium on Wearable Computers.* Los Alamitos, CA: IEEE Computer.

Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., & Wittig, F. (2001). Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In M. Bauer, P. Gmytrasiewicz, & J. Vassileva (Eds.), *UM2001, User modeling: Proceedings of the Eighth International Conference.* Berlin: Springer.

Olson, J. R., & Nilsen, E. (1987/1988). Analysis of the cognition involved in spreadsheet software interaction. *Human-Computer Interaction*, *3*(4), 309—349.

Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM*, *42*(11), 74–81.

Oviatt, S. (2000). Multimodal system processing in mobile environments. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology* (pp. 21–30).

Salvucci, D. D. (2001). Predicting the effects of in-car interfaces on driver behavior using a cognitive architecture. In J. A. Jacko, A. Sears, M. Beaudouin-Lafon, & R. J. Jacob (Eds.), *Human factors in computing systems: CHI 2001 conference proceedings* (pp. 120–127). New York: ACM.

Sodhi, M., Reimer, B., Cohen, J. L., Vastenburg, E., Kaars, R., & Kirschenbaum, S. (2002). On-road driver eye movement tracking using head-mounted devices. In *Proceedings of ETRA 2002: Eye Tracking Research and Applications Symposium* (pp. 61–68).

Wierwille, W. W. (1993). Visual and manual demands of in-car controls and displays. In B. Peacock & W. Karwowski (Eds.), *Automotive ergonomics* (pp. 299–320). London: Taylor and Francis.