A Study of Manual Gesture-Based Selection for the PEMMI Multimodal Transport Management Interface

Fang Chen¹, Eric Choi¹, Julien Epps¹, Serge Lichman¹, Natalie Ruiz¹,², Yu Shi¹, Ronnie Taib¹,², Mike Wu¹

¹Interfaces, Machines and Graphic Environments (IMAGEN)
National ICT Australia
Sydney, Australia

²School of Computer Science and Engineering
The University of New South Wales
Sydney, Australia

fang.chen@nicta.com.au

ABSTRACT

Operators of traffic control rooms are often required to quickly respond to critical incidents using a complex array of multiple keyboards, mice, very large screen monitors and other peripheral equipment. To support the aim of finding more natural interfaces for this challenging application, this paper presents PEMMI (Perceptually Effective Multimodal Interface), a transport management system control prototype taking video-based manual gesture and speech recognition as inputs. A specific theme within this research is determining the optimum strategy for gesture input in terms of both single-point input selection and suitable multimodal feedback for selection. It has been found that users tend to prefer larger selection areas for targets in gesture interfaces, and tend to select within 44% of this selection radius. The minimum effective size for targets when using ‘device-free’ gesture interfaces was found to be 80 pixels (on a 1280x1024 screen). This paper also shows that feedback on gesture input via large screens is enhanced by the use of both audio and visual cues to guide the user’s multimodal input. Audio feedback in particular was found to improve user response time by an average of 20% over existing gesture selection strategies for multimodal tasks.

Categories and Subject Descriptors

H.5.2 [Information Interfaces and Presentation]; User Interfaces–Input Devices and Strategies, Voice I/O, I.2.11 [Artificial Intelligence]; Multiagent systems.

General Terms


Keywords

Multimodal interaction, speech, manual gesture, multimodal fusion, multimodal output generation.

1. INTRODUCTION

During our recent informal studies of computer interfaces for incident handling in a local transport management centre, numerous problems with existing interfaces were encountered. These included input devices (e.g. mice, keyboards, touch screens) that were difficult to operate in conjunction with very large wall-mounted visual display units, inadequate search and selection mechanisms for geographical information systems, and generally cluttered and cognitively inefficient interfaces.

While aspects of this problem might be addressed using judicious modifications to the existing infrastructure, in this paper, we consider the possibilities for a new, multimodal approach to control room interfaces. The Perceptually Effective Multimodal Interface (PEMMI), introduced in this paper, aims to provide a more natural user experience for these operators via speech and gesture inputs and multimodal outputs, while attempting to preserve the robustness of interaction that they have come to expect from their existing interfaces.

Free-hand gesture as a mode of inputs is still a relatively new phenomenon in the fields of multimodal user interface and human-computer interaction (HCI) research. Despite insightful studies into their uses in command and control interfaces (e.g. [20]), there is a need for improved understanding of how they can best be applied to practical interfaces. Pen-based gestures are inherently more robust than computer vision-based manual gestures, offering clearly defined pen-up/pen-down events and exhibiting fewer tracking artefacts (such as jitter), due to the physical contact of the pen with the drawing surface. In this paper, strategies for manual gesture-based selection and feedback are evaluated, extending previous work on user preferred selection strategies [20]. The primary goal of [20] was the evaluation of selection strategies using vision and speech user interfaces on large-screens. Our experiments consider the minimum selection area for a target when using manual gesture interfaces; the use of multimodal vs. unimodal feedback in selection tasks; and user preferences for optimal selection area for a target.

Literature on Speech and Manual Gesture-based Interfaces

In applications where it is desirable for interaction to occur without touching or wearing any equipment, speech and gesture are two important modalities for which combined use dates back to the work of Bolt [2]. Bolt demonstrated that by fusing complementary information from deictic gestural input, speech input could become much more natural than had previously been considered feasible. The first known formal study of vision-based manual gesture and speech interface was carried out by [20] for a research prototype system dealing with crisis management in a simulated control room environment. Other studies have targeted systems that combine speech input with pointing gestures acquired from either magnetic trackers or visual markers [1, 9].

To our knowledge, there are no previous studies that examine the issues of applying multimodal user interfaces based on manual
gesture and speech to transport management. Although some human factors guidelines [12] have been developed for setting up a transport management centre, these guidelines are very general in nature and only consider factors related to traditional input and output devices, such as mouse and keyboard. Another related research effort [19] is the use of simulation-based learning for providing operators with an appreciation of the impact of incidents on traffic delay; however this focuses more on simulation and visualization for effective learning.

Literature on Vision-Based Manual Gesture Selection Strategies

Techniques for vision-based gesture tracking and recognition have proliferated in recent years (e.g. [11, 16, 20, 21]), although fewer researchers have experimented with the usability issues surrounding their application as an input mode. One previous study on vision-based gesture selection strategies [20] compared ‘point and wait’ (i.e. hold hand for a pre-determined fixed time to select current cursor position), ‘point and speak’ (i.e. speak to select current cursor position) and ‘point and shake’ (i.e. shake hand to select current cursor position) selection strategies. They found that users preferred the ‘point and wait’ strategy, which also gave the lowest error rate for the task tested despite requiring a slightly longer time-to-completion than other strategies. Their experiments also suggested that target position and size were two important factors in the design of vision-based gesture input, but did not give a figure for the minimum effective target size [20].

Literature on Multimodal Output Generation for Speech and Manual Gesture Interfaces

Human short-term memory is limited to just a few chunks of information [13], and the output generated by a given interface should manage the user’s cognitive load by optimizing the information flow over the available modalities. Information visualization and physiology research provides guidelines for multi-sensory output [15], for example the parallel exploitation of fovea and peripheral vision. A previous study of gesture-based selection [20] employed two types of unimodal visual feedback: (i) a cursor comprising an empty circle that is gradually filled from the centre during a ‘point and wait’ or ‘point and speak’ selection; and (ii) a multicolour cursor with a delayed linear fade designed to allow the user to see the history of their gesture trajectory during a ‘point and shake’ selection. Recent seminal studies on multimodal interaction have shown that users have a clear preference for multimodal interaction under situations of higher cognitive load [18], and this leads us to believe that transport management system operators may benefit from some aspects of the proposed PEMMI interface described herein.

Goals of the Current Study

The main objectives of the reported study were:

- To examine the lower limit of target sizes for gesture input, users’ preferences with respect to target sizes for different initial target size and spacing configurations;
- To evaluate the relative impact of multimodal feedback (particularly auditory vs. visual feedback) on the usability of the hand paused (‘point and wait’) gesture input.

2. THE PEMMI SPEECH AND GESTURE INTERFACE

The components of the PEMMI interface are shown in Figure 1. The entire system, including a mock transport management application, was implemented on two desktop PCs, networked using JADE (Java Agent Development Framework, http://jade.tilab.com/). Two types of messages flow through the system: a command message that requests a specific recipient to perform some tasks, and an event message that is sent to any interested recipients. For example, when the dialog management module receives a speech recognition engine completed event, it will issue a command to the output generation module to tell the user that speech input is currently inactive.

![Figure 1: Overview of the PEMMI interface.](http://www.example.com/figure1.png)

2.1 Gesture Input

Some previous gesture recognition implementations have been achieved using gloves or other body-worn aids (e.g. [1, 9]). Our approach dispenses with the inconvenience of these by tracking the user’s hand, using video-based skin colour detection. The skin colour extraction method employed in PEMMI is based on the Rule Induction algorithm described in [8], which, despite being a simple skin colour model, nevertheless outperforms other approaches like skin probability maps in most practical cases. Gesture tracking using disparity maps [11] or stereo cameras [21] are promising recent approaches, but may only give marginal improvements over single-camera skin colour detection once pattern recognition and tracking techniques are optimized.

In this work, a single hand was tracked and its location was used to point at locations on the large screen in order to select them. Three types of gesture events were recognized: raising the hand from its resting position (HAND_UP), pausing the hand during a pointing gesture (HAND_PAUSE), and lowering the hand to its resting position (HAND_DOWN). The hand up/down events were used to determine the beginning and end of a gesture input, while the hand paused events were used for single and multiple selection (the latter was based on the hand’s trajectory between two consecutive hand paused events), as summarized in Table 1. Note that the hand up events were also used to activate speech recognition.
In terms of choice of gesture-based selection strategy, the ‘hand paused’ selection approach used in PEMMI was informed by the ‘point and wait’ selection strategy that was found to provide superior performance relative to alternative strategies when evaluated for a similar application in [20]. In the application prototype, we found that a 500 ms delay was suitable for this style of selection, while 300 ms was used in [20]. Our use of an area selection strategy (described in Table 1 above) does not seem to have appeared in previous manual gesture prototypes, such as [20]. By way of comparison with pen-based gesture input, there is no need for a ‘hand paused’ approach, since this is replaced by the pen-up/pen-down inputs inherent in this input mode. In [5], where 3D gesture input is obtained via a PinchGlove, pen-up/pen-down is explicitly signalled using the PinchGlove.

2.2 Speech Input
A commercial command-and-control speech recognition component and Bluetooth wireless headset were used to provide robust recognition of a core set of critical commands. A small set of vocabulary command phrases was employed, which was informally observed to give near-perfect recognition accuracy. In order to deal with hesitations and out-of-vocabulary words in speech input, a set of garbage models was added to the speech recognizer to help capture non-meaningful speech utterances. A simple rule-based keyword spotter was used to extract the semantics of the spoken commands from the speech utterances.

2.3 Multimodal Fusion
Since the speech and gesture inputs in PEMMI contain predominantly complementary information, a semantic (late) fusion component [4] was employed. Informed by results from a previous study on integration patterns [7] showing a high occurrence of temporal overlap between speech and manual gesture inputs, the fusion component design assumes a high degree of speech/gesture simultaneity.

For purposes of speech recognition robustness, speech input does not commence until the user has raised his/her hand into the field of view of the camera. Thereafter, the user may select one or more items by pausing his/her hand as described above and optionally speaking. If no selection is desired, the user can lower his/her arm before or while speaking the command. The fusion component combines the two inputs once they are both complete. If the user has selected an area (i.e. multiple items) then the speech input is required to resolve the type of items required by the user.

2.4 Output Generation
Multimodal output generation is provided via map-based graphical output together with location-based artefacts such as camera or police station icons, an animated avatar with speech synthesis and earconas. A rabbit avatar gives visual feedback on the status of the speech and gesture recognition modules: dark glasses are worn when gesture recognition is disabled, and the long ears rise when the speech recognizer is listening to voice inputs.

Geographical data is graphically rich and complex, hence using an avatar for deictic purposes as proposed in the literature [6] has some drawbacks: overlaying an avatar on a map, or using arrows or similar artefacts impedes the vision of objects in the periphery of the focus of attention. However, the peripheral region of the retina is good at detecting movement [3], so reducing the avatar contribution to basic movement (e.g. raising or lowering long ears), allows users to note that the speech recognition is on, for example, while maintaining their attention on the primary task. Deictic functionality usually implies dynamic placement of the avatar on the screen depending on the location of the focus. Even well-managed, such movements contribute to a useless overload of the cognitive capacity of the operator. Thus, in the context of transport management, we fixed the position of the avatar in a corner of the screen and reduced its graphical attributes to moving ears and dark glasses. A speech bubble providing transcripts of the synthetic text-to-speech output is also attached to the avatar, but all users reported they never read its content.

2.5 Physical Layout
Figure 2 shows the physical implementation of PEMMI. The user stands about 1.5 metres in front of a large rear-projection screen measuring 2×1.5 metres. A low-cost web camera mounted on a tripod is used to capture the hand gestures, about a metre away from the arm. A small lamp is used to improve the lighting around the camera. The Bluetooth microphone is not seen in Figure 2 as it was worn on the left ear.

---

1 Automatic Speech Recognition is activated upon HAND_UP and stops automatically. HAND_DOWN occurs naturally at the end of a move.

---

Table 1: Types of Speech and Gesture Input.

<table>
<thead>
<tr>
<th>Input Type</th>
<th>Gesture Sequence Possible speech</th>
<th>Example Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech only</td>
<td>HAND_UP, HAND_DOWN “Alexandria”</td>
<td>Show map of specified suburb</td>
</tr>
<tr>
<td>Speech and gesture</td>
<td>HAND_UP, HAND_PAUSE, HAND_DOWN “Play this”</td>
<td>Select single landmark at specified point</td>
</tr>
<tr>
<td>Speech and gesture</td>
<td>HAND_UP, HAND_PAUSE, HAND_DOWN “Show cameras in this area”</td>
<td>Select group of landmarks within specified area</td>
</tr>
</tbody>
</table>
3. METHODS

3.1 Subjects
Twenty unpaid volunteers participated in the experiments, fifteen males and five females, between the ages of 19 and 50. All participants were right-handed. A different set of eight volunteers participated in the Feedback and Reaction Times experiment.

3.2 Apparatus
The setup for the experiment was physically very similar to the implementation shown in Figure 2, except that instead of the transport management application, a series of simple targets were displayed, as described in Section 3.3. The system comprised a wall-sized rear projection screen and a web cam mounted on a tripod near the user’s right arm, within comfortable reach of their right (pointing) hand. The gesture tracking software was run in real time (giving users instantaneous position feedback), based upon the implementation described in Section 2.1. The screen resolution used throughout the experiments is 1280x1024 pixels.

3.3 Experimental Design
Subjects were allocated around 20 minutes to complete all three tasks. A short questionnaire recording basic subjects’ details was completed before starting the experiments. The first task was very simple, and was used to familiarize participants with gesture input, since almost no participants had used gesture input previously. In total, three different tasks were used to gather various types of information on how gesture is used as a selection mechanism. The first task involved pointing and selection of the centre of a circular target. The second task began with user calibration of a selection area, followed by the sequential selection of circular targets in different configurations on the screen. The final task aimed to gather reaction time data for audio and/or visual feedback or no feedback at all in response to making a selection.

Selection Granularity (Minimum Selection Distance)
Subjects were presented with a black circular target in the centre of the screen, 20 pixels in radius. They were requested to point to the centre of the target and pause to select it. The trajectory of the gesture was recorded, and from this, two points were recorded: (i) the point in the trajectory with the minimum distance to the centre, and (ii) the distance from the centre of the target at the time of selection. This task was repeated four times, and was considered sufficiently simple to serve as a gesture practice task for the remainder of the study.

Target Spacing and Radius
Six circular targets were configured in a ring formation as shown in Figure 3 and subjects were instructed to select them in a fixed sequence, as described in [20], based on the International Organization for Standardization (ISO) Standard 9241-9: Requirements for non-keyboard input devices [10]. This test configuration was repeated 15 times, with 5 different distances between targets, and 5 different target sizes, with radii ranging from 9 to 105 pixels, as seen in Table 2.

<table>
<thead>
<tr>
<th>Target radius (pixels)</th>
<th>9</th>
<th>20</th>
<th>40</th>
<th>70</th>
<th>105</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target spacing (pixels)</td>
<td>220</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The first part of each test required the subject to ‘calibrate’ the selection area for each target. They were asked to designate the radius of a concentric circle around a target to define an area of selection they felt comfortable with for the purposes of reliably selecting that target, which could be anywhere outside the target or within the target. The proposed boundary of the selection area was shown with a red outline in Figure 4. This ‘calibration’ was made by the subject using gesture. Multiple attempts at calibration were permitted, and only once they were satisfied was the calibration confirmed by the experimenter and the measurement taken. The rationale behind the user-calibration method was to gain an insight into user preferences for target size with a device-free gesture pointing device.

Feedback and Reaction Times
In this task, the subjects were requested to perform five repetitions of the selection of a predefined pair of targets for each of four given types of feedback. Participants were requested to move their hands as soon as they felt a pause had been detected. A pause could be recognized through the presence of either audio and/or visual feedback. Audio feedback consisted of a short beep, while the visual feedback comprised a series of growing concentric ripples emanating from the location of the hand pause, over a total of two seconds. Inaccurate or failed inputs were discarded. The reaction time between the beginning of the beep and/or the first ripple of the visual output until the next significant move of the hand (i.e. greater than the cumulated camera jitter and trembling during the pause) was then measured. This experiment was repeated for four types of feedback, presented in the following order: audio-visual, audio only, visual only, no feedback (subjects were required to detect the pause without any cue).
4. RESULTS

4.1 Selection Granularity
This experiment aimed to measure the granularity (minimum selection distance) of gesture selections by monitoring:

- The point in a subject’s gesture trajectory that was closest to the centre of the target; and
- The distance of the actual point selected from the centre of the target.

As seen in Figure 5, where the results are averaged across all four repetitions, participants were consistently able to move more closely to the target centre than they were able to select. This is explained by both the difficulty in keeping the hand entirely stationary during the 0.5s selection time and the jitter in cursor position experienced during this interval. From this experiment, the target radius that can accommodate the actual distance at selection appears to be in the range of 15 to 40 pixels, depending on the individual user. Note that selection times are not taken into consideration here, meaning that some subjects may have performed their selection more quickly than others.

![Figure 5: Average distances from the centre of the target: the closest point the subject came to the centre (black) and the distance at which selection actually occurred (white).](image)

4.2 Target Spacing and Radius
The second task aimed to gain insight into user preferences for the required selection area around a target. The calibration distances, i.e. the maximum preferred selection distances from the centres of targets indicated by subjects are shown as a function of both target spacing and radii in Figures 6 and 7 respectively. In Figure 6, experimental results from the first column of Table 2 are given, while in Figure 7, experimental results from the first row of Table 2 are given, averaged across all subjects in each case. A one-way ANOVA analysis showed all these differences to be significant ($p < 0.001, F(4,90) = 11.58$).

It was expected that for a given target size, subjects would increase the selection area around the target as the distance between targets increased, in order to facilitate the selection as suggested by Fitts’ law. It was thought that the lower limit for the selection area would be bounded by the space available between targets and the upper limit would be bounded by subjects’ perceptions that they had defined a sufficiently large selection area. Figure 6 seems to confirm this hypothesis, with a monotonically increasing relationship between the calibration radius and spacing between targets, tending towards a more horizontal asymptote for larger spacings. Using Fitts’ Index of Difficulty $I_d$ as a coarse indicator, we observed that the selected calibration sizes result in a very mild $I_d$ increase when compared to the $I_d$ derived from the actual target size, confirming our expectations.

![Figure 6: Average calibration radius selected by subjects vs. spacing between the centres of each target and corresponding 95% confidence intervals (target radius = 9 pixels).](image)

It was also expected that subjects would choose a selection area proportional to the size of the target, with a lower boundary representing an area where the subjects could make gesture selections reasonably quickly. As can be seen in Figure 7, there is an almost linear relationship between the calibration radius and the target size. For target sizes smaller than a 20 pixel radius, a lower threshold of 40 pixels applies for the calibration radius. These values were found to be significantly different by a one-way ANOVA ($p<0.001, F(4,90) = 10.451$).

![Figure 7: Average calibration radius selected by subjects vs. the radius of each target and corresponding 95% confidence intervals (target spacing = 220 pixels).](image)
Another hypothesis motivating this experimental work was that subjects would only use a proportion of the area indicated by the calibration. To verify this, the actual selection distances, i.e. the distance from the centre of the required target at the time of selection, were graphed as a function of both target spacing and radius in Figures 8 and 9 respectively. In Figure 8, experimental results from the first column of Table 2 are given, while in Figure 9, experimental results from the first row of Table 2 are given, averaged across all subjects in each case. Figure 8 shows that the actual selection distance (solid) follows the trend of the calibration distance (dashed), but is much shorter than the latter. A similar characteristic is observed in Figure 9 for different target radii. The differences in average distance to the centre at time of selection, for different spacings, were found to be significant by a one-way ANOVA ($p<0.001$, $F(4,90) = 9.84$).

In Figure 10, experimental results from target radius 9 and the full set of target spacings (first column of Table 2) are given. When targets are small and very close together, difficult fine grained movement is required to select the targets. The task becomes easier and is completed faster when targets are placed a further apart. However, when the targets are far from one another, subjects need to execute a combination of coarse and fine grained movements: first to reach the vicinity of the target, then smaller movements and a pause to reach the centre of the target, once again increasing the completion time.

In Figure 11, experimental results from target spacing 220 pixels and the full set of radius sizes (first row of Table 2) are given, averaged across all subjects in each case. The completion times for the selection of a single target were found to mainly vary between about 2.8 and 5.6 seconds ($10^6$ and $90^6$ percentile values respectively). Smaller targets were harder to select, while larger targets were reached faster, as the distance to traverse from one to another is shorter, and no fine-grained movement is required.
The distances of the selection points from target centres as a function of the calibration distances that subjects selected for the same experiment are shown in Figure 12, across all subjects and all experiment configurations listed in Table 2. Linear regression analysis of these points revealed a gradient of 0.44, indicating that after subjects chose their preferred maximum selection distance during calibration, their actual selection distances during the ensuing tasks were generally around half of this maximum. 95% of all responses fell within a selection distance to calibration distance ratio of between 0.20 and 0.76. For individual users this range was even narrower, for example 0.36 to 0.59 in one user.

5.2 Target Spacing and Radius

In this experiment, it was expected that users would define large selection areas for smaller targets, and relatively smaller selection areas for larger targets. The size of the selection area was expected to be influenced by the proximity of other targets to each other in the configuration; if targets are closer together, the appropriate selection area would be confined to the target limits. We hypothesised that the calibration radius would always be greater than or equal to the target size, tending towards equality for larger sizes. It appeared that a subset of subjects consistently preferred calibration radii smaller than the target size, while a different subset consistently preferred calibration radii greater than the target size. As observed from Figure 7, on average, the subjects preferred a minimum calibration radius of about 40 pixels. Further studies will be needed to determine the impact of user perception and preferences on their gesture selection behaviour.

As expected, participants selected smaller preferred target radii (calibration distances) as both the target spacing and radii were decreased, as seen in Figures 6 and 7. Likewise, subjects performed their selections closer to the target centres as both the target spacing and radii were decreased, as expected. These results confirm the effect of target radius observed in [20], but also confirm the intuition that the target spacing is important.

5.3 Task Completion Times

Task completion times were long (around 3 times longer than in [20]) primarily due to a long response lag, and a slightly longer pause timeout. While the implementation could be improved to reduce this response lag, further studies will be required to determine the impact of response lags on overall user performance. An interesting pitfall was observed in the system settings: subjects tended to perform coarse movements over large screen distances, then “wait” for the pointer to ‘catch up’, before making finer grained movements towards the target centre. If the response lag is greater than or equal to the pause timeout, unwanted selections may result. The system response lag, measured as the delay between the subject and corresponding pointer movements on the screen, was re-assessed by manually monitoring a 25-frame per second DV video shot of the system in activity. The pointer was observed to begin moving about 0.45sec following the subject’s hand, and completes a move 0.78sec after the hand, regardless of the complexity of the subject’s movement.

5.4 Reaction Times

Unimodal feedback outperforms multimodal output in this instance: given the visual demand of the task, the visual sub-processing in working memory may become relatively overloaded with the additional visual cues; feedback is thus optimally received through the less taxed auditory sub-processor [14]. When asked about the feedback strategies, half of the subjects who participated in this experiment found the visual feedback distracting, while no participants took issue with the audio feedback.

5.5 Design Implications

The target size for vision-based gesture selection should not be less than 80 pixels in width or height relative to a 1280x1024 screen, at least to accommodate the users in this study. There appears to be a strong correlation between users’ preferred maximum selection distance (calibration distance) and the

<table>
<thead>
<tr>
<th>Audio-visual</th>
<th>Audio only</th>
<th>Visual only</th>
<th>No feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>624 ms</td>
<td>592 ms</td>
<td>735 ms</td>
<td>808 ms</td>
</tr>
</tbody>
</table>

Figure 12: Distance at time of selection vs. calibration distance, for all subjects over all experiments.

4.3 Feedback and Reaction Times

The reaction times, averaged across all eight participants, for each feedback strategy described in the previous section are given in Table 3. Since the subjects were all tested under the four conditions, a one-way repeated-measures ANOVA was conducted. It was found that the differences between the feedback conditions were significant (p < 0.001, F(3,21) = 8.276). A post-hoc Tukey HSD test (p < 0.05) revealed significant differences between all pairs of reaction times for the different strategies except between audio-visual and audio, and between visual and no feedback. Audio feedback produced the fastest reaction times for five subjects, while audio-visual feedback gave the fastest responses for the remaining three.

Table 3. Mean user reaction times to different types of feedback for a hand paused (‘point and wait’) gesture task.

5. DISCUSSION

5.1 Selection Granularity

Our data suggest that there is some degree of variation between individuals – the same subjects whose gesture trajectories travelled near to the target centres also made their selections closer to the target centres. The results shown in this paper must be regarded in the context of our implementation; other implementations may produce different results.
distance they generally used for selection. Asking users to select their preferred target radius, and adapting the interface to this seems a promising approach, since all subjects studied consistently performed selection at distances between 20% and 76% (typically 44%) of this preferred radius. Similar arguments for user adaptation can be found elsewhere in the multimodal literature (e.g. [18]). The results also support the use of audio feedback as a more effective alternative to visual cues in gestural interfaces, in particular, for eyes-busy tasks.

6. CONCLUSION
This paper has presented a speech and manual gesture-based multimodal prototype interface for interaction with graphical information on large screens, currently under consideration for transport management purposes. An investigation into the target sizes, spacing and feedback mechanisms for gesture-based input has been reported. This study indicates that a suitable minimum target size might be 80 pixels (on a 1280x1024 screen). A further result of the study was that after being asked to indicate their preferred selection radius around a target, subjects’ selections consistently occurred at distances of around 44% of these preferred radii. This suggests that a robust means of accommodating different users is to vary the sizes of the targets (and perhaps their spacing if needed), a degree of personalization rarely seen in existing mouse- or pen-based interfaces. Most of the relationships observed between calibration radii, selection distances and target spacings/radii concurred with expectations. Evaluation of audio and visual cues for feedback on gesture-based selection indicates that audio signals are a viable alternative to visual feedback, which may be helpful in eyes-busy applications.

The prototype described herein represents a first step for the longer-term PEMMI project. Future work will examine such issues as accuracy of speech and gesture-based task completion under different time or task constraints, adaptation of pause length in hand pause-based selection to accommodate individual user preferences, and the use of two-hand gestures for human-computer interaction.

7. ACKNOWLEDGMENTS
The authors would like to thank Chris Madden for coordinating the large-screen lab construction and Dr. Keith Nesbitt for helpful conversations on multi-sensory output.

8. REFERENCES