MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Enabling Interaction with Single User Applications through Speech and Gestures on a Multi-User Tabletop

Edward Tse, Chia Shen, Saul Greenberg, Clifton Forlines

TR2005-130 December 2005

Abstract

Co-located collaborators often work over physical tabletops with rich geospatial information. Previous research shows that people use gestures and speech as they interact with artefacts on the table and communicate with one another. With the advent of large multi-touch surfaces, developers are now applying this knowledge to create appropriate technical innovations in digital table design. Yet they are limited by the difficulty of building a truly useful collaborative application from the ground up. In this paper, we circumvent this difficulty by: (a) building a multimodal speech and gesture engine around the Diamond Touch multi-user surface, and (b) wrapping existing, widely-used off-the-shelf single-user interactive spatial applications with a multimodal interface created from this engine. Through case studies of two quite different geospatial systems Google Earth and Warcraft III we show the new functionalities, feasibility and limitations of leveraging such single-user applications within a multi user, multimodal tabletop. This research informs the design of future multimodal tabletop applications that can exploit single-user software conveniently available in the market. We also contribute (1) a set of technical and behavioural affordances of multimodal interaction on a tabletop, and (2) lessons learnt from the limitations of single user applications.

AVI 2006 (Advanced Visual Interfaces)

Enabling Interaction with Single User Applications through Speech and Gestures on a Multi-User Tabletop

Edward Tse[1,2], Chia Shen[1], Saul Greenberg[2] and Clifton Forlines[1]
[1]Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA, 02139, USA, +1 617 621-7500
[2]University of Calgary, 2500 University Dr. N.W. Calgary, Alberta, T2N 1N4, Canada +1 403 220-6087
[shen, forlines]@merl.com and [tsee, saul]@cpsc.ucalgary.ca

ABSTRACT
Co-located collaborators often work over physical tabletops with rich geospatial information. Previous research shows that people use gestures and speech as they interact with artefacts on the table and communicate with one another. With the advent of large multi-touch surfaces, developers are now applying this knowledge to create appropriate technical innovations in digital table design. Yet they are limited by the difficulty of building a truly useful collaborative application from the ground up. In this paper, we circumvent this difficulty by: (a) building a multimodal speech and gesture engine around the Diamond Touch multi-user surface, and (b) wrapping existing, widely-used off-the-shelf single-user interactive spatial applications with a multimodal interface created from this engine. Through case studies of two quite different geospatial systems – Google Earth and Warcraft III – we show the new functionalities, feasibility and limitations of leveraging such single-user applications within a multi user, multimodal tabletop. This research informs the design of future multimodal tabletop applications that can exploit single-user software conveniently available in the market. We also contribute (1) a set of technical and behavioural affordances of multimodal interaction on a tabletop, and (2) lessons learnt from the limitations of single user applications.

Categories and Subject Descriptors
H5.2 [**Information interfaces and presentation**]: User Interfaces – Interaction Styles.

General Terms
Design, Human Factors

Keywords
Tabletop interaction, visual-spatial displays, multimodal speech and gesture interfaces, computer supported cooperative work.

1. INTRODUCTION
Traditional desktop computers are unsatisfying for highly collaborative situations involving multiple co-located people exploring and problem-solving over rich spatial information. These situations include mission critical environments such as military command posts and air traffic control centers, in which paper media such as maps and flight strips are preferred even when digital counterparts are available [4][5]. For example, Cohen et. al.'s ethnographic studies illustrate why paper maps on a tabletop were preferred over electronic displays by Brigadier Generals in military command and control situations [4]. The 'single user' assumptions inherent in the electronic display's input device and its software limited commanders, as they were accustomed to using multiple fingers and two-handed gestures to mark (or pin) points and areas of interest with their fingers and hands, often in concert with speech [4][16].

AVI '06, May 23-26, 2006, Venezia, Italy.

While there are many factors promoting rich information use on physical tables over desktop computers, e.g., insufficient screen real estate and low image resolution of monitors, an often overlooked problem with a *personal* computer is that most digital systems are designed within single-user constraints. Only one person can easily see and interact with information at a given time. While another person can work with it through turn-taking, the system is blind to this fact. Even if a large high resolution display is available, one person's standard window/icon/mouse interaction – optimized for small screens and individual performance – becomes awkward and hard to see and comprehend by others involved in the collaboration [12].

For a computer system to be effective in such collaborative situations, the group needs at least: (a) a large and convenient display surface, (b) input methods that are aware of multiple people, and (c) input methods that leverage how people interact and communicate over the surface via gestures and verbal utterances [4][18]. For point (a), we argue that a *digital tabletop display* is a conducive form factor for collaboration since it lets people easily position themselves in a variety of collaborative postures (side by side, kitty-corner, round table, etc.) while giving all equal and simultaneous opportunity to reach into and interact over the surface**.** For points (b+c), we argue that *multimodal gesture and speech input* benefits collaborative tabletop interaction: reasons will be summarized in Section 2.

The natural consequence of these arguments is that researchers are now concentrating on specialized multi-user, multimodal digital tabletop applications affording visual-spatial interaction. However, several limitations make this a challenging goal:

1. **Hardware Limitations**. Most touch-sensitive display surfaces only allow a single point of contact. The few surfaces that do provide multi-touch have serious limitations. Some, like SmartSkin [20], are generally unavailable. Others limit what is sensed: SmartBoard's DViT (www.smarttech.com/dvit) currently recognizes a maximum of 2 touches and the touch point size, but cannot identify which touch is associated with which person. Some have display constraints: MERL's DiamondTouch [6] identifies multiple people, knows the areas of the table they are touching, and can approximate the relative force of their touches; however, the technology is currently limited to front projection and their surfaces are

relatively small. Consequently, most research systems limit interaction to a single touch/user, or by having people interact indirectly through PDAs, mice, and tablets (e.g., [16]).

2. **Software Limitations.** It is difficult and expensive to build a truly useful collaborative multimodal spatial application from the ground up (e.g., Quickset [5]). As a consequence, most research systems are 'toy' applications that do not afford the rich information and/or interaction possibilities expected in well-developed commercial products.

The focus of this paper is on wrapping existing single user geospatial applications within the multi-user, multimodal tabletop setting. Just as screen/window sharing systems let distributed collaborators share views and interactions with existing familiar single user applications [9], we believe that embedding familiar single-user applications within a multi-user multimodal tabletop setting – if done suitably – can benefit co-located workers.

The remainder of this paper develops this idea in three ways. First, we analyze and summarize the behavioural foundations motivating why collaborators should be able to use both speech and gestures atop tables. Second, we briefly present our Gesture Speech Infrastructure used to add multimodal, multi user functionality to existing commercial spatial applications. Third, through case studies of two different systems – Google Earth and Warcraft III – we analyze the feasibility and limitations of leveraging such single-user applications within a multi-user, multimodal tabletop.

2. BEHAVIOURAL FOUNDATIONS

This section reviews related research and summarize them in the form of a set of behavioural foundations.

2.1 Individual Benefits

Proponents of multimodal interfaces argue that the standard windows/icons/menu/pointing interaction style does not reflect how people work with highly visual interfaces in the everyday world [4]. They state that the combination of gesture and speech is more efficient and natural. We summarize below some of the many benefits gesture and speech input provides to individuals.

Deixis: speech refined by gestures. Deictic references are speech terms ('this', 'that', etc.) whose meanings are qualified by spatial gestures (e.g., pointing to a location). This was exploited in the Put-That-There multimodal system [1], where individuals could interact with a large display via speech commands qualified by deictic reference, e.g., "Put that…" (points to item) "there…" (points to location). Bolt argues [1] and Oviatt confirms [18] that this multimodal input provides individuals with a briefer, syntactically simpler and more fluent means of input than speech alone. Studies also show that parallel recognition of two input signals by the system yields a higher likelihood of correct interpretation than recognition based on a single input mode [18].

Complementary modes. Speech and gestures are strikingly distinct in the information each transmits, how it is used during communication, the way it interoperates with other communication modes, and how it is suited to particular interaction styles. For example, studies clearly show performance benefits when people indicate spatial objects and locations – points, paths, areas, groupings and containment – through gestures instead of speech [17][18][5][3]. Similarly, speech is more useful than gestures for specifying abstract actions.

Simplicity, efficiency, and errors. Empirical studies of speech/gestures *vs.* speech-only interaction by individuals performing map-based tasks showed that multimodal input resulted in more efficient use of speech (23% fewer spoken words), 35% less disfluencies (content self corrections, false starts, verbatim repetitions, spoken pauses, etc.), 36% fewer task performance errors, and 10% faster task performance [18].

Rich gestures and hand postures. Unlike the current deictic 'pointing' style of mouse-based and pen based systems, observations of people working over maps showed that people used different hand postures as well as both hands coupled with speech in very rich ways [4].

Natural interaction. During observations of people using highly visual surfaces such as maps, people were seen to interact with the map very heavily through both speech and gestures. The symbiosis between speech and gestures are verified in the strong user preferences stated by people performing map-based tasks: 95% preferred multimodal interaction *vs.* 5% preferred pen only. No one preferred a speech only interface [18].

2.2 Group Benefits

Spatial information placed atop a table typically serves as conversational prop to the group, creating a common ground that informs and coordinates their joint actions [2]. Rich collaborative interactions over this information often occur as a direct result of *workspace awareness:* the up-to-the-moment understanding one person has of another person's interaction with the shared workspace [11]. This includes awareness of people, how they interact with the workspace, and the events happening within the workspace over time. As outlined below, many behavioural factors comprising the *mechanics of collaboration* [19] require speech and gestures to contribute to how collaborators maintain and exploit workspace awareness over tabletops.

Alouds. These are high level spoken utterances made by the performer of an action meant for the benefit of the group but not directed to any one individual in the group [13]. This 'verbal shadowing' becomes the running commentary that people commonly produce alongside their actions. For example, a person may say something like "I am moving this box" for a variety of reasons:

- to make others aware of actions that may otherwise be missed,
- to forewarn others about the action they are about to take,
- to serve as an implicit request for assistance,
- to allow others to coordinate their actions with one's own,
- to reveal the course of reasoning,
- to contribute to a history of the decision making process.

When working over a table, alouds can help others decide when and where to direct their attention, e.g., by glancing up and looking to see what that person is doing in more detail [11].

Gestures as intentional communication. In observational studies of collaborative design involving a tabletop drawing surface, Tang noticed that over one third of all activities consisted of intentional gestures [23]. These intentional gestures serve many communication roles [19], including:

- pointing to objects and areas of interest within the workspace,
- drawing of paths and shapes to emphasise content,
- giving directions,
- indicating sizes or areas,
- acting out operations.

Deixis also serves as a communication act since collaborators can disambiguate one's speech and gestural references to objects and spatial locations [19]. An example is one person telling another person "This one" while pointing to a specific object. Deixis often makes communication more efficient since complex locations and object descriptions can be replaced in speech by a simple gesture. For example, contrast the ease of understanding a person pointing to this sentence while saying 'this sentence here' to the utterance 'the 4th sentence in the paragraph starting with the word deixis located in the middle of the column on page 3'.

Gestures as consequential communication. Consequential communication happens as one watches the bodies of other's moving around the work surface [22][19]. Many gestures are consequential *vs.* intentional communication. For example, as one person moves her hand in a grasping posture towards an object, others can infer where her hand is heading and what she likely plans to do. Gestures are also produced as part of many mechanical actions, e.g., grasping, moving, or picking up an object: this also serves to emphasize actions atop the workspace. If accompanied by speech, it also serves to reinforce one's understanding of what that person is doing.

Simultaneous activity. Given good proximity to the work surface, participants often work simultaneously over tables. For example, Tang observed that approximately 50-70% of people's activities around the tabletop involved simultaneous access to the space by more than one person [23]**.**

Gaze awareness. People monitor the gaze of a collaborator [13][14][11]. It lets one know where others are looking and where they are directing their attention. It helps one check what others are doing. It serves as visual evidence to confirm that others are looking at the right place or are attending one's own acts. It even serves as a deictic reference by having it function as an implicit pointing act. While gaze awareness is difficult to support in distributed groupware technology [14], it happens easily and naturally in the co-located tabletop setting [13][11].

2.3 Implications

The above points clearly suggest the benefits of supporting multimodal gesture and speech input on a multi-user digital table. This not only is a good way to support individual work over spatially located visual artefacts, but intermixed speech and gestures comprise part of the glue that makes tabletop collaboration effective. Taken all together, gestures and speech coupled with gaze awareness support a rich multi-person choreography of often simultaneous collaborative acts over visual information. Collaborators' intentional and consequential gesture, gaze movements and verbal alouds indicate intentions, reasoning, and actions. Participants monitor these acts to help coordinate actions and to regulate their access to the table and its artefacts. Participant's simultaneous activities promote interaction ranging from loosely coupled semi-independent tabletop activities to a tightly coordinated dance of dependant activities.

While supporting these acts are good goals for digital table design, they will clearly be compromised if we restrict a group to traditional single-user mouse and keyboard interaction. In the next section, we describe an infrastructure that lets us create a speech and gesture multimodal and multi-user wrapper around these single-user systems. As we will see in the following case studies, these afford a subset of the benefits of multimodal interaction.

3. GESTURE SPEECH INFRASTRUCTURE

Our infrastructure is illustrated in Fig. 1. A standard Windows computer drives our infrastructure software, as described below.

The table is a 42" MERL Diamond Touch surface [6] with a 4:3 aspect ratio; a digital projector casts a 1280x1024 pixel image on the table's surface. This table is multi-touch sensitive, where contact is presented through the DiamondTouch SDK as an array of horizontal and vertical signals, touch points and bounding boxes (Fig. 1, row 5). The table is also multi-user, as it distinguishes signals from up to four people. While our technology uses the Diamond Touch, the theoretical motivations, strategies developed, and lessons learnt should apply to other touch/vision based surfaces that offer similar multi user capabilities.

Speech Recognition. For speech recognition, we exploit available technology: noise canceling headset microphones for capturing speech input, and the Microsoft Speech Application Programmers' Interface (Microsoft SAPI) (Fig. 1, rows 4+5). SAPI provides an *n*-best list of matches for the current recognition hypothesis. Due to the one user per computer limitation in Microsoft SAPI, only one headset can be attached to our main computer. We add an additional computer for each additional headset, which collects and sends speech commands to the primary computer (Fig. 1, right side, showing a 2nd headset).

Gesture Engine. Since recognizing gestures from multiple people on a table top is still an emerging research area [25][26]**,** we could not use existing 3rd party gesture recognizers. Consequently, we developed our own Diamond Touch gesture recognition engine to convert the raw touch information produced by the DiamondTouch SDK into a number of rotation and table-size independent features (Fig. 1, rows 4+5 middle). Using a Univariate Gaussian clustering algorithm, features from a single input frame are compared against a number of pre-trained hand and finger postures. By examining multiple frames over time, we capture dynamic information such as a hand moving up or two fingers moving closer together or farther apart. This allows applications to be developed that understand both different hand postures and dynamic movements over the Diamond Touch.

Input Translation and mapping. To interact with existing single user applications, we first use the GroupLab WidgetTap toolkit [8] to determine the location and size of the GUI elements within

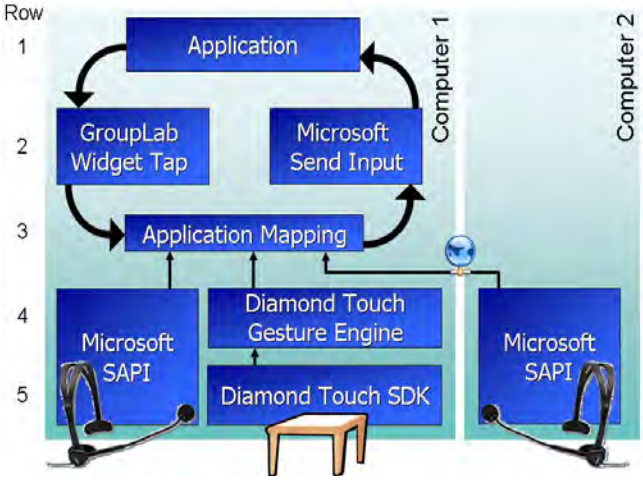

Figure 1. The Gesture Speech Infrastructure

it. We then use the Microsoft Send Input facility to relay the gesture and speech input actions to the locations of the mapped UI elements (Fig. 1, rows 1, 2 and 3). Thus speech and gestures are mapped and transformed into one or more traditional GUI actions as if the user had performed the interaction sequence via the mouse and keyboard. The consequence is that the application appears to directly understand the spoken command and gestures. Section 5.5 elaborates further on how this mapping is done. If the application allows us to do so, we also hide the user interface GUI elements so they do not clutter up the display. Of importance is that application source code is neither required nor modified.

4. GOOGLE EARTH and WARCRAFT III

Our case studies leverage the power of two commercial single user geospatial applications: Google Earth (earth.google.com) and Blizzard's Warcraft (www.blizzard.com/war3). The following sections briefly describe their functionality and how our multimodal interface interacts with them. While the remainder of this paper primarily focuses on two people working over these applications, many of the points raised apply equally to groups of three or four.

4.1 Google Earth

Google Earth is a free desktop geospatial application that allows one to search, navigate, bookmark, and annotate satellite imagery of the entire planet using a keyboard and mouse. Its database contains detailed satellite imagery with layered geospatial data (e.g., roads, borders, accommodations, etc). It is highly interactive, with compelling real time feedback during panning, zooming and 'flying' actions, as well as the ability to tilt and rotate the scene and view 3D terrain or buildings. Previously visited places can be bookmarked, saved, exported and imported using the places feature. One can also measure the distance between any two points on the globe.

Table 1 provides a partial list of how we mapped Google Earth onto our multimodal speech and gesture system, while Fig. 2 illustrates Google Earth running on our multimodal, multi user table. Due to reasons that will be explained in §5.4, almost all speech and gesture actions are independent of one another and immediately invoke an action after being issued. Exceptions are 'Create a path / region' and 'measure distance', where the system waits for finger input and an 'ok' or 'cancel' utterance (Fig. 1).

4.2 Warcraft III

Warcraft III is a real time strategy game. It implements a command and control scenario over a geospatial landscape. The landscape is presented in two ways: a detailed view that can be panned, and a small inset overview. No continuous zooming features are available like those in Google Earth. Within this setting, a person can create *units* comprising semi-autonomous characters, and direct characters and units to perform a variety of actions (e.g., move, build, attack). While Google Earth is about navigating an extremely large and detailed map, Warcraft is about giving people the ability to manage, control and reposition different units over a geospatial area.

Table 2 shows how we mapped Warcraft III onto speech and gestures, while Fig. 3 illustrates two people interacting with it on a table. Unlike Google Earth and again for reasons that will be discussed in §5.4, Warcraft's speech and gesture commands are often intertwined. For example, a person may tell a unit to attack, where the object to attack can be specified before, during or even after the speech utterance.

Table 1. The Speech/Gesture interface to Google Earth

Speech commands		Gesture commands	
Fly to *<place name>*	Navigates to location, eg., Boston, Paris	**One finger move / flick**	Pans map directly / continuously
Places *<place name>*	Flys to custom-created places, e.g., MERL	**One finger double tap**	Zoom in 2x at tapped location
Navigation panel	Toggles 3D Navigation controls, e.g., rotate	**Two fingers, spread apart**	Zoom in
Layer *<type>*	Toggles a layer, e.g., bars, banks	**Two fingers, spread together**	Zoom out
Undo layer	Removes last layer	**Above two actions done rapidly**	Continuous zoom out / in until release
Reorient	Returns to the default upright orientation	**One hand**	3D tilt down
Create a path *<points>***Ok**	Creates a path that can be travelled in 3D	**Five fingers**	3D tilt up
Tour last path	Does a 3D flyover of the previously drawn path	**Bookmark**	Pin + save current location
Create a region *<points>*	Highlight via semi-transparent region	**Last bookmark**	Fly to last bookmark
Measure Distance	Measures the shortest distances between two	**Next bookmark**	Fly to previous bookmark

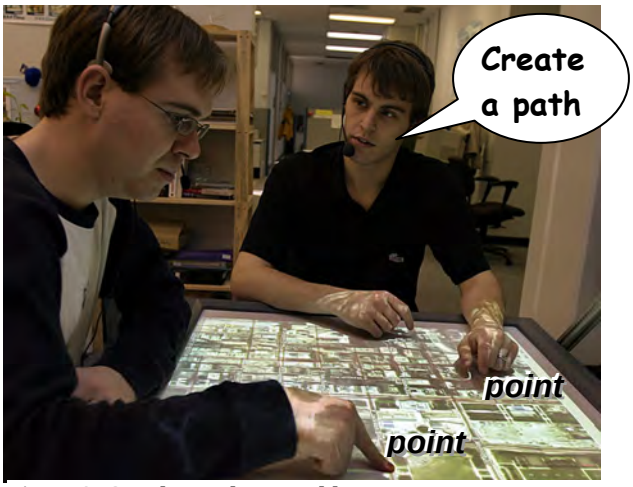

Figure 2. Google Earth on a table.

5. ANALYSIS and GUIDELINES

From our experiences implementing multi-user multi-modal wrappers for Google Earth and Warcraft III, we encountered a number of limitations that influenced our wrapper design, as outlined below. When possible, we present solutions to mitigate these limitations, which can also guide the design of future multi-user multi-modal interactions built atop single user applications.

This section is loosely structured as follows. The first three subsections raise issues that are primarily a consequence of constraints raised by how the single user application produces *visual output*: upright orientation, full screen views, and feedthrough. The remaining subsections are a consequence of

Table 2. The Speech/Gesture interface to Warcraft III

Speech commands		Gesture commands	
Unit <#>	Selects a numbered unit, e.g., one, two	**One hand**	Pans map directly
Attack / attack here [point]	Selected units attack a pointed to location	**One finger**	Selects units & locations
Build <object> **here [point]**	Build object at current location, e.g., farm, barracks	**Two fingers**	Context – dependant move or attack
Move / move here [point]	Move to the pointed to location	**Two sides of hand**	Select multiple workers in an area
[area] Label as unit <#>	Adds a character to a unit group	**Next worker**	Navigate to the next worker
Stop	Stop the current action		

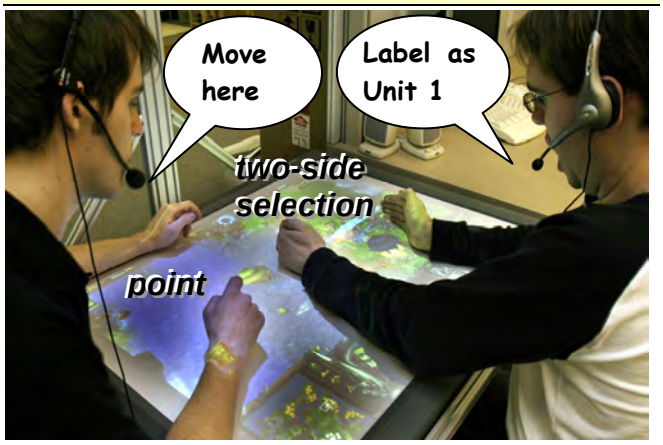

Figure 3. Two people interacting with Warcraft III.

constraints raised by the application consider *user input:* interacting speech and gestures, mapping, and turntaking.

5.1 Upright Orientation

Most single user systems are designed for an upright display rather than a table. Thus all display items and GUI widgets are oriented in a single direction usually convenient for the person seated at the 'bottom' edge of the display, but would be upside down for the person seated across from them. As illustrated in the upside down inset figure, a screenshot from Google Earth, problems introduced include text readability (but see [24]), difficulties in comprehending incorrectly oriented 3D views, inhibiting people from claiming ownership of work areas [15], and preventing people from naturally adjusting orientation as part of their collaborative process [15]. Similarly, the layout of items on the surface usually favors a single orientation, which has implications for how people can see and reach distant items if they want to perform gestures over them.

Warcraft III maintains a strictly upright orientation; while people can pan, they cannot rotate the landscape. Critical interface features, such as the overview map, are permanently positioned at the bottom left corner, which is inconvenient for a person seated to the right who wishes to navigate using the overview map. Google Earth has similar constraints: its navigation panel (exposed by a speech command) is at the very bottom, making its tilt GUI control awkward to use for anyone but the upright user. While Google Earth allows the map to be rotated, text labels atop the map are *not* rotated. In both systems, 3D perspective is oriented towards the upright user. A tilted 3D image is the norm in Warcraft III. While Google Earth does provide controls to adjust the 3D tilt of a building on the map, the viewpoint always remains set for the upright user.

Some of these problems are not solvable as they are inherent to the single user application, although people can choose to work side by side on the bottom edge. However, speech appears to be an ideal input modality for solving problems arising from input orientation and reach, since users can sit around any side of the table to issue commands (*vs.* reach, touch or type).

5.2 Full Screen Views

Many applications provide a working area typically surrounded by a myriad of GUI widgets (menus, palettes, etc.). While these controls are reasonable for a single user, multiple people working on a spatial landscape expect to converse over the scene itself. Indeed, one of the main motivations for a multimodal system is to minimize these GUI elements. Fortunately, many single user applications provide a 'full screen' view, where content fills the entire screen and GUI widgets are hidden. The trade-off is that only a few basic actions are allowed, usually through direct manipulation or keyboard shortcuts (although some applications provide hooks through accessibility APIs).

Because Warcraft III is designed as a highly interactive game, it already exploits a full screen view in which all commands are accessible through keyboard shortcuts or direct manipulation. Thus speech/gesture can be directly mapped to keyboard/mouse commands. In contrast, Google Earth contains traditional GUI menus and sidebars: 42% of the screen real estate is consumed by GUI items on a 1024x768 screen! While these elements can be hidden by toggling it into full screen mode, much of Google Earth's functionality is only accessible through these menus and sidebars. Our solution uses full screen mode, in which we map multimodal commands to action macros that first expose a hidden menu or sidebar, perform the necessary action on it (via WidgetTap and Send Input), and then hide the menu or sidebar (see §5.5). When this stream of interface actions is executed in a single step, the interface elements and inputs are hidden.

5.3 Feedback and Feedthrough

Feedback of actions is important for single user systems. *Feedthrough* (the visible consequence of another person's actions) is just as important if the group is to comprehend what another person is doing [7]. True groupware systems can be constructed to regulate the feedback and feedthrough so it is appropriate to the acting user and the viewing participants. Within single user systems, we can only use what is provided.

Fortunately, both Google Earth and Warcraft III are highly interactive, immediately responding to all user commands in a very visual and often compelling manner. Panning in both produces an immediate response, as does zooming or issuing a 'Fly to' command in Google Earth. Warcraft III visually marks all selections, re-enforcing the meaning of a gestural act. Warcraft III also gives verbal feedback. For example, if one says the 'Move

here' or 'Attack here' voice command and points to a location (Table 2), the units will respond with a prerecorded utterance such as "yes, master" and will then move to the specified location.

In both systems, some responses are animated over time. For example, 'Fly to, Calgary from a distant location will begin an animated flyover by first zooming out of the current location, flying towards Calgary, and zooming into the centre of the city. Similarly, panning contains some momentum in Google Earth, thus a flick gesture on the table top will send the map continually panning in the direction of the flick. In Warcraft III, if one instructs 'Unit one, build farm' <here>, it takes time for that unit to run to that location and to build the farm. These animations provide excellent awareness to the group, for the feedthrough naturally emphasises individual actions [12].

Animations over time also provide others with the ability to interrupt or modify the ongoing action. For example, animated flyovers, continuous zooming or continuous panning in Google Earth can be interrupted by a collaborator at any point by touching on the table surface. Similarly a 'stop' voice command in Warcraft III can interrupt any unit's action at any time.

Feedback, even when it is missing, is also meaningful as it indicates that the system is waiting for further input. For example, if one says 'Unit one move' to Warcraft III, the group will see unit one selected and a cross hair indicating that it is waiting for a location to move to, but nothing will actually happen until one points to the surface. This also provides others with the ability to interrupt, and even to take over the next part of the dialog (§5.6).

5.4 Interacting Speech and Gestures

Ideally, we would like to have the system respond to interacting and possibly overlapping speech and gesture acts, e.g., 'Put that' <*points to object*> 'there' <*points to place*> [1]. This is how deixis and consequential communication works. It may even be possible to have multiple people contribute to command construction through turn taking (see §5.6). However, the design of the single user application imposes restrictions on how this can be accomplished.

Google Earth only allows one action to be executed at a time; no other action can be executed until that action is completed. For example if a person performs simultaneous keyboard and mouse interactions only the keyboard commands will be performed. The design consequence is that we had to map most spoken and gestural actions into separate commands in Google Earth (Table 1). As mentioned, with the exception of the 'create a path/region' and 'measure distance' command, gestures and speech do not interact directly. Some gesture and speech commands move or zoom to a location. Other speech commands operate in the context of the current location, usually the center of the screen. For example, 'bookmark' only acts on the screen center; while a person can position the map so the location is at its center, they cannot say 'Bookmark' and point to a location off to the side.

In contrast, Warcraft III is designed to be used with the keyboard and mouse in tandem, i.e., it can react to keyboard and mouse commands simultaneously. This makes it possible to use intermixed speech and deixis for directing units. Our mapping uses speech in place of keyboard commands, and gesture in place of mouse commands, e.g., saying 'Unit 1, move here' while pointing to location.

By understanding the sometimes subtle input constraints of the single user application, a designer can decide if and where intermixing of speech and gestures via mapping is possible.

5.5 Mapping

Complementary Modes. Our behavioural foundations state that speech and gesture differ in their ability to transmit and communicate information, and in how they interact to preserve simplicity and efficiency [17][5][3]. Within Google Earth and Warcraft III (Tables 1 & 2), we reserve gestures primarily for spatial manipulations: navigation, deixis and selections. 'Abstract' commands are moved onto the speech channel.

Mapping of Gestures. Many systems rely on abstract gestures to invoke (i.e., mode change into) commands. For example, a two fingered gesture invokes an 'Annotate' mode in Wu's example application [25]. Yet our behavioural foundations state that people working over a table should be able to easily understand other people's rich gestural acts and hand postures as both consequential communication and as communicative acts. This strongly suggests that our vocabulary of postures and dynamics must reflect people's natural gestures as much as possible (a point also advocated in [25][26]).

Because we reserve gestures for spatial manipulations, very little learning is needed: panning by dragging one's finger or hand across the surface is easily understood by others, as is the surface stretching metaphor used in spreading apart or narrowing two fingers to activate discrete or continuous zooming in Google Earth. Pointing to indicate deictic references, and using the sides of two hands to select a group of objects in Warcraft III is also well understood [17][5][3]. Because most of these acts work over a location, gaze awareness becomes highly meaningful. However, the table's input constraints can restrict what we would like to do. For example, an upwards hand tilt movement would be a natural way to tilt the 3D map of Google Earth, but this posture is not recognized by the DiamondTouch table. Instead, we resort to a more abstract one hand / five finger gesture set to tilt the map up and down (Table 1).

Mapping of Speech. A common approach to wrapping speech atop single user systems is to do a 1:1 mapping of speech onto system-provided command primitives. This is inadequate for a multi-user setting: a person should be able to rapidly issue semantically meaningful commands to the table, and should easily understand the meaning of other people's spoken commands within the context of the visual landscape and their gestural acts. In other words, speech is intended not only for the control of the system, but also for the benefits of one's collaborators. If speech were too low level, the other participants would have to consciously reconstruct the intention of the user. The implication is that speech commands must be constructed so that they become meaningful 'alouds'.

Within Google Earth, we simplified many commands by collapsing a long sequential interaction flow into a macro invoked by a single well formed utterance (Table 1). For example, with a keyboard and mouse, flying to Boston while in full screen mode requires the user to: 1) use the tool menu to open a search sidebar, 2) click on the search textbox, 3) use the keyboard to type in 'Boston, MA' followed by the return key, and 4) use the tool menu to close the search sidebar. Instead, a person simply speaks the easily understood two-part utterance 'Fly to' 'Boston'. We

also created 'new' commands that make sense within a multimodal multi-user setting, but that are not provided by the base system. For example, we added the ability for anyone to undo layer operations (which adds geospatial information to the map) by creating an 'Undo Layer' command (Table 1). Under the covers, our mapping module remembers the last layer invoked and toggles the correct checkbox in the GUI to turn it off.

Intermixing of Speech and Gesture. We explained previously that a strength of multimodal interaction is that speech and gestures can interact to provide a rich and expressive language for interaction and collaboration. Because of its ability to execute simultaneous commands, Warcraft III provides a good example how speech and gesture can be mapped to interact over a single user application. Our Warcraft III speech vocabulary was constructed as easily understood phrases: nouns such as 'unit one', verbs such as 'move', action phrases such as 'build farm' (Table 2). These speech phrases are usually combined with gestures describing locations and selections to complete the action sequence. For example, a person may select a unit, and then say 'Build Barracks' while pointing to the location where it should be built. This intermixing not only makes input simple and efficient, but makes the action sequence easier for others to understand.

5.6 Turn taking

Single user applications expect only a single stream of input coming from a single person. In a multi-user setting, these applications cannot disambiguate what commands come from what person, nor can they make sense of overlapping commands and/or command fragments that arise from simultaneous user activities.

In shared window systems, confusion arising from simultaneous user input across workstations is often regulated through a *turn taking* wrapper interposed between the multiple workstation input streams and the single user application [9][10]. Akin to a switch, this wrapper regulates *user pre-emption* so that only one workstation's input stream is selected and sent to the underlying application. The wrapper could embody various turn taking protocols, e.g., explicit release (a person explicitly gives up the turn), pre-emptive (a new person can grab the turn), pause detection (explicit release when the system detects a pause in the current turn-holder's activity), queue or round-robin (people can 'line up' for their turns), central moderator (a chairperson assigns turns), and free floor (anyone can input at any time, but the group is expected to regulate their turns using social protocol) [10].

In the distributed setting of shared window systems, technical enforcement of turn taking is often touted since interpersonal awareness is inadequate to effectively use social mediation. Our two case studies reveal far richer opportunities for social regulation of turn-taking in tabletop multimodal environments.

Ownership through Awareness. We noticed that unlike distant-separated users of shared window systems, co-located tabletop users were aware of moment by moment actions of others and thus were far better able to use social protocol to mediate their interactions. Alouds arising from speaking into the headset let others know that one had just issued a command so they could reconstruct its purpose; thus people are unlikely to verbally overlap one another, or to unintentionally issue a conflicting command. Through consequential communication, people see that one is initiating, continuing or completing a gestural act; this strongly suggests one's momentary 'ownership' of the table and thus regulates how people time appropriate opportunities for taking over. The real time visual feedback and feedthrough provided by both Google Earth and Warcraft emphasises who is in control, what is happening, when the consequences of their act is completed, and when it is appropriate to intercede.

Interruptions. We noticed that awareness not only lets people know who is in control, but also provides excellent opportunities for interruptions. That is, a person may judge moments where they can stop, take over and/or fine-tune another person's actions. Eye gaze and consequential communication helps people mutually understand when this is about to happen, enabling cooperation *vs.* conflict. We already described how animations initiated by user actions (e.g., unit movement in Warcraft or the animated flyovers in Google Earth) can be stopped or redirected by a spoken command ('Stop') or a gestural command (touching the surface).

Assistance. Awareness also provides opportunities for people to offer assistance. Indeed, the interruptions mentioned above are likely a form of assistance, i.e., to repair or correct an action initiated by another person. Assistance also occurs when multiple people interleave their speech and gestures to compose a single command. For example, we previously mentioned in §5.5 how multi modal commands in Warcraft III are actually phrases, where phrases are chained together to compose a full command. As one person starts a command ('unit one', 'move') another can continue by pointing to the place where it should move to. Similarly, the 'create a path' and 'create a region' spoken commands in Google Earth expect a series of points: all members of the group can contribute these points through touch gestures.

The Mode problem. In spite of the above, people can only work within the current mode of the single user application. While one can take over (through turn taking) actions within a mode, two people cannot work in different modes at the same time. For example, in Warcraft III it is not possible for multiple people to control different units simultaneously.

In summary, while our experiences with our case studies suggest that social regulation of turn taking suffices for two people working over a multi modal, multi user tabletop (since the group has enough information to regulate themselves), there could be situations in which technical mediation is desired. Examples could include larger groups (to avoid accidental command overlap and interruptons), participants with different roles, or conflict situations. This proved fairly easy to do by incorporating a turn taking layer to the Application Mapping module in our infrastructure (Fig. 1). This module already knows which user is trying to interact with the system by touch or speech, and can detect when multiple people are contending for the turn. Decision logic or coordination policies [10][21] can then decide which input to forward to the application, and which to ignore (or queue for later). The logic could enforce turn taking policies at different levels of granularity.

- *Floor control* dictates turns at a person level, i.e., a person is in control of all interaction until that turn is relinquished to someone else.
- *Input control:* one input modality has priority over another modality, e.g., gesture takes priority over speech commands.
- *Mode control* enforces turn taking at a finer granularity. If the system detects that a person has issued a command that enters

a mode, it blocks or queues all other input until the command is complete and the mode is exited. For example, if a person opens the navigation panel or begins a tour flyover in Google Earth, all input is blocked until the flyover is completed.

- *Command control* considers turn taking within command composition. If the system detects that a person has issued a phrase initiating a command, it may restrict completion of that command to that person, e.g., if a person selects a character in Warcraft III, the system may temporarily block others from issuing commands to that character. Alternately, other people may be allowed to interleave a subset of command phrases to that character, e.g., while they can gesture to enter points via to Google Earth's 'Create a Path' command, only the initiator can complete that command with the spoken 'Ok'.

6. CONCLUSIONS

This paper described how and why we enabled speech and gestural interaction with commercial single user applications on a multi-user tabletop. By surveying the literature on groupware and multimodal interactions, we presented key behavioural affordances that motivate and inform the use of multimodal, multi user table top interaction. These behavioural affordances are applied in practice to implement two existing geospatial systems (Google Earth and Warcraft III) atop a common Gesture Speech Infrastructure. From our experiences, we derived a detailed but generalized analysis of issues and workarounds, which in turn provides guidance to future developers of this class of systems.

This work represents an important first step bringing multimodal multi-user interaction to a table display. By leveraging the power of popular single user applications, we bring a visual and interactive richness to table top interaction that can not be achieved by a simple research prototype. Consequently, demonstrations of our systems to the creators of Google Earth, real world users of geospatial systems including NYPD officers with the Real Time Crime Center, and Department of Defence members have evoked overwhelming positive and enthusiastic comments, e.g., "How could it be any more intuitive"?

For our next steps, we are studying 'true' multi-user, multimodal tabletop systems which will serve as stand-alone applications, and as an interactive layer placed atop single-user systems.

Illustrative video: Visit http://grouplab.cpsc.ucalgary.ca/tabletop/

Acknowledgements. We are grateful for the support from our sponsors: ARDA, NGA, NSERC, Alberta Ingenuity and iCORE.

7. REFERENCES

[1] Bolt, R.A., Put-that-there: Voice and gesture at the graphics interface. *Proc ACM Conf. Computer Graphics and Interactive Techniques Seattle*, 1980, 262-270.

[2] Clark, H. *Using language*. Cambridge Univ. Press, 1996.

[3] Cohen, P. Speech can't do everything: A case for multimodal systems. *Speech Technology Magazine*, 5(4), 2000.

[4] Cohen, P.R., Coulston, R. and Krout, K., Multimodal interaction during multiparty dialogues: Initial results. *Proc IEEE Int'l Conf. Multimodal Interfaces*, 2002, 448-452.

[5] Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L. and Clow, J., QuickSet: Multimodal interaction for distributed applications. *Proc. ACM Multimedia*, 1997, 31-40.

[6] Dietz, P. and Leigh, D. DiamondTouch: a multi-user touch technology. *Proc ACM UIST*, 2001, 219-226.

[7] Dix, A., Finlay, J. Abowd, G. and Beale, R. Human-Computer Interaction. 2nd ed. Prentice Hall, 1998.

[8] Greenberg, S. and Boyle, M. Customizable physical interfaces for interacting with conventional applications. *Proc ACM UIST*, 2002, 31-40.

[9] Greenberg, S., Sharing views and interactions with single-user applications. *Proc ACM COIS*, 1990, 227-237

[10] Greenberg, S. Personalizable groupware: Accommodating individual roles and group differences. *Proc ECSCW*, 1991,17-32,

[11] Gutwin, C., and Greenberg, S. The importance of awareness for team cognition in distributed collaboration. In E. Salas, S. Fiore (Eds) *Team Cognition: Understanding the Factors that Drive Process and Performance*, APA Press, 2004, 177-201.

[12] Gutwin, C. and Greenberg, S. Design for individuals, design for groups: Tradeoffs between power and workspace awareness. *Proc ACM CSCW*, 1998, 207-216

[13] Heath, C.C. and Luff, P. Collaborative activity and technological design: Task coordination in London Underground control rooms. *Proc ECSCW*, 1991, 65-80

[14] Ishii, H., Kobayashi, M. and Grudin, J. Integration of interpersonal space and shared workspace: ClearBoard design and experiments. *ACM TOIS*, 11 (4), 1993, 349-375.

[15] Kruger, R., Carpendale, M.S.T., Scott, S. and Greenberg, S. Roles of orientation in tabletop collaboration: Comprehension, coordination and communication. *J CSCW*, 13(5-6), 2004, 501-537.

[16] McGee, D.R. and Cohen, P.R., Creating tangible interfaces by augmenting physical objects with multimodal language. *Proc ACM Conf Intelligent User Interfaces*, 2001, 113-119.

[17] Oviatt, S. L. Ten myths of multimodal interaction, *Comm. ACM*, 42(11), 1999, 74-81.

[18] Oviatt, S. Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction* 12, 1997.

[19] Pinelle, D., Gutwin, C. and Greenberg, S. Task analysis for groupware usability evaluation: Modeling shared-workspace tasks with the mechanics of collaboration. *ACM TOCHI*, 10(4), 2003, 281-311.

[20] Rekimoto, J. SmartSkin: An infrastructure for freehand manipulation on interactive surfaces. *Proc ACM CHI*, 2002.

[21] Ringel-Morris, M., Ryall, K., Shen, C., Forlines, C., Vernier, F. Beyond social protocols: Multi-user coordination policies for co-located groupware. *Proc ACM CSCW*, 262-265, 2004.

[22] Segal, L. Effects of checklist interface on non-verbal crew communications, NASA Ames Research Center, Contractor Report 177639. 1994

[23] Tang, J. Findings from observational studies of collaborative work. *Int. J. Man-Machine. Studies.* 34 (2), 1991, 143-160.

[24] Wigdor, D., Balakrishnan, R. Empirical investigation into the effect of orientation on text readability in tabletop displays. *Proc ECSCW*, 2005.

[25] Wu, M., Shen, C., Ryall, K., Forlines, C., and Balakrishnan, R. Gesture registration, relaxation, and reuse for multi-point direct-touch surfaces. IEEE Int'l Workshop Horizontal Interactive Human-Computer Systems (TableTop). 2006.

[26] Wu, M. and Balakrishnan, R. Multi-finger and whole hand gestural interaction techniques for multi-user tabletop displays. *Proc ACM UIST*, 193-202. 2003.