An Exploration of Gesture-Speech Multimodal Patterns for Touch Interfaces

Prasenjit Dey, Sriganesh Madhvanath, Amit Ranjan, Suvodeep Das[†]
Hewlett-Packard Labs , 24, Salarpuria Arena, Adugodi, Hosur Road, Bangalore-560030, India
[†]3rd Floor, 267 KHB Colony, 5th Block, Kormangala, Bangalore-560095, India
{pdey, srig, amitr}@hp.com, suvodeepdas@gmail.com

ABSTRACT

Multimodal interfaces that integrate multiple input modalities such as speech, gestures, gaze, and so on have shown considerable promise in terms of *higher task efficiency, lower error rates and higher user satisfaction*. However, the adoption of such interfaces for real-world systems has proved to be slow, and the reasons may be both technological (e.g. accuracy of recognition engines, fusion engines, authoring) as well as usability-related. In this paper, we explore a few patterns of "command and control" style multimodal interaction (MMI) using touch gestures and short speech utterances. We then describe a multimodal interface for a photo browsing application and a user study to understand some of the usability issues with such interfaces. Specifically, we study walk-up use of multimodal commands for photo manipulations, and compare this with unimodal multi-touch interactions. We observe that there is a learning period after which the user gets more comfortable with the multimodal commands, and the average task completions times reduce significantly. We also analyze temporal integration patterns of speech and touch gestures. We see this as the first of many studies leading to more detailed understanding of user preferences and performance for using MMI, which can help inform the judicious use of MMI in designing interactions for future interfaces.

Categories and Subject Descriptors

H.5.2 [User In-terfaces]: User interface management systems; Voice I/O; Natural language; D.2.2 [**Software Engineering**]: Design Tools and Techniques - *user interfaces;*

General Terms

Algorithms, Human Factors

Keywords

multimodal systems, framework, authoring, usability

1. INTRODUCTION

Multimodal interfaces that integrate multiple input modalities such as speech, gestures, gaze and so on have been a focus of research in recent years because of their promise of *higher task efficiency, lower error rates and higher user satisfaction*, compared to unimodal ones [1, 2]. The interest in multimodal interaction is also being driven by the increasing ubiquity of inexpensive sensors such as touch sensors, web-cameras, microphones, and accelerometers in personal devices, and the increased computational power of these devices that is allowing vision and speech processing to be performed in real time.

However, the penetration of these interfaces in real-world systems has not been significant. Part of the problem has to do with the *technological* issues such as the accuracy and performance of input modality recognizers, efficiency and robustness of fusion engines, platform capabilities etc. Further, authoring multimodal application interfaces can be a complex task for the average application developer. The other equally important part of the problem has to do with the *usability* and design of these interfaces. Deeper understanding of user interaction patterns and usage preferences among different modalities for different application scenarios and related tasks is required to be able to design these new multimodal interfaces.

In this paper, we explore simple patterns of combining speech and multi-touch gestures into multimodal commands, and apply these patterns to define a multimodal command vocabulary for a photo browsing application. This is accomplished using *Mist*, a multimodal framework that we have designed and developed that supports easy and rapid authoring of multimodal commands via graphical user interface. We then use this application as a testbed to study user preferences and performance in using multimodal interactions, as compared to purely unimodal multi-touch interactions. We also present our observations of multimodal temporal integration patterns of different users, which can help improve the design of temporal fusion.

The paper is organized as follows. We first discuss multimodal patterns with references to previous work in Section 2. The Mist framework and authoring interface is described briefly in Section 3. Empirical analyses of user preferences and performance, and comparison with existing unimodal multi-touch interactions are presented in Section 4. We conclude the paper with discussions and next steps in Sections 5 and 6 respectively.

2. MULTIMODAL PATTERNS

Many multimodal systems have been proposed for research purposes with various target applications as focus. For example, MATCH [3] was primarily designed to be a city guide, with map navigation being the primary task using multimodal interaction. SmartKom [4] was designed for various uses such as a smart kiosk and a smart home companion. The primary tasks were information navigation and organization tasks such as booking movie tickets, organizing the home living environment etc. QuickSet [5] was one of the very early research prototypes with a focus on battlefield planning and emergency response. These early systems used speech along with stylus or touch as the input modalities, and established patterns of combining these modalities. Beginning with Put-that-there [6], the most common pattern of combining speech and gesture has been to use spoken commands and *deictic* pointing gestures (via touch, stylus, glove etc) to indicate the objects being referred to. QuickSet introduced other patterns in which stylus gestures and speech were placed on a more even footing and could be used unimodally as well. There has also been research into multimodal conversational interfaces incorporating conversational gestures, speech and other modalities such as eye-gaze and facial expressions, but these are typically not used for command-style interaction with computer systems.

Our research focus has been to support short multimodal command interactions for walk-up interactions with platforms such as PCs and kiosks using available modalities such as touch, speech, hand gestures and so on. One of the common multimodal interaction patterns may be called *speech-primary interaction*, wherein speech contains most of the information including the command, and deictic (pointing) gestures are used to indicate the referents in the utterance such as icons or other visual information, e.g. "*delete (↗) this and this (↗)* ". Here the arrows indicate pointing gestures made towards specific on-screen objects (by touching, or pointing from a distance) during the course of the utterance.

A second pattern of interaction that we have devised is what we call *gesture-primary interaction*, wherein the gesture is primary, and speech is used to qualify or parameterize the gesture, e.g., doing a two-finger "*pinch zoom*" touch gesture on a displayed photograph while uttering "*two times*", in order to see the photograph enlarged two times. In the above examples, the gestures may be touch and captured using a touch sensor, or non-touch and captured using a camera. Speech may also be used to provide related but different interpretations to the same gesture so as to restrict the vocabulary of gestures to a small set.

3. MIST MULTIMODAL FRAMEWORK

In order to be able to easily and rapidly implement such interactions, we have designed and developed a multimodal interaction framework called Mist. As described earlier, Mist is meant to enable application developers to build multimodal command and control interfaces for their applications, by focusing on the multimodal interactions, and abstracting the many complexities of input modality interpretation, management and fusion. The Framework runtime runs as a service in the background. In order to enable specific multimodal interactions, the application developer authors the multimodal interactions using the Framework's authoring tool, and codes handlers that listen to and respond to the corresponding interaction events from the Mist service. The framework is described in detail in [7].

3.1 Authoring Interface

Mist provides an authoring tool that allows developers and interactions designers to define new multimodal commands via a GUI (Figure 1). This enables rapid prototyping and testing of new interactions.

A multimodal command involving, say a multi-touch gesture and speech input can be defined by selecting the appropriate touch gesture (e.g "*zoom enlarge*") from the supported touch gesture vocabulary from a drop down menu, and entering the accompanying speech utterance (e.g. "*two times*") in a text box. The user can then give a name to the command for future reference. The newly defined multimodal command gets added to a common repository of commands as shown in Figure 1, and can be associated/ disassociated with future applications, thus facilitating reuse. The interactions for an application once defined are in turn parsed to create the application-specific *multimodal vocabulary (*or *grammar)* used by the framework during runtime.

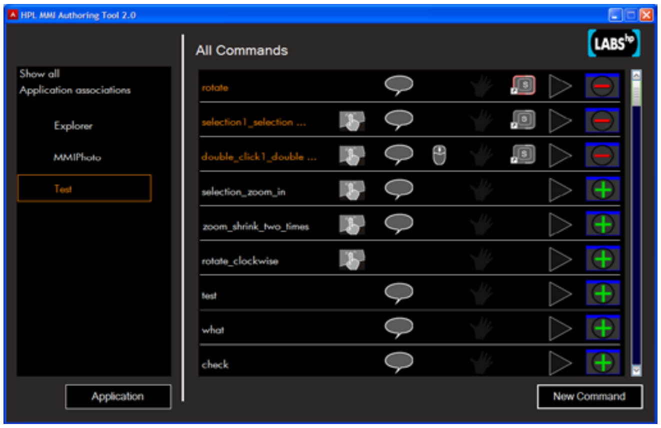

Figure1: Mist Authoring Tool UI

4. EVALUATION

4.1 Goals

Our goal for user evaluation was to evaluate the following usability aspects of multimodal interactions for a photo-browsing application: (i) performance, (ii) preferences, and (iii) patterns of temporal integration. We used familiar interactions such as unimodal multi-touch as our baseline for comparing performance and preferences. We also sought to collect subjective satisfaction data for a more holistic assessment of the use of multimodal interactions in designing interfaces. In order to design better multimodal fusion engines, we collected data to understand the multimodal temporal integration patterns for use of touch and speech (i.e., the temporal sequence of touch and speech inputs when using a synergistic multimodal command).

4.2 Apparatus

We used a HP TouchSmart PC running Windows 7 for our evaluation. A photo browsing application was developed and multimodal interactions that combined *speech* and *multi-touch* were integrated using the Mist framework. The list of multimodal commands is shown in Table 1. As mentioned earlier, unimodal multi-touch interactions for photo manipulation was used as the baseline for comparison.

4.3 Participants

Ten participants were recruited from within the office environment for the experiment. They were from the 20-35 year age group. They had used some form of touch enabled device like a computer or a phone. This age groups' familiarity with baseline unimodal multi-touch interactions ensured that unfamiliarity with the interface did not impact the task completion times unduly. Participants received compensation in the form of a gift certificate of nominal value.

	Function	Touch Gesture	Speech
Rotation	Clockwise	 Select object	"Rotate clockwise" (90°)
	Anti-clockwise	Select object	"Rotate anti-clockwise" (90°)
	Straighten photo	Select object	"straighten"
	Invert photo	Select object	"invert"
Zooming In/Out	Zooming In	Select object	"zoom in"
			"2 times"
	Zooming Out	Select object	"zoom out"
			"2 times"

Table 1: Multimodal commands enabled in the application

4.4 Experimental Procedure

Three tasks were designed for the users to perform one after the other with a gap of 5 mins between tasks: *Task1* (Rotate the photo clockwise by 90 degrees and zoom-in two times the current size); *Task2* (Rotate the photo back by 90 degrees and zoom-out two times to bring it back to original state); *Task3* (Align the photograph so that it is perfectly horizontal, right side up, and zoom-in two times).

Task1 was designed as a simple photo manipulation task which could be efficiently carried out even with unimodal multi-touch interactions. *Task2* was designed to be repetition of *Task1*, to observe whether users used discrete multimodal commands compared to unimodal multi-touch gestures requiring significant finger translation. *Task3* is a task that requires exactness and was designed to observe user performance and preference for precise multimodal commands as opposed to approximate unimodal multi-touch ones.

Before beginning the experiment, sufficient training in the different multimodal commands was provided to the participants, and a 10 minute warm-up session was given or participants to try out these interactions before starting the experiment. For the unimodal multi-touch interactions, since all the users were familiar with the interactions, no training was required but the users were given time to warm-up using those interactions as well.

Users were asked to carry out the three tasks in succession using each set of commands - *multimodal* (MMI) and *unimodal multi-touch* (MT) - and task completion times were measured. Subsequently a feedback session was conducted to collect subjective evaluation scores on a scale of 1 to 7, indicating strong disagreement to strong agreement respectively.

4.5 Design

A *within-subjects* design was used wherein participants were randomly assigned to two groups of 5 participants each. The first group performed the experiment using MMI first and then MT, whereas the other group did it in the reverse order. At least 24 hours elapsed between the use of each of the techniques for each group to limit interference between the techniques.

4.6 Results

4.6.1 Performance

Figure 2 presents an analysis of task completion times for the 10 participants for the three tasks. While the use of MMI resulted in acceleration of about 20% from *Task2* to *Task3*, no significant acceleration in task completion time was noticed for MT interactions. This may be attributed to of the familiarity gained by the user in using MMI from performing *Task 1* and *Task 2*, as opposed to prior familiarity of the users with MT. On the average MT outperformed MMI for *Task1* and *Task2* since these were simple manipulation tasks for which users were familiar with the use of MT. For *Task 3*, which required more exact manipulation, we found no significant difference in performance between MMI and MT.

From this data there is no clear indication of the efficiency of one technique over the other. While MT enjoys the advantage of familiarity among users, MMI has the advantage of being capable to precise commands such as "straighten", "increase contrast" etc. without going through menus. Similarly, Analysis of Variance also does not indicate significant effect of technique on the task completion times ($F_{1,58} = 7.45 > F_{critic} = 7.09, p < 0.01$).

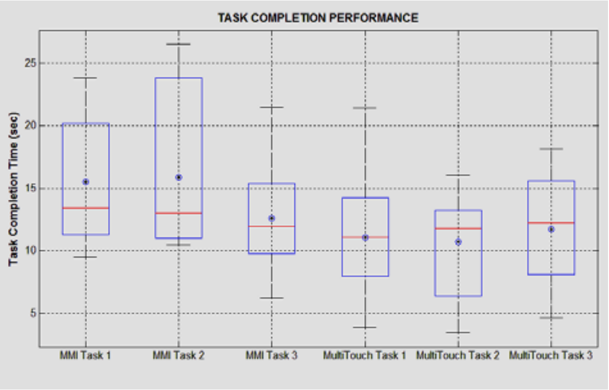

Figure 2: Box plot of the task completion times for each task and each technique. The mean value is indicated by the circle

4.6.2 Subjective Satisfaction

The MT interactions imposed a cognitive load on the users to move or manipulate the object accurately, and hence the users tended to fumble at first with this interaction.

MT was found to be well-suited for the manipulation of collages of images stacked one on top of another, but had the problem of inadvertent commands being executed when users were attempting to execute a unimodal multi-touch command on a specific object. For example sorting a stack of objects on the display was very easy with both hands, but while zooming-in/out, rotate command was also triggered intermittently. After some time, it was observed that the users got better at doing fluid rotate and zoom movements using MT.

With MMI, users seemed to prefer touching an object first and then uttering the speech component of the command. Users who did not prefer to touch first felt that in touching first, they had to do a touchdown and hold the object before saying the speech command, which could result in the target starting to move, which in turn would adversely affect the command execution.

Participants felt that for precise operations e.g. rotating a photograph clockwise by 90 degrees or straightening a picture, MMI was much better. Users noted that the one hindrance was that they had to memorize the commands exactly and this left no room for errors, which led to stress.

With gesture primary multimodal interactions e.g. zoom-out gesture accompanied by speech " two times ", users tended to do the gesture and speech and felt that the command was executed successfully even though the zoom-out command was executed on the picture totally in gesture and without the speech command (this was because there was no visual feedback on the actual zoom-out percentage).

Users perceived MT to have a more engaging and fun element with instantaneous and continuous feedback, as opposed to discrete MMI commands which were executed abruptly and quickly. Users perceived MMI to be more accurate but MT to be more intuitive. Some users also raised apprehensions about the delays between the touch and speech that the system can tolerate. MMI was perceived to be more open to errors. However they overcame these apprehensions with some practice.

4.6.3 Temporal Integration Patterns

As mentioned earlier, another objective of our study was to understand the temporal patterns of combining speech and gesture as part of multimodal commands. Figure 3 shows a plot of the temporal integration patterns for multimodal interactions across all users. These patterns are measured by the delay between (the system receiving) the starting of touch and speech events, measured in milliseconds. The positive values indicate that the user used *speech before touch* (SBT) while negative values indicate that the user used *speech after touch* (SAT). We see from the plot that most users - about 72%- used the SAT temporal integration pattern, while the remaining 28% followed the SBT pattern. The mean delays for SAT and SBT were approximately 600 ms and 500ms respectively. The maximum delays for SAT and SBT were found to be approximately 1600 ms and 1200 ms respectively.

The data indicates that the fusion algorithm used to combine speech and gesture events has to address both SAT and SBT patterns of temporal integration. However, in taking care of the larger delays of the order of ±1500 ms, the challenge would be to deliver real-time feedback to the user.

5. DISCUSSION

Though there has been a lot interest in multimodal interfaces recently and many research prototypes have been developed, many *technological* and *usability* related issues still need to be addressed to bring these interfaces into real-world systems. Our Mist framework addresses some of the technological issues.

Unimodal multi-touch (MT) represents a significant improvement over mouse-based interactions in that multiple objects can be manipulated at a time, and more direct manipulation is possible. In the same vein, multimodal interfaces in theory have clear advantages over unimodal ones. However, in our limited experiments we did not observe clear user preferences for MMI over MT. On the contrary MT was perceived to be more intuitive and engaging, owing to the availability of instantaneous and continuous feedback. Though speech is useful for circumventing menus, we also saw that people forget speech commands very easily or use a lot of variants of the same speech command, which is detrimental to speech recognition accuracy and consequently the success of the interface. Nevertheless, discrete multimodal commands that can perform precise tasks such as “straighten”, “invert” etc., were found to be very useful by users. Clearly an important advantage of MMI over *unimodal speech input* is the ability to use input speech – generally very prone to errors because of ambient noise – reliably because of mutual disambiguation with gestural input.

Another interface design and usability challenge is to be able to inform users of the MMI commands available to them, without getting in the way of the interaction. Before MMI interfaces can become mainstream, more in-depth understanding of these aspects of feedback and affordances is needed, even if technology solutions are available for building robust multimodal interfaces.

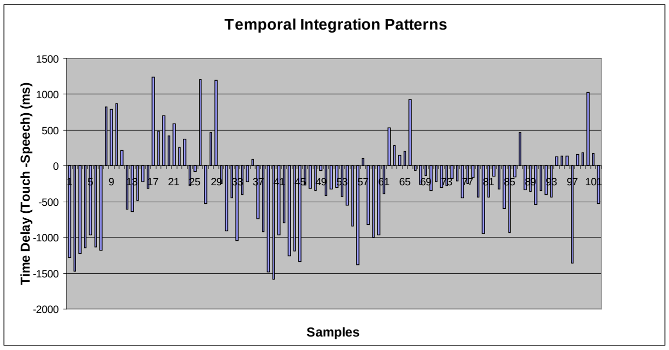

Figure 3: Plot of the temporal integration patterns for the multimodal interactions

6. SUMMARY AND NEXT STEPS

In this paper, we explored the use of simple gesture-speech multimodal patterns for a photo browsing application and presented findings from a user study to understand some of the usability issues with such interfaces. We briefly described the Mist Multimodal Framework for rapid development of multimodal applications. The framework provides an authoring environment to develop new multimodal applications very easily, and a runtime that provides the functionalities of running a multimodal application.

In the user study, though users’ task completion times using multimodal interactions improved with use, they did not outperform unimodal multi-touch. Users showed a preference for unimodal multi-touch for fluid interactions, but liked multimodal interactions for precise and discrete manipulations. The users also displayed a variety of temporal integrations patterns of speech and touch which can inform the design of more robust temporal fusion algorithms.

In the future, we plan to conduct more in-depth user evaluation for other multimodal patterns and application scenarios. We see this as the first of many studies leading to more detailed understanding of user preferences and performance for using multimodal interactions, which can help inform the judicious use of such interactions in future interfaces. We also plan to improve upon the multimodal fusion engine, and enhance the authoring tool to support these patterns. We are also working on improving the accuracies of the individual recognizers, and on integrating user context information such as presence, and expression, to enable even richer multimodal interaction.

7. REFERENCES

[1] Oviatt, S.L. *Multimodal interfaces, Handbook of Human-Computer Interaction* (revised edition), (ed. by J. Jacko & A. Sears), Lawrence Erlbaum Assoc: New Jersey, 2006.

[2] Oviatt, S. L., “Mutual Disambiguation of Recognition Errors in a Multimodal Architecture”, *Proc. of the ACM CHI 99 Human Factors in Comp. Sys. Conf.*, pp. 576-583.

[3] Johnston, M., et. al., “MATCH: an architecture for multimodal dialogue systems”, In *Proc. of the 40th Annual Meeting on Assoc. For Comp. Ling.*,, 2001, pp. 376-383.

[4] Wahlster, W., *Smartkom: Foundations of Multimodal Dialogue Systems (Cognitive Technologies)*. Springer-Verlag New York, Inc, 2006

[5] Cohen, P. R., et. al., “QuickSet: multimodal interaction for simulation set-up and control”, In *Proceedings of the Fifth Conference on Applied Natural Language Processing* (Washington, DC, March 31 - April 03, 1997). Applied Natural Lang. Conf., Assoc. Comp. Linguistics, NJ, 20-24.

[6] R. A. Bolt, “Put-that-there: Voice and gesture at the graphics interface,” In *International Conference on Computer Graphics and Interactive Techniques*, July’ 80, pp. 262–270.

[7] Mist Framework, HP Labs Technical Report in preparation.