Where is "*it*"? Event Synchronization in Gaze-Speech Input Systems

Manpreet Kaur
Rutgers University
CAIP Center
Piscataway, NJ 08854,
+1-732-445-5099
mkaur@caip.rutgers.edu

Marilyn Tremaine
New Jersey Institute of Technology
Department of Information Systems
Newark, NJ 07102
+1-973-596-5824
tremaine@njit.edu

Ning Huang, Joseph Wilder, Zoran Gacovski, Frans Flippo, Chandra Sekhar Mantravadi
Rutgers University, CAIP Center
Piscataway, NJ 08854,
+1-732-445-5099
{huangn,wilder,zorang,fflippo, sekhar}@caip.rutgers.edu

ABSTRACT
The relationship between gaze and speech is explored for the simple task of moving an object from one location to another on a computer screen. The subject moves a designated object from a group of objects to a new location on the screen by stating, "*Move it there.*" Gaze and speech data are captured to determine if we can robustly predict the selected object and destination position. We have found that the source fixation closest to the desired object begins, with high probability, ***before*** the beginning of the word "*Move*". An analysis of all fixations before and after speech onset time shows that the fixation that best identifies the object to be moved occurs, on average, 630 milliseconds before speech onset with a range of 150 to 1200 milliseconds for individual subjects. The variance in these times for individuals is relatively small although the variance across subjects is large. Selecting a fixation closest to the onset of the word "*Move*" as the designator of the object to be moved gives a system accuracy close to 95% for all subjects. Thus, although significant differences exist between subjects, we believe that the speech and gaze integration patterns can be modeled reliably for individual users and therefore be used to improve the performance of multimodal systems.

Categories and Subject Descriptors
H5.2 [**Information Interfaces and Presentation**]*: User Interfaces, Input devices and strategies.*

General Terms
Measurement, Design, Experimentation, Human Factors.

Keywords
Multimodal Interfaces, Eye-Tracking, Multimodal Fusion, Gaze-Speech Co-occurrence.

1. INTRODUCTION
This paper describes a study in which we combined speech and gaze in a multi-threaded (data and speech are received and processed simultaneously) interactive system designed for one simple task, to move an object to a new location using one single phrase, "*Move it there*." The purpose of the study was to determine precisely how eye movements occurred in relationship to speech. A key goal in our analysis was to learn how robustly we could identify the object the person wanted to move when they referred to the object via anaphoric reference. Knowing the time relationship between eye movements and speech utterances will allow us to build computer algorithms that use the onset times of spoken words to ascertain the context of the speech and also, possibly, to provide a natural mechanism for identifying and manipulating screen objects while the user talks about them. By natural, we mean that gaze will primarily be used to acquire information about the visual scene, not as a consciously controlled input device.

This paper is organized as follows. In the next section we discuss related research in the areas of multimodal system fusion, eye fixation determination and speech and gaze interfaces. This is followed by a description of PICC, the speech-gaze multimodal system we have built. In Section 4 we present the study we conducted using PICC. In Sections 5 and 6, we analyze the data from the study and present our conclusions. This is followed by a brief discussion of future work.

ICMI '03, November 5-7, 2003, Vancouver, British Columbia, Canada.

2. RELATED WORK
2.1 Multimodal Systems
The key focus of multimodal systems in prior work has been on the integration of gestures and speech [27]. Oviatt et al. [21] describe a system that effectively combines speech and pen-based gestures to manipulate objects on a map. In an overview article on the integration of speech and gesture-based multimodal systems Oviatt et al. [20] argue and demonstrate that using speech and pen input is more efficient than mouse-based selections for a variety of common computer tasks. A key advantage of pen-based multimodal systems is the ability to unambiguously identify the objects being manipulated, both through events of significant duration (dwell time of pen on object) and of distinct motor activity (pen point position on screen or pad). Using gaze direction to identify screen objects is more difficult. Because gaze is a human input device, eyes jump randomly about the screen, picking up elements of the scene to build a mental image of the visual picture. Dwell times are short and there are many of them [11] [28] [33]. Thus, identifying which dwell time and location represents an event that is related to speech is a difficult task.

Gaze as a part of the multimodal input system was first used in the landmark multimodal system by Bolt [2]. Bolt combined hand gesture, gaze and speech to select and manipulate objects on a large wall display. Any gaze selection had to dwell on a desired object for a long enough time to make a unique selection. Others have overcome this requirement to "stare" at a desired object to generate a gaze event by using another event such as mouse movement to determine what gaze position should be used. Zhai et al. [34] jump the mouse cursor across the screen to where the eye is looking every time the hand moves to the mouse. Unfortunately, although this approach is seemingly useful in saving mouse movement time, they found no significant differences in performance. More recent attempts to combine gaze and speech in a "natural" fashion include the work of Maglio et al. [17], in which the researchers use the location of gaze to determine to whom a person is talking. This technique is also used by Vertegaal et al. [32] to indicate a focus of attention to a distant user in remote collaboration scenarios. Salvucci and Anderson [25] use gaze direction to help interpret spoken protocols. Although all of these studies are multimodal, only Zhai et al. [34] use gaze as a computer input event. In the other two systems, gaze direction is used to convey information to other humans. In the system by Zhai, mouse movements provide precision input, not gaze. As previously mentioned, this is, no doubt, due to the difficulty of determining what gaze event to use as input.

2.2 Gaze Input Systems

Advances in the development of eye trackers make it possible to rapidly and non-invasively measure gaze location, leading to the possibility of using gaze as an input mechanism. The use of the eyes as a means of specifying input is appealing for a number of reasons:

- The gaze location is the only reliable predictor of the locus of visual attention.
- Gaze can be used as a "natural" mode of input that bypasses the need for learned hand-eye coordination such as that required to use a mouse, a joystick or a screen touchpad.
- Gaze selection of screen objects is expected to be significantly faster than hand-eye coordination [28].
- Gaze allows for hands-free interaction. This advantage has already been employed in the design of interfaces for motor-disabled users [3].

Not surprisingly, initial attempts to use gaze as a computer input mechanism have attempted to build mechanisms that use gaze in a manner similar to that of the mouse [7]. Although gaze movement is much faster than that of a mouse, there are several disadvantages to this approach:

- There is no analogy to the mouse button - eye blinks and fixations have been used, but are hard to learn [31].
- Clicking and dragging objects is not possible.
- Unlike the mouse, gaze control is always active. Intentional and non-intentional selections are hard to distinguish [11].
- The human eye does not move smoothly but jumps from position to position. Fixing the eyes on a specific position for even a short time is fatiguing for the user.

The focus of early eye-tracking research to determine human cognitive processing suggests how gaze could be used. This research assumes that humans will look at those places on the screen that reflect their cognitive processing. Thus, if their cognitive processing is that of manipulating objects on a computer screen, then the aggregation of where they look should contain information about what they intend to do. Nielsen [19] advocates this approach to interface design as do Starker and Bolt [29]. The advantage of aggregating and interpreting raw data at a higher semantic level is two-fold. Users do not need to learn difficult eye movement control, nor do they need to pay conscious attention to the routine cognitive processing steps required to carry out screen object manipulation. This approach has been tried successfully by Salvucci [24] who used hidden Markov models to predict a user's intent when "eye typing." We take a similar approach in this paper assuming that users are "looking at" objects and places they are talking about.

2.3 Speech and Gaze Multimodal Systems

Humans generate information through multiple channels. As they speak, they gesture with their hands and generate facial expressions for additional emotive content. Eyes are used both to express emotion and to create a focus of attention. Each of these activities is synchronized with the other to convey a complete information message. Speech recognition, hand gesture recognition and eye pattern interpretation all have high error rates when used for computer input. Given the difficulty of interpreting user intent from any one of these activities, an advantage can possibly be gained by aggregating the interpretation from each activity. This combination approach also allows the user to select the best input for performing a task, e.g., selecting an object with a gesture rather than a spoken explanation.

A key issue in combining any of these human activities is to identify the synchronicity of the input events. This is important when combining speech and gaze input because of the high data rate and data variability for both human activities. Recognized input events are likely to occur within the 60 to 100-millisecond level, a much finer resolution than gesture events.

Researchers have already studied the timing relation of speech and gaze events in the social domain, where humans use gaze direction to exchange floor control [5] and indicate focus of attention [32]. Psycholinguistic studies have further strengthened the assumption that speech and gaze are used synchronously. Tanenhaus et al. [30] conducted a study in which they recorded eye movement patterns during spoken sentence comprehension to test the effects of relevant visual context on the rapid mental processes that accompany spoken language comprehension. Subjects heard spoken commands to move objects in a physical display, and it was seen that visual context influenced spoken word recognition and mediated syntactic processing and their eye movements were correlated with different parts of the sentences. Maglio et al. [17] provide stronger evidence for the relationship between gaze and speech. In their study (used to develop a user interface that anticipated user needs) they measured the distribution of gaze patterns in relationship to speech commands given in a natural office setting. Gaze direction was focused on the office object implied in the spoken command, e.g., the printer when the subject said, "*give me a copy*." Kaur [13] examined the relationship between gaze and speech in a computer domain and demonstrated that the gaze direction that provided information on the user's intent preceded the spoken description of this intent.

She also found that the time relation between speech onset time and gaze direction varied depending on the linguistic structure of the spoken command and individual differences in speaking speed.

Synchronizing gaze and speech events to interpret user intent has been tried in a number of applications. Hatfield et al. [8] describe a gaze-speech system for air force pilots that allows them to manipulate their displays. Their paper focuses on the eye movement protocols and associated dialog commands that can be used to generate a set of mission planning tasks. A key issue addressed in their work is the type of linguistic reference to use in tasks, e.g., specific or approximate deictic reference. Their work notes that the synchronization of speech and gaze is important but they do not address how this is best accomplished.

Koons et al. [14] describe a system that combines speech, gesture and eye tracking. Their system fuses time-stamped information from the various modalities through a process of multiple frame instantiation, by finding deictic (gesture and gaze) frames that have acceptable temporal proximity to the speech object frame. They do not explore the region of acceptable temporal proximity. Thus, although other research exists on fusing speech and gaze, no detailed exploration of the time synchronicity has been conducted. The present paper therefore attempts to answer the following three questions about gaze-speech multimodal systems:

1. What is the time relationship between a deictic reference and accompanying gaze patterns?
2. How robust is this relationship, i.e., can it be used in software algorithms to accurately predict the intended screen location?
3. Does the relationship hold across users or is it unique to each user, i.e., is a user required to train a speech-gaze system to his or her eye–speech patterns?

3. THE PICC MULTIMODAL SYSTEM

In this section we give a brief overview of the multimodal platform used for our study. PICC (Portable Interactive Command Console) is a test-bed for crisis management in the field. It is a portable unit carried in a jeep or all-terrain vehicle. PICC users work collaboratively in deploying equipment or manpower that is represented by symbols on a screen map. Figure 1 shows a prototype of PICC.

Figure 1. Portable Interactive Command Console (PICC)

Initial development of the PICC was sponsored by NSF under the STIMULATE program and by the US Army (CECOM) under the Future Combat Systems Command and Control program. PICC uses a gimbal-mounted eye-tracker (ISCAN RK726) to capture gaze direction and a steerable microphone array to capture speech. Embedded in the system are speech recognition, speech synthesis and language understanding modules. These are multi-threaded with the gaze input [18] to ensure parallel processing of the speech and gaze inputs.

4. STUDY DESIGN

To answer our questions about the time relationship between deictic speech and gaze positioning, we had subjects move objects on a computer screen using our PICC system while we collected data on their speech and gaze patterns. We then analyzed the data from this study to determine both the time relationship between speech and gaze and the strength of the relationship.

In order to run the study, we already had to have an algorithm in place that assumed a relationship between speech and gaze, so that when a subject directed a screen object to be selected and moved, the computer system would move the correct object most of the time. This is problematic because a weak algorithm is likely to have a large number of errors. Ours did. Users of the first version of PICC moved objects with an accuracy of only 60 percent. The initial algorithm we used worked as follows: We took the last eye fixation closest to the utterance of the word "it" and looked at the closest object to that eye fixation. If no object was found within a circle of a pre-set circumference of the eye fixation, the system asked the user for clarification. If more than one object was found within this circle, the object closest to the eye fixation was chosen. If a conflict existed between two or more possible candidate objects, the system asked the user to redo the task. There were several problems with this algorithm that led to our high error rate.

1. The eye jumps about considerably, making multiple eye fixations [10] [22]. The closest eye fixation when the referent is stated may not be the correct one.
2. Multiple objects were often very close together. Furthermore, unintentional overlaps occurred when two items were inadvertently placed too close together with a previous move command. This made it impossible to accurately determine the desired object.
3. Other items were moving on the screen because the interface supported collaboration among multiple sites. Such movement may divert the eye fixation.
4. Other variations in the display such as a complex map scene containing contrasting colors or a variety of patterns also impacted eye fixations [4].

Because of these problems we adjusted our object selection algorithm, based on our prior research[13], to select the eye fixation closest to the utterance of the word "*Move*" and then to pick the object closest to that eye fixation. Our plan was to run experiments to measure the relationship between eye fixations and speech commands using the PICC speech-gaze interface. Results from the experiments could then be used to develop better algorithms that would use timestamps from the speech understanding system combined with eye-tracker data to ascertain the screen object referred to in the speech commands. We assume that these algorithms will have parameters that will vary under

different display conditions and with different grammatical utterances.

To gather data for this analysis, we ran the following study. A set of nineteen colored balls was arranged in a hexagon shaped cluster so that each ball was equidistant from the other. The cluster was displayed in one of three positions on the left hand side of the screen (middle, top and bottom). A square gray box with a green shamrock-shaped object in it was displayed in three possible positions on the right side of the screen (middle, top, bottom). One ball in the cluster was colored green. The other balls were colored green-blue. The user was instructed to move the green ball to the location designated by the gray square by saying, "Mo*ve it there*." (We did not use the phrase, "*Move that there"* which is the correct English phrase for referring to an undetermined object because all of our subjects were non-native English speakers and some had trouble with pronouncing two words in succession that began with *"th."* The cursor for the eye tracker was displayed on the screen and the user knew that the software would use information from the eye-tracker to determine both the object to be moved and where to move it. Figure 2 gives an illustration of the trial screen that was displayed to the subject.

Balls that were to be moved were located in the middle layer of the cluster of balls so that they were neither on the edge nor in the center of the cluster. The size (1^o visual angle radius) and spacing (1^o edge-to-edge) of the objects in the cluster was based on the accuracy of the eye-tracker and recommendations from the vision literature [12] [15]. All eye movements were designed to be from left to right [16].

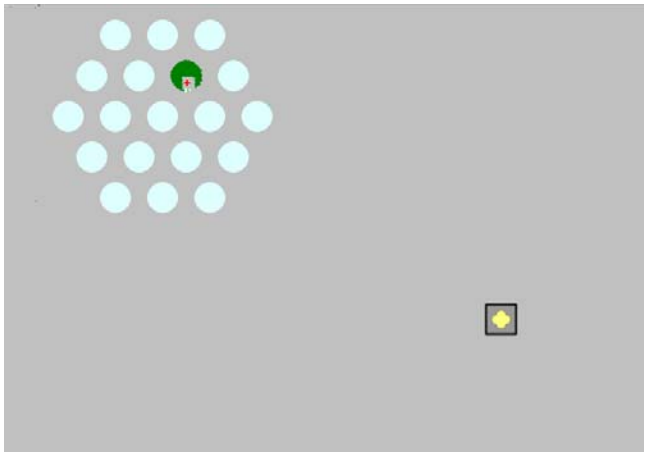


Figure 2. Example of an object cluster and source destination used in the study. The darker circle is the object to be selected. The square on the right is the destination to which the object is to be moved.

All subjects participated in two training sessions before the experiment session. The first training session was with the speech recognizer, IBM Via Voice. Subjects read text for 15 minutes to train the recognizer to their voice characteristics. The second training session involved the use of the eye-tracker. Subjects sat with their head mounted in a chin rest at a fixed distance of 23 inches from the computer screen. Subjects first participated in a five-point calibration scheme provided by ISCAN. We modified the calibration procedure so that subjects used speech, i.e., said "*next*" to move to the next calibration point. After the first calibration the subject engaged in the 16-point calibration test shown in Figure 3. Each of the 16 points is randomly displayed, one at a time, and the subject is directed to move his or her gaze to the displayed point. In Figure 3, the circles are the calibration points and the squares are the subject's gaze points.

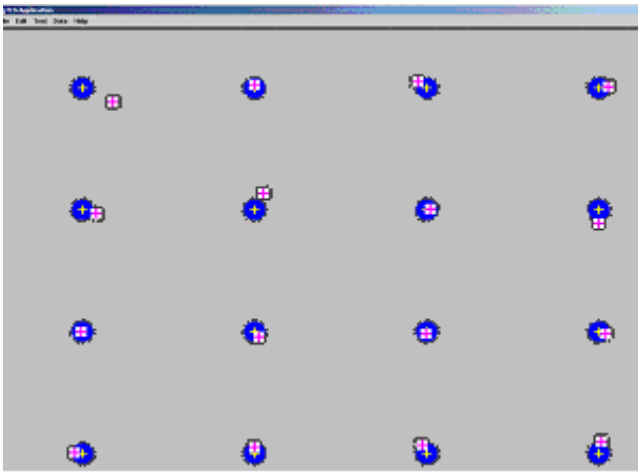


Figure 3. Typical results from a 16-point calibration test. The circles are the calibration points and the smaller squares are the user's gaze points.

The calibration tests gives the subject practice in moving the gaze cursor to a desired object and in giving speech commands to the system. The 16 point calibration test is repeated every 10 trials in the experiment to capture and correct any calibration drifts that might occur in the study.

After the initial calibration test is complete, the experiment trials begin. Each subject performs 54 trials. Each trial consists of an initial fixation circle that appears in the center of the screen. Subjects are told to look at this circle. After a short delay (1000 msec), the task screen is displayed with the source object cluster and the destination location marker displayed at the same time [9]. The subjects must move the designated object by looking at the desired object and the destination location, and giving the speech command "*Move it there.*" The system computes the gaze fixations for the source object and destination location using the algorithm described in the next section and moves the object. Once the object is moved to its destination, the next trial begins. Trials are randomly chosen to have the object cluster and the destination marker either in the top, middle or bottom position of the left and right side of the screen, respectively.

Speech and eye movement events are time-stamped and recorded for offline data analysis. The complete experiment session is also video recorded for offline analysis of error trials and extraneous events that might have contributed to the errors.

A total of five subjects (4 males and 1 female) participated in this study. All the subjects were graduate students in computing related fields, The subjects ranged in age from 20-30 years, and all had normal or corrected vision of 20/20. Each subject completed 54 trials for the study, resulting in a total data set of 270 trials.

5. RESULTS AND ANALYSES

As described in the previous section, we have collected time-stamped gaze and speech data for a simple "*Move it there*" task for 5 subjects, with 54 trials for each subject. Our goal was to determine where the eye fixation that represents "*it"* is during the speech command "*Move it there*". For fixation determination we have used our clustering algorithm, which is a modified dispersion algorithm [25]. This algorithm was used during experiment runtime and for offline data analysis because other studies suggested high accuracy and because of its runtime efficiency. The following is the pseudocode for this algorithm:

For a string of arriving gaze points returned by the eye tracker
(1) Set the start point equal to the first point arriving

(2) Set the mean location of the cluster forming the eye fixation to the location of the start pint

(3) For each new point, calculate its distance D to the current mean location

If D < threshold (1°visual angle), then

Add new point to the cluster

Calculate the new mean location of the cluster

Repeat (3)

Else if number of cluster points > 6 then

Consider the cluster an eye fixation

Else

Discard the cluster

Set the new start point to be the last point of the previous cluster+1
Go to (2)

We used this algorithm to determine the gaze fixation for the source object to be moved (source fixation), as well as the fixation that identified the location of the object's destination (destination fixation). This paper presents a detailed analysis of the timing relationship between speech and gaze for the source fixation ***only***. Figure 4 shows a typical scanpath for one complete trial for one subject. The numbered clusters mark all the fixations found in the trial by the cluster algorithm during offline data analysis. Our challenge is to identify which fixation refers to the source object.

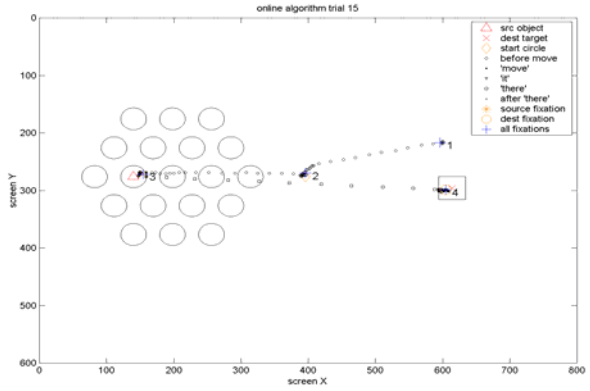


Figure 4. A typical scan path of a trial

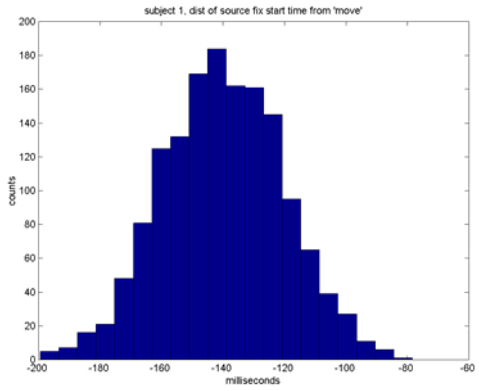
In order to identify the fixation that would best describe the source object, we need the distribution of the fixation start and end times. For our offline data analysis, we assumed apriori knowledge of the "desired" object and used the fixation closest to the desired object as the "correct" fixation. We examined the time relationship of this fixation to the various speech events, and observed it to be closest in time to the onset of the word "*Move*". Although all subjects completed 54 trials for the experiment, we discarded trials where the speech recognizer did not recognize the spoken command in the first attempt because we were concerned that eye-movement patterns might change for these trials. The number of trials per subject is shown in Table 1. We have performed an ANOVA test (alpha = 0.01, two tail F-critical (4, 228) = 3.41) to analyze inter-subject differences in the fixation's start time, end time and duration. Although significant, these inter-subject differences are not large; only 20-50 msec. Table 1 shows the average start times of a fixation (in relation to speech onset time) for each subject.

Table 1. Fixation's start time mean for object selection relative to speech onset time.

Subject	N	Mean fixation start, msec.	Std dev. of mean distribution (bootstrap)	Probability Fixation starts before 'move'
1	50	-140	20	0.839
2	46	-641	50	0.967
3	43	-1232	64	0.998
4	54	-722	51	0.972
5	40	-416	31	0.982

In order to increase the reliability of our estimates, we applied the bootstrap algorithm, which provides improved small-sample-based estimates of parameters of distributions by creating artificial data sets [23]. We used this method for computing the means and standard deviations of the experimental data, iterating 1500 times.

The analysis shows that the source fixation closest to the desired object begins, with high probability, ***before*** the beginning of the word, "*Move*." The bootstrap mean, standard deviation for the mean and the probability that the fixation starts before the onset of "*Move*" are given in Table I for the five subjects. Figure 5 shows the distribution of the mean values of source fixation start, end, and mid time, as well as fixation duration, for subject 1, based on 1500 iterations of the bootstrap algorithm.



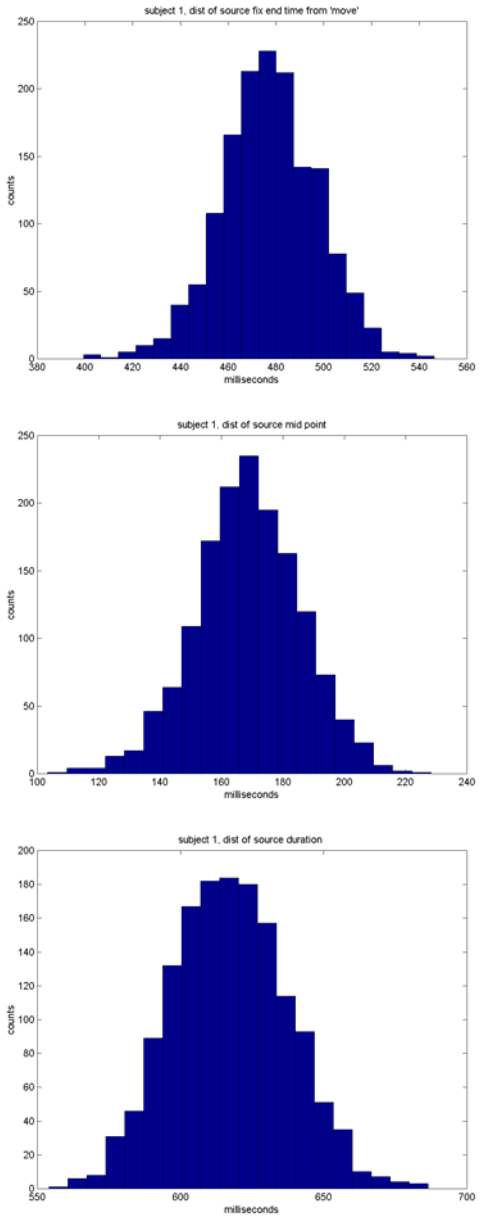


Figure 5. Distribution of fixation's: a) start time, b) end time, c) mid-point time, d) duration, (t=0 is onset of "move").

Figure 6 shows the time relationship between parts of the speech command and the corresponding eye movements (saccades and fixations) for one trial.

Once the source fixation has been reliably identified for each individual using the time relationship in Table 1, we can use the mid-time of that fixation as a reliable measure for finding the spatial referent for the desired object. It is seen that the deviation of individual gaze points from the midpoint of the source fixation is much smaller than the duration of the fixation. This is borne out in Table 2, where the standard deviation of gaze points about the midpoint is less than 23% of the duration of the fixation for all 5 subjects (the corresponding probabilities of individual samples falling within the fixation are also shown in the table).

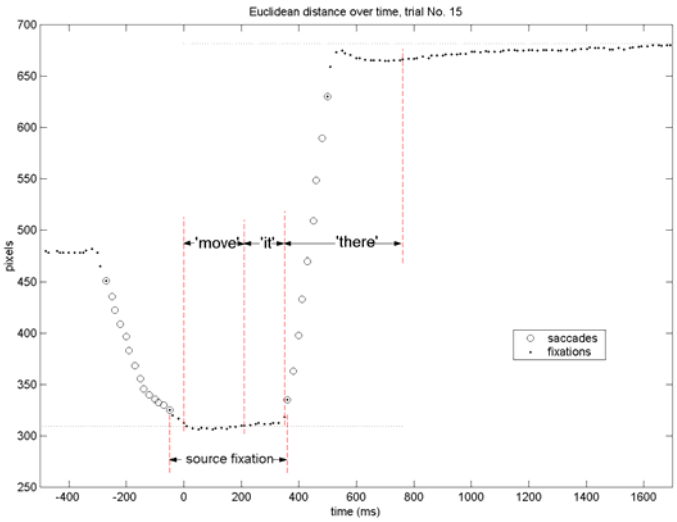


Figure 6. The relationship between saccades, fixations and speech

Consequently, knowing the location of that midpoint for an individual will enable that fixation to be identified with high probability, i.e. that midpoint will indicate where the speaker's attention was focused when speech began.

Table 2. Relationship between fixation's duration and mid-point relative to speech onset time, P(M) – probability of calculated mid-point falling within source fixation.

Subject	Fixation Mid-point		Mean fixation duration	P (M)
	Mean	Std dev		
1	169	121	617	0.989
2	-219	179	844	0.982
3	-550	226	1363	0.997
4	-174	190	1095	0.996
5	61	106	954	0.999

Overall, we can see that the fixation closest to the beginning of the utterance "*Move*" is, with high probability, the fixation closest to the object the user wished to move. This is borne out by the results shown in Table 3. Errors can be attributed to a combination of equipment calibration drift, performance of the chosen fixation detection algorithm, and unaccounted variations in subject behavior. Further analysis will help to classify these errors as to type.

Table 3. Errors in selecting the source object for each subject.

	Source Object	
Subjects	**No. Errors**	**% Errors**
Subject 1	5/50	10
Subject 2	5/46	10.9
Subject 3	3/43	7.0
Subject 4	2/54	3.7
Subject 5	0/40	0

6. CONCLUSIONS AND FUTURE WORK

In this paper we have described our prototype system for gaze and speech fusion for multimodal human-computer interaction. We have conducted experiments to determine precisely the time relationship between speech and gaze events in a simple "select and move" type multimodal task: "Move it there". This time relationship is important for the robust fusion of gaze and speech data streams in order to implement more "natural" and reliable gaze based interfaces.

We have found that the source fixation closest to the desired object begins, with high probability, ***before*** the beginning of the utterance "Move". Using the fixation closest to the onset of the word "move" we have observed accuracy close to 95%, up from approximately 60% in our initial implementation of the system, which used the fixation closest to the referent "*it*" as the source fixation.

We have seen significant differences between subjects but we believe that the speech and gaze integration patterns can be modeled reliably for individual users and can be used to improve the performance of the multimodal systems. We also observed significant training effects in our experiment. Two of the five subjects can be categorized as experts, while for one subject this was the first exposure to this system. The overall error rate was highest for the naïve subject. Fatigue seems to be an important consideration for a system of this type. We observed that most of the errors happened in the later parts of the experiment session for all the subjects, indicating that fatigue occurs relatively quickly in this type of multimodal environment.

Lastly, it is not only "*it*" that we want to accurately determine, but also, the location of "*there*." In our initial examination of the destination fixation, we observed that the fixation starts after the onset of the speech referent, "*there.*" Unfortunately, our examination of the timestamps received from the speech recognition system (IBM ViaVoice 4.0) indicates that we are only reliably receiving the onset times of the start of the spoken phrase. Thus, we could not perform the same calculations for the location of *"there*." For a gaze – speech interface to work effectively with multiple linguistic variations, we require accurate time stamps for the onset of each work in a spoken phrase. These onset times are probabilistic determinations as are our gaze fixations, and current off-the-shelf speech recognizers do not give this type of data.

We plan to extend the current work to include the destination fixation relationship in our analysis. We will also perform a cross validation of our results by examining from the video recordings of the experiment. Additionally, we will do a detailed analysis of the error trials to determine the cause of the errors. One of the important factors that affect the performance of gaze-based systems is the algorithm used to identify the fixations from the eye movement data. While much work has been done on identifying fixations and saccades in the vision literature, we are not so interested in identifying pre-defined parameters of gaze behavior as in identifying that pattern of gaze behavior which best predicts our object and destination locations. Some algorithms attempt to aggregate the largest number of gaze points to identify a fixation. We may want to have fewer gaze points that are closer in time to the start of a speech element. Thus, we need to reanalyze existing algorithms that have been studied to see if they best match our criteria for identifying a gaze fixation. Finally, it is our intent to continue this work with more linguistic variation and multimodal tasks performed in a range of environments to see how general our results are.

7. ACKNOWLEDGMENTS

We acknowledge the financial support we received for this work from CECOM at Fort Monmouth, New Jersey and from the Rutgers CAIP Center.

8. REFERENCES

[1] Bernsen, N. O. and Dybkjær, L.: Is speech the right thing for your application? In Proceedings of the International Conference for Spoken Language Processing, ICSLP'98, Sydney. Australian Speech Science and Technology Association 1998, 3209-3212.

[2] Bolt, R. A. *The Human Interface.* Lifetime Learning Publications, Belmont, CA, 1984.

[3] Corno, F., Farineti, L. and Signorile, I. A cost-effective solution for eye-gaze assistive technology. In *Proceedings of the ICME 2002 IEEE Conference on Multimedia and Expo,* IEEE Press, Piscataway, NJ, 2002.

[4] Curry, R., Hung, G. K., Wilder, J. and Julesz, B. Context effect of common objects on visual processing. *Optometry and Vision Science Vol.72*, 1995, 452-460.

[5] Dabbs, J. M., Jr., Evans, M. S., Hooper, C. H., & Purvis, J. A. Self monitors in conversation: Patterns of speech and gaze. *Journal of Personality and Social Psychology Vol. 39*, 1980, 278-284.

[6] Farid, M. M. and Murtagh, F. Eye-movements and voice as interface modalities to computer systems. In *Proceedings of OPTO Ireland,* SPIE Press, Bellingham, WA, September 5-6, 2002, CD-ROM.

[7] Glenstrup, A. J. and Engell-Nielsen, T. Eye controlled media: present and future state. Technical report, University of Copenhagen, Denmark, 1995.

[8] Hatfield, F., Jenkins, E. A., Jennings, M. W. and Calhoun, G. Principles and guidelines for the design of eye/voice interaction dialogs. In *Proceedings of the IEEE 3rd Symposium on Human Interaction with Complex Systems, HICS/96,* Dayton, OH, August 25-27, 1996, 10-19.

[9] Hung, G. K., Wilder, J., Curry, R., and Julesz, B. Simultaneous better than sequential for brief presentations. *Journal of the Optical Society of America Vol. 12*, 1995, 441-449.

[10] Hung, G. K., Wilder, J., Weiss, F. and Curry, R. K, Random and direct path eye movements during target search. *Medical Science Research Vol. 21*, 1993, 389-391.

[11] Jacob, R J. K. The use of eye movements in human-computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems, Vol. 9* (3), 1991, pp. 152-169.

[12] Kapoula, Z., and Robinson, D. A., "Saccadic undershoot is not inevitable: saccades can be accurate," *Vision Research Vol. 26*, 1986.735-743,

[13] Kaur, M. *Integration of Gaze and Speech for Multimodal Human-Computer Interaction*. Unpublished Ph.D. dissertation, Department of Biomedical Engineering, Rutgers, the State University, 2000, 142 pages.

[14] Koons, D. B., Sparrell, C., J., and Thorisson, K.R. Integrating simultaneous input from speech, gaze, and hand gestures. In M. Maybury (Ed.) *Intelligent Multimedia Interfaces*, MIT Press, Menlo Park, CA, 1993, 252--276.

[15] Kowler, E. and Blaser, E. The accuracy and precision of saccades to small and large targets. *Vision Research Vol. 35* (12), 1995, 1741-1754.

[16] Lin, W., Kaur, M., Tremaine, M., Hung, G. and Wilder, J.. Performance analysis of an eye-tracker.In *Proceedings of the SPIE Conference on Machine Vision Applications, Architectures and Systems Integration V*, 1999, CD-ROM.

[17] Maglio, P. P., Matlock, T., Campbell, C. S., Zhai, S. and Smith, B. A. Gaze and speech in attentive user interfaces. In *Proceedings of the Third International Conference on Multimodal Interfaces, ICMI 2000,* Beijing, China, October 14-16, 2000, 1-7.

[18] Mantravadi, C. S., Wilder, J., Grove, D. and Yuan, X. A Java-based multimodal human-computer interface architecture. In *Proceedings of ICICS-2001*, Singapore, IEEE Press, Piscataway, NJ, 2001, CD-ROM.

[19] Nielsen, J. Noncommand user interfaces. *Communications of the ACM*, 36:83-99, 1993

[20] Oviatt, S., Cohen, P., Wu, L., Vergo, J., Duncan, L., Subh, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J. and Ferro, D. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. *Human Computer Interaction, Vol. 15* (4), 2000, pp. 263-322.

[21] Oviatt, S., DeAngeli, A. and Kuhn, K. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of the ACM CHI'97 Conference on Human Factors in Computing Systems,* ACM Press, New York, NY, March 22-27, 1997, pp. 415-422.

[22] Pelz, J. B., Canosa, R. and Babcock, J. Extended tasks elicit complex eye movement patterns. In *Proceedings of the ACM ETRA 2000 Eye Tracking Research and Applications Symposium,* ACM Press, New York, NY, 2000, 37-44.

[23] Press, W.H., Vettering, W.T., Teukolsky, S. A., and Flannery, B. P., *Numerical Recipes in C, Second Edition*, Cambridge University Press, 1992, 691-692.

[24] Salvucci, D. D. Interring intent in eye-based interfaces: Tracing eye movements with process models. In *Proceedings of the ACM CHI'99 Conference on Human Factors in Computing Systems.* ACM Press, New York, NY, May 15-20, 1999, 254-261.

[25] Salvucci, D. D. and Anderson, J. R. Intelligent gaze-added interfaces. In *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems*, ACM Press, New York, NY, 2000, 273-280.

[26] Salvucci, D. D. and Goldberg, J. H. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the ACM ETRA 2000 Eye Tracking Research and Applications Symposium,* ACM Press, New York, NY, 2000, 71-78.

[27] Sharma, R., Pavlovic, V. I. and Huang, T. S. Toward multimodal human-computer interfaces. In *Proceedings of the IEEE, Vol. 86,* (5), May 1998, 853-869.

[28] Siebert, L. E. and Jacob, R. J. K. Evaluation of eye gaze interaction. In *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems,* ACM Press, New York, NY, April 1-6, 2000, 281-288.

[29] Starker, I. and Bolt, R. A., A gaze-responsive self-disclosing display. In *Proceedings of the CHI'90 Conference on Human Factors in Computing Systems*, ACM Press, New York, NY, 1990, pp. 3-9.

[30] Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K. and Sedivy, J. Integration of visual and linguistic information during spoken language comprehension. *Science, Vol. 268,* 1995, pp. 1632-1634.

[31] Velichkovsky, B. M. and Hansen, J. P. New technological windows into mind: There is more in eyes and brains for human-computer interaction. In *Proceedings of the CHI'96 Conference on: Human Factors in Computing Systems*, ACM Press, New York, NY, 1996, 496-503.

[32] Vertegaal, R., Slagter, R., van der Veer, G. and Nijholt, A. Eye gaze patterns in conversations: There is more the conversational agents than meets the eyes; In *Proceedings of the CHI 2001 Conference on Human factors in Computing Systems,* New York, NY, ACM Press, 2001, 301–308.

[33] Ware, C. and Mikaelian, H. H. An evaluation of an eye tracker as a device for computer input. In *Proceedings of the ACM CHI+GI'87 Conference on Human Factors in Computing Systems*, ACM Press, New York, NY, 1987,. 183-188.

[34] Zhai, S., Morimoto, C. and Ihde, S. Manual and gaze input cascaded (MAGIC) pointing. In *Proceedings of the ACM CHI'99 Conference on Human Factors in Computing Systems,* ACM Press, New York, NY, May 15-20, 1999, 246-253