ADAPTIVE MULTIMODAL INTEGRATION OF SPEECH AND GAZE

By

CHANDRA SEKHAR MANTRAVADI

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Electrical and Computer Engineering

written under the direction of

Prof. Joseph Wilder and Prof. Marilyn Tremaine

and approved by

New Brunswick, New Jersey

[October, 2009]

ABSTRACT OF THE DISSERTATION

Adaptive Multimodal Integration of Speech and Gaze

By Chandra Sekhar Mantravadi

Dissertation Directors:
Prof. Joseph Wilder and Prof. Marilyn Tremaine

Speech has been used as the foundation for many human/machine interactive systems to convey the user's intent to the system. However, other input mechanisms, commonly called modalities, such as gaze, touch, and hand gestures, have been explored as a means of providing a more robust interaction in environments where speech alone is not adequate. By combining the inputs from multiple, complementary modalities, none of which is perfectly reliable, a better understanding of the user's true intent can be imparted to the system. In this dissertation, the effectiveness of using gaze (where someone is looking) to aid speech in providing the user's intent to the machine is explored. To create a speech and gaze integration model, two human factors experiments were conducted to collect data for building this model. The first experiment had the user read a single word displayed on a screen, and the second experiment required the user to read a designated word from a menu of words. Speech onset time and the user's gaze patterns data were captured and analyzed to understand the timing relations between the two modalities. A set of gaze/speech features were extracted from the data and used to predict the location of the word that the user read. The best features and the best model for predicting the location of the target word were found through an iterative trial and error process. A linear model was able to predict the gaze location of the target as well as any of the non-linear models considered. The linear system representation was then used to create an adaptive model using the Row Action Projection (RAP) technique. The RAP adaptation model was found to predict the user’s intent with higher probability for the majority subjects than the non-adaptive approaches. The RAP model adapted to the speech/gaze patterns of each individual user as well as the variation in a single user’s interaction behavior over time. It was also found that the feature set used for successfully identifying the target in Experiment 1, a simple isolated word task, was different than that used in Experiment 2, a more complex menu selection task, suggesting that task complexity was an important consideration in the design of a speech/gaze interface. In summary, this dissertation has shown that an adaptive gaze and speech integration model is better than speech or gaze performance alone.

DEDICATION

To

Vardhani, Bala Subrahmanyam

Kameswari Devi, Kriti Saraswati, and Chinmay Sravankumar

ACKNOWLEDGEMENTS

First I would like to thank Prof. Joe Wilder and Prof. Marilyn Tremaine for their invaluable help and support throughout my dissertation work. It has been really a long and complex journey for me that I couldn’t have completed without the help of either of them.

I would like to give special thanks to Prof. Wilder for his enormous patience in helping me all the years I spent on my dissertation work. He is an amazing professor and scholar. I am extremely lucky to have had him as my advisor. He helped in every step of the way, and I will remember this support for my entire life. I feel that his insight and advice significantly helped shape my thought processes and clarify my research. I can not thank him enough, ever!

I would also like to thank Prof. Tremaine for her insightful ideas, planning and discussions during my dissertation work. She is a great professor and scholar that I am glad to have as an advisor. She guided my dissertation at the level of detail it needed.

I would also like to give special thanks to Prof. Mammone who stood by me and supported my dissertation research approach. He helped immensely during my modeling work and also in the final stages of my defense. He is another excellent professor I can never forget in my life.

I would also like to thank Prof. Rabiner for his insistence on quality and making sure my dissertation was an accurate demonstration of its conclusions. It was his insistence on writing and research presentation clarity that helped to make this dissertation as good as it is. If there are problems in presentation or argument, they are only because I did not interpret his advice adequately. I thank Prof. Gwizdka for taking time to walk through my dissertation in detail and provide me with his invaluable comments. I would also like to thank Prof. Marsic for his time and effort in helping to shape my research.

Next, I would like to really thank my manager at work, Mr. Kenneth Settle, who helped

me all these years. He supported me at work giving me time to work on my dissertation and also gave me moral support in this long journey. He is one of the few people I will always remember till my last breath. I would like to thank ISCAN for providing me with the equipment and updates as and when needed.

Definitely not least, I would like to thank my family and friends for their enormous support in making this dream come true for me. I have to say that my mother believed that I would finish despite the many years it has taken me. Without her faith in me I doubt that I would ever have finished. I would like to make special mention of my wife and kids who must have missed me all these years. I couldn’t have finished without their invaluable support. Also, I would like to thank Dr. Sharma and Dr. Sharada for their invaluable help and support during the final days of my defense as well as throughout my life. I would like to thank all my subjects who helped me carrying out the experiments without which I wouldn’t have this dissertation.

Table of Contents

ABSTRACT OF THE DISSERTATION ii

DEDICATION iv

ACKNOWLEDGEMENTS v

Table of Contents vii

List of Figures x

List of Tables xii

1. Introduction 1

2. Multimodal Systems and Integration Techniques 5

2.1. Introduction 5

2.2. Multimodal Systems 5

2.3. Factors Influencing Information Fusion 9

2.4. Review of Fusion Techniques 13

2.5. Design Requirements Considered in this Research 20

2.6. Summary 23

3. Experiment 1: One Word Task 25

3.1 Introduction 25

3.2 Hypotheses 25

3.3 Task Description 25

3.4 Experimental Procedure 28

3.4.1 Speech Training and Gaze Calibration 28

3.4.2 Subject Population 30

3.5 Summary 31

4. Experiment 1: Results 32

4.1 Introduction 32

4.2 Fixation Features 33

4.3 Dominant Gaze Features 35

4.4 Need for Adaptation 37

4.5 Linear Prediction Model 38

4.6 Adaptive Prediction Model 43

4.7 Comparison of Linear and Adaptive Prediction 48

4.8 Performance Comparisons of Speech, Gaze, Adaptive and Non-Adaptive Approaches to Target Detection 50

4.8.1 Target Detection Approaches Compared 50

4.8.2 Statistical Analysis: Paired Samples t-Tests 51
4.8.3 Interface Usability Performance 55
4.9 Summary 58
5. Experiment 2: Menu Task 60
5.1 Introduction 60
5.2 Hypotheses 60
5.3 Task Description 61
5.4 Experimental Procedure 68
5.4.1 Gaze Calibration 68
5.4.2 Subject Population 70
5.5 Summary 70
6. Experiment 2: Results 71
6.1 Introduction 71
6.2 Dominant Gaze Features in Menu Task 71
6.3 Adaptation Coefficients 74
6.4 Experiment 2 Performance 77
6.4.1 Effect of font size and spacing on performance 77
6.4.2 Statistical Analysis: Paired Samples t-Tests 78
6.4.3 Interface Usability Performance 81
6.4.4 Kruskal-Wallis Test for Spacing and Font Size Effects 84
6.4 Summary 86
7. Task Modeling 90
7.1. Introduction 90
7.2 ‘fi’ – better task modeler than attention predictor? 90
7.3 Summary 92
8. Summary of Contributions 94
8.1. Introduction 94
8.2. Contributions 94
9. Conclusions and Future Work 98
Appendix A. Review of Multimodal Systems 101
Appendix B. Word List in Experiment 1 103
Appendix C. Word List in Experiment 2 105
Appendix D. System Description 107
Appendix E. Data Capture and Validation 111
E.1. Raw Gaze Data 111
E.2. Speech Machine Data 112

E.3. Experiment 1 Data Capture 116
E.4. Experiment 2 Data Capture 122
E.5. Gaze/Speech Data Preprocessing 128
E.6. Scanpath Analysis 131
E.7. Search Discriminants 138
E.8. Interaction Scanpaths 141
E.9. Experiment Limitations/Design Choices 148
E.9.1. Effect of Speech Training/Recognition 150
E.9.2. Effect of Gaze Calibration 153
E.9.3. Effect of Word Timestamps on Fixation Identification 153
REFERENCES 155
Curriculum Vita 167

List of Figures

Figure 1 No Interference Word Display 27

Figure 2 Gaze Calibration Instructions 29

Figure 3 Gaze Calibration Panel 29

Figure 4 Gaze Calibration Accuracy Panel 30

Figure 5 Dominant Gaze Features for all Subjects in One Word Task 36

Figure 6 Dominant Gaze Variations Over Time 37

Figure 7 Speech and Gaze Interaction Model, S_s = speech start time, -TM1 and TM2 bound the time window (+/- 1500msec) around S_s. 39

Figure 8 RAP Algorithm – An x_0 converging to a solution x_3 after 3 projections (1 iteration) 46

Figure 9 Pseudo code of RAP Algorithm 47

Figure 10 Comparison of Linear and Adaptive Prediction 48

Figure 11 L2 distance of predicted location from the target 49

Figure 12 Target Detection Probabilities in One Word Task 51

Figure 13 Interface Performance for all 39 subjects in One Word Task 57

Figure 14 Interface Performance for (9 native and 9 non-native) Speakers in One Word Task 58

Figure 15 Menu Item Letter 63

Figure 16 Inner 4x4 Array (the inner array is NOT highlighted in the actual experiment screen) 63

Figure 17 Small-Font Small-Spacing 65

Figure 18 Large-Font Large-Spacing 65

Figure 19 Upper-Left Screen 66

Figure 20 Lower-Right Screen 66

Figure 21 Upper-Right Screen 67

Figure 22 Lower-Left Screen 67

Figure 23 Gaze Calibration Instructions 68

Figure 24 Gaze Calibration Accuracy Panel 69

Figure 25 Dominant Gaze Feature for All Subjects in Menu Task 72

Figure 26 Dominant Gaze for each Session in Menu Interaction 73

Figure 27 RAP Coefficients for Predicting x coordinate of the gaze location (x, y) 75

Figure 28 RAP Coefficients for Predicting y coordinate of the gaze location (x, y) 76

Figure 29 Session Performances (Gaze/RAP) in Menu Interaction (fixed parameters) 77

Figure 30 Session Performances (Gaze/RAP) in Menu Interaction (optimal parameters) 78

Figure 31 Usability curves (fixed parameters/all subjects) 82

Figure 32 Usability curves (optimal parameters/all subjects) 82

Figure 33 Interface Performance for 5 Native/5 non-Native Speakers (fixed parameters) 83

Figure 34 Interface Performance for 5 Native/5 non-Native Speakers (optimum parameters)....................84

Figure 35 Fixation Intensity Profiles in Experiment 1 (Simple Task) with different dispersion thresholds (DT)..91

Figure 36 *fi* as a Task Complexity modeler instead of attention predictor..92

Figure 37 System Installation for Speech/Gaze Interaction Experiments...108

Figure 38 ISCAN Interface 109

Figure 39 Raw Gaze Data 112

Figure 40 Scanpath classification using Recognition Categories in Constraint-Free Task........................117

Figure 41 Preprocessing of Event Data...128

Figure 42 Data Alignment Algorithm 129

Figure 43 Gaze Sample Local Timestamp Differences..131

Figure 44 Gaze Sample Global Timestamp Differences..131

Figure 45 Good Scanpath Gaze Interpolation 133

Figure 46 Bad Scanpath Gaze Interpolation..134

Figure 47 Bad Scanpath Gaze Interpolation (blink interpolated)...135

Figure 48 Bad Scanpath Parameters (blink interpolated)...135

Figure 49 Scanpath Detailed Analysis 137

Figure 50 Scanpath Detailed Analysis of Small Word...139

Figure 51 Scanpath Detailed Analysis of Medium Word...139

Figure 52 Scanpath Detailed Analysis of Large Word...140

Figure 53 Scanpath of Word Reading Experiment – raw gaze reaching cursor.................................142

Figure 54 Scanpath of Word Reading Experiment – raw gaze before speech start............................142

Figure 55 Scanpath of Word Reading Experiment – raw gaze after speech start...............................143

Figure 56 Scanpath of Word Reading Experiment – interpolated gaze reaching cursor.....................143

Figure 57 Scanpath of Word Reading Experiment – interpolated gaze before speech start.................144

Figure 58 Scanpath of Word Reading Experiment – interpolated gaze after speech start...................144

Figure 59 Scanpath of Menu System Interaction – raw gaze reaching cursor....................................145

Figure 60 Scanpath of Menu System Interaction – raw gaze before speech start 145

Figure 61 Scanpath of Menu System Interaction – raw gaze after speech start 146

Figure 62 Scanpath of Menu System Interaction – interpolated gaze reaching cursor........................146

Figure 63 Scanpath of Menu System Interaction – interpolated gaze before speech start 147

Figure 64 Scanpath of Menu System Interaction – interpolated gaze after speech start 147

Figure 65 Word Length Statistics in Constraint-Free Interaction 149

Figure 66 Recognizer Word Timestamp Analysis 154

List of Tables

Table 1 Experiment 1 Subject Population Characteristics....31

Table 2 Fixation Features....35

Table 3 Different Target Detection Probabilities....50

Table 4 μ ⁄ σ of Target Detection Probabilities in Experiment 1 (all subjects)....52

Table 5 μ ⁄ σ of Target Detection Probabilities in Experiment 1 (only native subjects)....53

Table 6 *p*-values of paired samples t-Test in Experiment 1....53

Table 7 Menu System Experiment Sessions....62

Table 8 Experiment 2 Subject Population Characteristics....70

Table 9 Adaptation Coefficients....74

Table 10 μ ⁄ σ of Target Detection Probabilities in Experiment 2 (all subjects)....79

Table 11 μ ⁄ σ of Target Detection Probabilities in Experiment 2 (non-native vs. native subjects)....79

Table 12 *p*-values of paired samples t-Test in Experiment 2....79

Table 13 Spacing (in pixels) effect on target detection probability in Experiment 2....85

Table 14 Font-size (in points) effect on target detection probability in Experiment 2....86

Table 15 One syllable words....103

Table 16 Two syllable words....103

Table 17 Three syllable words....104

Table 18 Four and Five syllable words....104

Table 19 Speech Recognition Categories....114

Table 20 Speech Machine Event Tuple....114

Table 21 Scanpath Recognition Analysis in One Word Task....118

Table 22 Performance of the Speech Recognizer in One Word Task....119

Table 23 Raw Speech Machine Data – Marker Display / Dismissal in Experiment 1....120

Table 24 Raw Speech Machine Data – Word Display / Speech Start in Experiment 1....120

Table 25 Raw Speech Machine Data – Speech Start / Stop in Experiment 1....121

Table 26 Raw Speech Machine Data – Recognizer Finalization in Experiment 1....121

Table 27 Raw Speech Machine Data – Recognizer API events / Finalization in Experiment 1....122

Table 28 Recognition Performance in Menu System for a subset of users in Experiment 2....124

Table 29 Raw Speech Machine Data – Marker Display / Dismissal in Experiment 2....125

Table 30 Raw Speech Machine Data – Word Display / Speech Start in Experiment 2....126

Table 31 Raw Speech Machine Data –Speech Start / Stop in Experiment 2....127

Table 32 Raw Speech Machine Data – Recognizer Finalization in Experiment 2....127

Table 33 Raw Speech Machine Data – Recognizer API events / Finalization in Experiment 2....127

Table 34 Gaze Categories..132

Table 35 Search Region Discriminants’ Variables...140

Table 36 Search Region Discriminants 141

Table 37 Words Statistics...148

1. Introduction

Human-to-human communication can take many forms (*e.g.*, by various hand gestures, speech, gaze, and touch). When these mechanisms are used to provide input to a computer system, they are typically referred to as a communication *modality*. When humans use more than one modality to communicate to a computer, the computer system is called a multimodal system. After keying (keyboard typing) and mousing (mouse pointing), speech has proven to be the single most widely used modality to communicate with computers ([2], [3], [4], and [5]). In speech-based systems, users say what they want and get an appropriate response from the computer. Although speech recognition systems [116] have been around a long time, these systems work well primarily in controlled environments with a pre defined grammar. The dialogue between the user and the computer (i.e., voice response system) is not *natural human-to-human conversation*. It is a code language established between humans and computers to perform a series of tasks. Increasingly, automatic speech recognition systems are being introduced into human/machine systems to handle these routine interactions. Some of these speech-based systems can also resolve acoustic signal level ambiguities [17] and improve the conversation with the machine by converting user's speech input into queries. In less controlled environments, however, speech-based systems perform less well due to a variety of interferences, such as high levels of background noise, the high variability in human speech patterns, and the difficulty users have with an interface that requires them to use a specific command grammar or restricts their request to a limited number of utterances. For such applications, the notion of augmenting speech with additional modalities (as well as improving a machines' *language understanding* capabilities, a significant area of research by itself) is postulated in this research to be likely to improve speech recognition and, in the process, the overall human/computer interaction.

Other modalities of human-to-human communication interaction, besides speech, such as gaze (where a person is looking), touch, gestures, *etc.* are typically used to complement speech. In some cases, these modalities can convey the meaning without speech (e.g., pointing at something unusual may communicate as much as saying “look at that!” Recent advances in eye tracking have made it possible to use gaze as an additional input modality in many multimodal systems. These speech and gaze systems are designed to solve specific problems (*e.g.*, for cellular phones [13], [67]) and to operate in highly controlled environments. However, gaze and speech combinations in a more natural user environment suffer from a variety of factors that make it difficult to interpret results. Visual variability in natural environments may affect a user’s gaze patterns and variations in attending to other visual activities may also create a high variability in gaze behavior. For example, fatigue related changes in a user’s gaze patterns over time, the type of task the user is performing, the display size on which the user observes the task can significantly influence the interaction patterns in gaze and speech ([11], [19], [25], [48], [55], [56], and [57]). Therefore, it is necessary to develop an integration model that can track user and task variations and adapt to them over time [58].

There are a number of applications where a speech/gaze system might provide a significant improvement over a speech-alone system, in particular where hands-free communication is required to convey the user’s intent. One example is in the operating room, where a surgeon, while operating, wants to consult with a pathologist at a remote location while the surgeon and the pathologists are both looking at displays showing the same X-Ray. The object of the surgeons gaze at a point on his display will be marked by a cursor on the pathologist’s display, while they discuss what needs to be done. There are also a number of examples of in-vehicle navigation and peripheral function systems where hands-free interactions are required in noisy environments, e.g. in helicopters, trucks and cars. Research into in-car systems has gained popularity in recent years, *e.g.*, a menu search interface investigated in Italy [63]. A car navigation system is a typical example of a human/computer interactive system which contains all the parameters of interest for designing gaze/speech multimodal systems [43]. Adding gaze input to an in-car system is not expected to impact driving safety because the plan is not to require any conscious control of the gaze activity, that is, the measurement of gaze will be unobtrusive. A driver already performs many visual tasks in a car besides looking at the road (*e.g.*, looking at side mirrors and the audio system controls) taking eyes off the road for a considerable amount of time, up to 1.6 seconds ([130] Wierwille 1993). Gaze, which is a very fast modality compared to the other modalities, should actually reduce the visual distractions produced by manual adjustment tasks such as controlling the audio system. The speech application already controls peripheral functions in a car environment in a nearly-natural way. However, using speech alone runs into issues because high levels of noise in a car (e.g., due to a number of occupants talking at once while the driver is attempting to issue a command to the Speech/Gaze system) cannot be filtered out even with microphone arrays and noise canceling systems. Other issues affecting speech recognition performance include the large and diverse distribution of accents in today's world and the high variability in the spoken interactions (*e.g.*, a tired user, a user with a cold *etc.*). In addition, a driver's cognitive attention is likely to be elsewhere so that speech commands may be forgotten, garbled or both. Also, any cognitive load and conversation inaccuracies in a speech-only system could potentially impact the driver's performance.

In summary, the goals of this dissertation are to:

- **discover the basic speech and gaze interrelationships for the task of reading a single word from a computer screen**
- **discover the speech and gaze interrelationships for the task of reading a word from a menu displayed on a computer screen**
- **illustrate the effectiveness of using gaze to improve speech recognition**
- **develop an adaptive algorithm for fusing speech and gaze modalities**
- **demonstrate that the adaptive algorithm is more effective than simpler models or speech alone**

- **demonstrate that the adaptive algorithm can be used in a real-time, time-critical multimodal system**

To accomplish these goals, two human factors experiments are performed in which speech/gaze data involved in human/computer interactions are collected and analyzed. Also, prediction/adaptation models are developed leading to the design of an adaptive algorithm for these speech/gaze interactions.

The dissertation is organized as follows. Chapter 2 gives a review of the literature examining the current multimodal systems and the algorithms currently developed for combining different modalities. Chapter 3 presents the experiment design, system setup and data collection process for Experiment 1, the task of reading a single word from a computer screen. Chapter 4 discusses the results obtained in Experiment 1 and postulates and defines the possible features that can be used to predict gaze/speech recognition for this relatively simple task. In addition to demonstrating the advantages of a Speech/Gaze system over speech alone, the intent is to demonstrate that adaptive prediction is necessary and improves the overall prediction accuracy in ascertaining the user's intent for a simple one-word task. Chapter 5 presents the experiment design, system setup and data collection process for Experiment 2, the task of reading a word from a menu displayed on a computer screen. Chapter 6 discusses the results obtained in Experiment 2 and postulates and defines the possible features that can be used to predict gaze/speech recognition for this more complex task. As with the simple task, in addition to demonstrating the advantages of a Speech/Gaze system over speech alone, the intent is to demonstrate that adaptive prediction in a Speech/Gaze system is necessary and improves the overall prediction accuracy in ascertaining the user's intent for a more complex menu selection task. Chapter 7 examines the effectiveness of the adaptive gaze/speech model via comparison to the other generated models. Chapter 8 lists the dissertation contributions and Chapter 9 concludes with proposed future work.

2. Multimodal Systems and Integration Techniques

2.1. Introduction

In this chapter, definitions of the basic terminology used in describing multimodal systems are presented. This is followed by a review of how eye-tracking has been used successfully in association with computer systems for such areas as alerting the computer system to the user's intent or providing user interaction information for the better design of computer systems. This is followed by a review of other multimodal systems that combine a variety of modalities, finally looking at the current work that has been done in fusing gaze with speech for use as multimodal input to a computer. Because the approach in this dissertation focuses on adaptive fusion, a review of multimodal integration techniques is presented. It discusses both their advantages and potential disadvantages of each approach in light of the focus of this dissertation. Finally, the design requirements for a multimodal gaze/speech system are presented. These requirements are based on various multimodal systems studied in this dissertation [see Appendix A for a comprehensive listing of these systems].

2.2. Multimodal Systems

The physical act of a human interacting with a machine is a complex phenomenon in which the human can use a variety of muscle-controlled mechanisms to communicate with the machine and, in turn, the machine can display its return communication by a variety of techniques which can be understood by the sensory systems of the human. If a human communicates with a machine using only one means of input, *e.g.*, by typing information on a keyboard, the input system is then referred to as a unimodal input system. If, however, a human communicates with a machine using multiple means of input, *e.g.*, by speaking and by typing on a keyboard, the input system is then called a multimodal system. Typically, machine output is not referred to as multimodal although a machine can generate sounds, use visual displays and modify touched areas to use more than one human sensory system for communicating. Thus, a human-machine system is referred to as multimodal if the input from the human uses more than one computer input mechanism. Note that this definition of modality is different from that used for human modalities. When a modality is referred to in a human, it typically means a different sensing system for input, *e.g.*, the hearing modality *vs.* the seeing (visual) modality. Because of the nature of machines having multiple mechanisms for input, each of these is considered a separate modality. For example, a mouse input is one modality, but if a second mouse were used, making the system a two-handed input system, the machine system would be multimodal, having two inputs that are of exactly the same form. In this dissertation, therefore, when the word “modality” is used, it implies another input channel to the machine, not a different operational behavior in the human, although differences in this behavior will generally be the case. Also, note that because a human is producing the inputs using the same cognitive systems for the multiple input productions, it is not surprising that the inputs are in some way interrelated. That is, a gesture by a human is likely to be correlated with the human’s speech. Thus, multimodal input systems needs to address the interaction of the user with the machine using various modalities (*e.g.*, speech, gaze, gestures, *etc*.) and also the interaction of the modalities in the human generation of the input.

The term “gaze” refers to where a person is looking. In terms of interaction with a computer system, this typically means “where” on the screen display the user’s central visual focus is placed. With today’s high resolution eye trackers, the x-y position of gaze can refer to a single pixel. This ability to measure gaze precisely means that it can be used as an input mechanism for a computer system with some caveats. Since people use their eyes to acquire information, the eyes tend to jump around a computer screen extensively and often are not in complete conscious control of a user. Thus, although gaze can serve as a separate input modality for a computer, its use can be noisy and error prone. Gaze is a human behavior that has been studied extensively

([72], [97], and [99]). Many commercial eye/gaze trackers are available for use in conjunction with computer systems. They are thus available for combining with speech input to create multimodal systems. Gaze patterns have been used to predict the cognitive state of the user [51], as an input mechanism (*e.g.,* gaze typing [125]) especially in cases where the user is severely motor handicapped. Gaze patterns have also been used as a way of determining the intent of the user, e.g., by knowing who is intending to speak next and turning the appropriate microphone on, in studying how individuals search web pages ([84] and [85]), in error detection in task completion, as an aid to communicating intent in desktop video conferencing, in the detection of user attention in human/robot interaction [121], in supporting virtual/remote environments [81], as a way of performing a meeting analysis [111], as a user attention predictor ([100], [101], and [106]), a way of estimating the effect of different computer-based events in various user tasks, to assess the effect on user search in multi-resolution displays, in the analysis of common behavioral patterns [66], as a usability analysis tool, to analyze the effect of screen clutter, e.g., multiple animated displays on the TV news, and in a variety of other application domains. Tien and Atkins [75] developed a real-time gaze selection interface which demonstrates the feasibility of using gaze as an input mechanism in real-time systems. There exists a variety of data analysis tools for processing gaze data and processing the set of patterns created by eye movements. Monk and Watts [91] even found that gaze is a more reliable data channel than speech when video quality is poor. Clearly, after speech, gaze is gaining in popularity as an acceptable input modality and a variety of computer systems are now being built with the addition of eye movement measuring capabilities, suggesting that the use of gaze is feasible, inexpensive and easy to implement.

Speech has been integrated with gaze and other modalities ([5], [10], [12], [13], [14], [28], [33], [49], [71], and [87]) to design custom applications in several different application domains. Although speech recognition performance is continuously improving over the years, Tan *et. al.* [44] found that speech alone is not effective as an input modality. Faria [133] demonstrated that speech recognizers are typically heavily biased towards the specific accent of a speaker. Miniotas *et. al.*, [62] used speech in combination with gaze to demonstrate that a practical interface can be built that performs similarly to current computer interfaces with keyboard/mouse. Gaze has been used in conjunction with speech for data entry systems, user attention prediction ([100], [101], and [106]), spoken language processing, dialogue systems [96], discourse segmentation, the understanding of ocular expressions, redundancy / complementarity measures ([104] and [110]), pointing mechanisms [115], selection strategies [98], the investigation of natural conversational dialogues, the support of collaborative/virtual environments ([94], [112], and [120]), and as a reference resolution or disambiguation of speech ([76], [89], [95], [105], [107], [113], [114], [119], and [131]). This supports the idea that the eye patterns generated by gaze have useful information that can be used to disambiguate the speech modality.

Although many attempts have been undertaken to develop multimodal applications, only a few systems have developed an integration model for combining the multiple modalities to improve the overall recognition performance of user input. Only a limited few use an adaptive integration model for fusing modalities. Since the beginning of multimodal system development [15], many frameworks/systems emphasized the creation of a single architecture for solving input and output problems for two or more modalities. These frameworks/systems concentrated on the application development framework, with little or no emphasis on solving the adaptive modality fusion problem [58]. This dissertation treats this adaptive fusion as central to human/machine understanding. It focuses on being adaptive because of the high variability of behavior across humans and the high variability of behavior in an individual over time. The fusion part is important because, as indicated earlier, humans do not typically perform multiple motor movements in isolation from each other but in tandem. Thus, if one changes, the other parameters being measured also change. A few studies ([5], [8], and [38]) have concentrated on multimodal integration or fusion problems as central to human/machine understanding. QuickSet [33] emphasized the need for fusion architecture for multimodal integration. Oviatt [58] has pointed out the need for adaptive fusion to build an effective multimodal system. Several recently developed multimodal architectures focus on one or more modalities in communicating with the user [Appendix A]. These architectures support modalities like speech, gaze, touch, pen, gestures, etc. Many other multimodal systems have been built to address a wide range of modalities employing different integration techniques, user perception policies, human machine dialogue management mechanisms ([21], [41], [108], and [109]), output representations, interruption management [46], and life-like agents/robots [86] *etc.* Thus, the multimodal field is acknowledging that fusion models are an appropriate way to handle multimodal input, but not much research has been done on adaptive fusion systems. This dissertation therefore focuses on creating an adaptive fusion system for gaze and speech input.

2.3. Factors Influencing Information Fusion

Information fusion, or simply *fusion*, can be defined as the process of combining information from different multimodal inputs to create a meaningful decision which can be interpreted by the machine to carry out the task. It can also be called *multimodal integration.* Information fusion (or interchangeably multimodal integration) is a complex process which depends on several factors like characteristics of the modalities involved in the process, characteristic behavior patterns of users, the interrelationships between the modalities, *etc*. Typically, human-to-human interaction involves a single modality for low complexity tasks (e.g., speech) and two or more modalities for higher complexity tasks (e.g., speech and pointing) ([19] and [25]). However, some low complexity tasks may require more than one modality. In addition to task complexity, modality integration also depends on the task characteristics, the individual’s personal dominant integration pattern (i.e., the most frequently used modalities and the manner in which they are combined in communication) [48], the individual’s capacity to assimilate information and act on the surrounding environment [25], the history of information assimilated over the task time, and the information aging/decaying model employed. Considering these factors, modality integration is not limited to a simple correlation of incoming sensory signals which a person uses to make a meaningful fused decision. Therefore, to develop an effective multimodal interface, the information fusion architecture should consider (among several other factors):

- **developing an integration pattern suitable to the individual user** ([48], [55], [56], and [57])

  For example in a map-based flood management system with 6 male and 9 female users of ages from 66 to 86, the users performed three tasks with low/medium/high complexities using speech and pen inputs. The experiment(s) showed that users have different integration patterns. Some users used a single modality one at a time, *i.e., a sequential integration pattern* and some users used both modalities in producing *a simultaneous integration pattern.* Some users used both *sequential* and *simultaneous* integration patterns in carrying out the tasks. This suggests that the fusion architecture should be able to integrate input modalities adaptively (sequentially or simultaneously) depending on the interaction pattern of each user/task.

- **accounting for the user's ability to assimilate, retain, and retrieve information** ([55] and [56])

  The ability of a user in understanding a task can be related to the reaction time among several other factors. Seniors and Children differ in their reaction times due to age differences and exhibit different integration patterns in interacting with the machine. The reaction times also differ in carrying out tasks of different complexity within the same age group.

- **accounting for all possible user's integration patterns** ([12] and [48])

  Oviatt *et. al*. suggests that only 20% of the human-machine interaction patterns are of a point-and-speak nature and they depend on the individual. These factors illustrate the necessity of a fusion architecture that allows for many different integration patterns.

- **incorporating dynamic integration patterns for a single user** ([48], [55], and [57])

  Oviatt *et. al.* demonstrated in a speech and pen multimodal system that not all the users interact multimodally always. Users differ in their use of integration patterns due to several factors like task complexity, fatigue etc. Users typically develop a few integration patterns in interacting with the machine and they get fixed onto those patterns in carrying out their tasks. However, the usage of any pattern depends on the specific interaction constraints, and hence, the system should be able to identify the pattern dynamically in order to accurately predict the user's intent.
- **differing integration patterns based on the characteristics of the input modalities**

  The ICARE system [9] provides a conceptual model of categorizing the modalities into *elementary components* and *composition components*. Elementary components include low-level physical layer abstraction of the device corresponding to the modality and the interaction language components for logical level abstraction of the modality. Composition components provide the fusion mechanism through the concepts of *Complementarity* (*i.e.*, combining complementary data from two or more modalities close in time), *Redundancy* (*i.e.*, redundant information from two or more modalities close in time), *Assignment*, and *Equivalence* properties. *Assignment* and *Equivalence* are modeled as linkages between components instead of any specific properties of modalities. Some modalities/components are completely sufficient in expressing the user's intent while some other modalities require a complementary modality to complete the expression of the user. The fusion architecture should be able to handle the varying characteristics of the modalities in carrying out the user's task.
- **extracting correlations from multimodal inputs at the signal/feature level and subsequent semantically higher levels** [33]

  Multimodal systems can be broadly classified as two types namely those that fuse information at the signal level and those that fuse at the semantic level. Different modalities can be combined at signal/feature or semantic level in making a fused decision. Signal/feature based fusion architectures work better for closely coupled modalities like speech and lip movements while semantic fusion architectures scale well and support a wide range of application domains. The fusion architecture should be flexible enough to handle multimodal integration at all possible feature and semantic levels.

- **accounting for the task description and complexity** [19]

  Not all the tasks require multimodal interactions and not all users will be using multimodal interactions for the same task. In a speech/pen interface, three different types of tasks namely *general action tasks, selection tasks,* and *spatial location tasks* exhibit differences in user behavior in generating multimodal interactions. Spatial location tasks require a high percentage of multimodal interactions while general action tasks do not require high percentage of multimodal interactions. Selection tasks require a moderate percentage of multimodal interactions. Another speech/gesture multimodal system found that increase in task complexity and hence cognitive load, decreases the redundancy of information contained in modalities requiring all modalities to be used in a high complexity task.

- **accounting for the history of modality information during fusion** [58]

  Users typically exhibit different integration patterns depending on their natural behavior, task complexity and other ambience factors. However, they always select multimodal interactions that they have used before and this behavior pattern will be further refined and recalled often when the task is repeated. The multimodal system should be able to understand the distinct interaction patterns of a user based on the history of prior interactions and be able to predict the current interaction accurately.

- **compensating for inadequate training data for individual modalities** [1]

  A multimodal system is not guaranteed to have sufficient training data for all modalities involved in the interaction always. For example, a user may not have trained the system enough for the best results with speech recognition or may not have calibrated the gaze for accurate gaze data. The system should be able to predict the user's intent accurately even when there is not enough training data under these conditions.

- **managing the context and uncertainty of individual modalities and tasks** [27]

  Only few application domains exhibit special requirements on multimodal systems. Several multimodal integration mechanisms exist for map related applications, while a very few have concentrated on graphic design applications. In a graphic design application DPD, cross-channel correlations between speech and gesture are employed to build a fusion strategy based on parsing techniques. In particular the integration strategy takes care of managing the context and uncertainty for graphic design applications.

- **User fatigue and other ambience factors** [19]

  Users may exhibit different integration patterns depending on the prior knowledge of the system and fatigue levels. So, adaptive fusion architectures need to focus on modeling user fatigue and any ambience parameters of the application domain.

Thus, information fusion is a complex modality mixing process from an engineering point of view. It is a highly *complex adaptive cognitive process* dependent on the user's interaction, the command being executed, and the modalities involved, *etc*. So a dynamic rather than a simple static modality integration process is required.

2.4. Review of Fusion Techniques

The following describes some of the fusion techniques currently used in recent multimodal systems [Appendix A]:

- **Timing of fusion**

  Timing of fusion refers to the time when multiple modalities are combined by the multimodal system to make a logical decision in understanding the user's intent. Two techniques based on timing of fusion, namely early fusion (i.e., signal/feature level) and late fusion (i.e., semantic level) are used to integrate modalities. Early fusion means the modalities are combined at a very low level without much meaning derived from the modality data. Late fusion implies that modalities are combined after there is some semantics incorporated into the modality data. For example, when processing the gaze data, raw eye coordinates or fixations (i.e., centroids of clusters of raw eye coordinates) on the screen can be processed. A limitation of these techniques is that they do not allow for the ability to change the timing of the fusion or provide fusion at all possible semantic levels. Any adaptive fusion technique should not really be concerned with the exact timing of the fusion but instead it should automatically incorporate timing of fusion into the adaptation model. Moreover, a fixed timing may not suit all modalities, user behaviors, tasks, *etc.*

- **Decision level fusion**

  This technique employed in some systems [26] falls under the late fusion strategy. While a one second interval for fusing two modalities may be useful for speech and facial expressions, it is not a suitable strategy for general multimodal fusion. For example, in one second, gaze being a very fast modality can produce a large number of fixations (semantic level). The one second granularity is too large for determining the correct reference point of gaze on the screen. So an adaptive fusion model cannot simply be tied to fuse modalities at the decision level alone.

- **Unification-based fusion** ([12], [14], [90], [93], and [117])

  This, more widely used, technique integrates individual modality features (expressed in Typed Feature Structures *i.e.,* hierarchical collection of typed attribute/value pairs) into a single feature-set to be passed on to the next semantic layer. Most of the unification-based methods use temporal constraints which is rather simple and limiting to create a fused decision. Different constraints like temporal proximity can be employed to create an enriched semantic expression to a semantically higher layer. Fusion architecture should allow for the *specification of these constraints* suited for specific applications. While these techniques express the data representation/communication from signal level to decision level, several adaptive aspects (*e.g.*, user, task adaptation *etc.*) central to the fusion problem still need to explored to solve the generalized modality fusion.

- **Fuzzy Logic Model of Perception (FLMP)** ([20], [59], and [60])

  FLMP which is mathematically equivalent to Bayes' theorem is based on the concept that some computations in the brain are analogous to Bayes' theorem. It is based on a neural network model which assumes that the modality integration occurs in overlapping stages (*evaluation*, *integration* and *decision*) with streaming data between any two stages. While this is a feasible model for adaptive fusion, it is not clear whether it is computationally feasible. And a real-time implementation of an adaptive system may be more challenging. So the fusion architecture should consider a computationally feasible approach for a real-time adaptive integration model.

- **Frame / slots based fusion approach**

  This approach [5] falls under semantic fusion mechanisms with data structures similar to Typed Feature Structures (TFSs), where the fusion manager attempts to discover the *target*, *action* and *parameters* of a particular task. These attributes (target, action, and parameters) form the slots of a frame. The fusion manager tries to capture these attributes from a parse tree filled in by the context providers. The Context Provider is analogous to a data acquisition module for a specific modality. It is the fusion manager which fills in the slots of the frame by appropriately resolving the ambiguities and refining the attributes' information using redundant modalities. Although this provides a rapid application development framework, it needs an adaptive model for managing the multimodal integration.

- **Multimodal chart parsing techniques** [35]

  Casting multimodal integration as a parsing problem, *multimodal chart parsing techniques* have evolved to unify individual modalities to form an integrated decision. A chart parsing technique can be summarized as a union of discrete and linearly ordered input constituents using a rules-base. However, multimodal input streams do not fit the criteria. So a variation of chart parsing technique *multichart* has been proposed. All these chart parsing techniques are centralized around speech. Moreover, a rules based system may not be suitable to adapt to variations in user, task, and ambience conditions.

- **Multi-chart parsing**

  A multi-chart parsing strategy at the semantic level [32] fuses input modalities (i.e., speech and gesture) based on rules and manages a pool of TFSs, where new elements can be added to the pool and some can be removed. In integration iterations, not all elements are always included. This kind of rules based multimodal system may not easily adapt to variations in user, task, and ambience conditions.

- **Members Teams Committee** (MTC)

  A MTC technique uses a statistical, symbolic/semantic fusion mechanism as in the QuickSet ([29] and [33]) architecture. With *mode conditional input feature density functions* for integrating input modalities, QuickSet uses temporal, statistical, and semantic fusion strategies in that order. This technique unlike others includes many aspects of a feasible adaptive model. But it is not clear whether it can provide a computationally feasible and extensible adaptive model.

- **Hybrid approaches**

  Some hybrid approaches ([34] and [80]), which combine data driven and knowledge-based methods with rule-based methods, are aimed at integrating specific modalities like speech and gestures. These methods are limiting for a real-time adaptive model because they are based on a static or dynamic rules-set. The rule itself doesn’t have the concept of adaptation. The rules need to be exhaustive enough to account for all possible user, task, and ambience conditions which may not be possible.

- **Human-communicational-rhythm-based model**

  One interesting technique, human-communicational-rhythm-based model [36], found that humans communicate in a rhythmic manner. It models the rhythm in human-to-human communications to find the correlations between speech and gestural inputs. It uses a “tri-state rhythm model (swing-subside-wait)” to segment multimodal input streams and correlate them before passing on to the next semantic layer of understanding. Although it may be able to account for a majority of human-machine interactions, clearly it may not be feasible to predict the user’s intent accurately when the expected rhythm is missing in the interaction. Moreover, the rhythm may be disturbed by several factors like distractions, fatigue, *etc*.

- **Context-based and semantics-based multimodal integration** ([38] and [64])

  The PETE/COMIC system uses context-based multimodal integration, a rule-based integration approach, in which the user and machine have a *local turn context* containing the information of input modalities, history of modality events, and the dialogue state. The fusion technique is not clearly separated from the semantics of understanding i.e., it only provides an integrated decision while dialogue management handles the real conversation state. Another contextual multimodal integration technique uses entropy based techniques along with contextual information. Another semantics-based integration technique using speech/gesture system found that the multiple modalities become more complementary than redundant as the cognitive load increases. Semantic information is explored in yet another fusion technique using subspace learning techniques. Like other rule-based techniques, these semantic techniques are also not suitable for building an adaptive fusion model because the rules may not be sufficient to accurately model all possible variation of user, task, and ambience conditions. And moreover, the semantic information is highly task/application dependent.

- **Data-flow-based maximum entropy technique** ([42] and [57])

  The data-flow-based maximum entropy technique, which uses a maximum entropy framework, classifies the data from low-level signal to high level semantics into three features *bag-of-words, contextual features,* and *prepositional feature* for multimodal integration. Although it may be extensible to multiple sets of features, the classification of three kinds of features seems rather limiting. Also, it is not clear how these can handle an adaptive model of human-machine interaction.

- **Layered HMM technique** ([2] and [47])

  A Layered HMM technique is a cascaded network, where each layer is responsible for a specific temporal granularity. Each layer tries to analyze the information from input modalities at different temporal granularities to resolve any spatial and deictic references. While this may be another feasible approach for solving the adaptive fusion problem, it is restricted to the temporal domain and doesn't seem to account for task variations.

- **Gestalt principles of grouping information** [48]

  Techniques based on Gestalt's principles of grouping information have been used to analyze the production and perception of multimodal integration patterns. Humans adapt to the machines' recognizers easily and quickly so that the system understands our commands in contrast to machines understanding humans. For example, a user would increase the duration of an utterance or pause carefully between words/utterances for the speech recognizer to recognize them. Studies conducted in also describe a similar increase in utterance duration. These principles may guide the development of an adaptive integration model but empirical knowledge of human machine interaction is necessary to build a comprehensive integration model. So, human factors experiments need to be carried out for every modality which interacts with the machine. These principles in conjunction with the empirical knowledge may form the basis of an adaptive integration model. However, this dissertation employs a computationally feasible model with empirical knowledge instead of using these theoretical principles.

- **Finite State Model based methods**

  Finite state methods have been employed for multimodal input parsing, understanding and semantic feature extraction ([30] and [52]). Although these methods can potentially provide a basis for adaptive integration, it is not clear if these methods are computationally feasible with ever increasing modalities. In order to build a practical adaptive integration model, one has to choose a computationally less intensive and simpler model.

- **Active Memory Model**

  An Active Memory Model vision system [53] learns and retains information about objects of a scene in a multimodal system. It aims to represent the knowledge in the real-world scene from different sources as a systematic set of memory elements. These memory elements are organized and maintained by a memory infrastructure. Each memory element for a real-world object contains a *hypothesis* representing uncertainty, reliability, created/updated timestamps *etc*. The memory element's hypothesis allows creation of an information decaying model around the memory element. The active memory model vision system understands the real-world scene and creates a memory infrastructure around it from a visual system point of view alone. This model is complex and may not lend itself easily to all modalities.

- **Customized modality integration techniques** ([14] and [27])

  Customized modality integration techniques like spatial integration techniques, assume the completion of multimodal activity before fusing modalities. Another technique tries an optimal multimodal integration strategy specifically for graphic design tasks. Such techniques while useful to integrate specific modalities (e.g., speech/sketch and speech/pen) effectively, are not useful for a wide variety of integration patterns required for the majority of multimodal applications.

2.5. Design Requirements Considered in this Research

Historically, the multimodal integration problem has been perceived as a unification problem i.e., it merges the incoming information streams into a semantic information stream using different types of constraints. Lately, different approaches are being considered, treating it as a statistical problem of integrating independent mode feature densities, casting it as a parsing problem to create a higher level decision etc. Based on the aforementioned integration strategies it can be inferred that different integration techniques are employed based on the modalities involved and for specific applications. Some integration techniques are more suitable for some modalities and/or applications than others. Even modality characteristics play a vital role in the integration techniques required for a multimodal interface. A general architecture for multimodal system development should concentrate on developing a framework of fusion/integration mechanisms suitable for all modalities accounting for different characteristics. Thus, among several other factors, a general adaptive fusion architecture should be able to handle:

- **User-based, modality-based and task-based integration strategies**

  Integration patterns differ based on user behavioral patterns, different modalities used in the system, and different tasks the system has to perform. Any multimodal system cannot assume these to be static properties of the system because these can change with time. It should learn the user behaviors quickly and adapt to the ever changing scenarios of interactions.

- **Dynamic detection and planning of modalities’ usage**

  System errors and usage patterns of modalities can render one or more modalities unusable leaving the system to operate based on the available modalities. Moreover, the modalities and their characteristics are continuously/rapidly changing requiring the multimodal systems to be flexible to incorporate them with ease. The integration model should account for the presence or absence of information from various modalities and adapt to the availability of modalities dynamically.

- **Maintaining a modality, user, and task history for continuous adaptation**

  User behaviors are fairly predictable from a past history and the interaction patterns become more subtle when higher concentration levels are required. But the patterns repeat from past history of interactions. Thus, maintaining a history of interaction patterns will help quickly predict the new interactions.

- **Flexible data representation and information processing at different semantic levels**

  Information fusion is not a simple process of combining the information from multiple modalities at signal or semantic level for each interaction. Interaction modeling involves combining the information at various granularities at signal and semantic level. This requires different data representations and different processing strategies at various decision levels. Fusion architectures should provide all possible data representation and information processing mechanisms to allow for fusing information at any signal and semantic level.

- **Ambience conditions**

  Apart from users, tasks, and modalities the environment also plays a huge role in multimodal system effectiveness. Ambience noise is a very significant factor in rendering some modalities unusable sometimes. For example, when there are multiple acoustic sources near a multimodal system, the speech recognition may not accurately interpret the user’s speech commands. When ambient illumination changes significantly, the gaze data may not be recordable rendering the gaze modality unavailable. The multimodal system should be able to detect these ambient conditions through the modalities and intelligently decide not to use the modality that is affected in the decision making process.

Oviatt et. al., ([83] and [128]) emphasized the need for an adaptive information fusion in multimodal systems and demonstrated the strong need to have empirical knowledge to build practical models which can predict multimodal integration patterns. They have designed a system to study user adaptation in a speech/pen interface instead of adapting the system to the users' behavioral patterns [129]. FAME [88] architecture proposed a conceptual framework for building adaptive multimodal interfaces but it is too general to account for all cognitive processes that drive various multimodal interactions. Adaptive speech-only interfaces [92] are built to integrate speech into any existing applications. Perakakis [134] studied inactivity times in multimodal interactions and showed that users would use a modality that suits them for an efficient expression while exhibiting a bias towards speech. Matt and Pantic [68] designed an adaptive affective interface which contains many modalities including speech and gaze but did not treat the integration model separately. The interface adapts to the user's behavior at a macro level rather than understanding the low-level interaction models. Leah Findlater and Joanna McGrenere [74] designed an adaptive interface for small screens but it is not a multimodal system. Their multimodal system adaptation criterion is entirely different from the regular interface design and similar principles may not be applied in speech/gaze systems. Gajos et. al. [82] designed an adaptive toolbar interface to restructure the user interface based on user behavior but even it did not look into the integration model of speech and gaze. Moreover, the experiment task involves additional stimuli to the subject which could potentially change the user behavior. Also, task difficulty may influence the gaze behavior [124] and subsequently impact the speech/gaze adaptive model. Apart from the task complexity, an adaptive integration model must be able to compensate for user head movements. Moreover the far field speech recognition itself poses several problems [135] in recognizing speech accurately and thus requiring an additional modality for interface effectiveness. Another category of interfaces called *attentive interfaces* change the information present to the user dynamically but do not adapt to the users/tasks *etc*.

A majority of the multimodal systems mentioned earlier in this dissertation addressed the fusion problem as an integral part of the system development. A majority of the above fusion techniques are custom created and do not have a *generalized adaptive fusion model* behind the fusion of modalities. Even the self adaptive software systems [24] have not focused on adapting the system from *a speech/gaze cognitive process* standpoint. The *generalized adaptive fusion model* is a vast research area in itself and requires empirical knowledge from several disciplines apart from theoretical models. Theoretical foundations from psychology [47], general human machine interaction behaviors [70], and empirical knowledge from various disciplines concerning each modality involved need to come together to create a generalized adaptive fusion model.

Even the systems that included gaze are designed to suit particular application domains. Even the systems which included speech and gaze did not employ adaptive integration of speech and gaze in a multimodal system. In order to fully understand the speech and gaze interaction one has to explore the cognitive process in speech and gaze interactions ([39] and [54]). This research aims to create an adaptive integration model for speech and gaze by exploring speech and gaze interaction in general and as applied to real-world applications. It uses an adaptive technique, Row Action Projection (RAP, described later) [61] which is a computationally feasible approach coupled with the empirical knowledge derived from the human factors experiments to build a cognitive model for speech and gaze interactions. In later sections, two human factors experiments are described along with the RAP-based adaptive fusion model illustrating that the addition of the gaze modality to a speech interface will enhance the overall effectiveness of the system.

2.6. Summary

Several multimodal systems and integration techniques discussed thus far are either custom designed for the system at hand or have not treated the *adaptive integration* with the empirical knowledge factored into it. Speech, pen and gesture based multimodal systems have been built but do not use the information in a synchronous/adaptive manner. This dissertation aims to gather the empirical knowledge in speech and gaze integration to build an adaptive fusion model for using speech and gaze simultaneously in a multimodal system. It also studies the effect of display parameters like font-size, spacing, and location of objects on screens in multimodal interfaces that include speech and gaze and looks for optimal values for these parameters. In the next chapter the research will be described by first presenting the methodology that has been employed to gather the empirical data needed to analyze the gaze and speech patterns.

3. Experiment 1: One Word Task

3.1 Introduction

Speech recognition can itself pose an immense challenge to accurately recognize the user's spoken words because of numerous spoken languages, different accents/dialects of a language, different pronunciations of a single word by different users, difficulty in distinguishing the user's voice from multiple acoustic sources in the environment, and complex ambient conditions etc. Gaze is even more intractable because of its highly unpredictable nature. So, in order to understand the speech and gaze interrelationship, one has to first extract the very low level or fundamental behavior of speech/gaze interaction. The task in Experiment 1 has been designed to extract this low level behavior. The task avoids distraction (e.g., other objects) on a display screen when the subject is speaking a word. In this chapter the hypotheses, task, and procedure for Experiment 1 are described.

3.2 Hypotheses

Experiment 1 is expected to provide the fundamental design parameters to be used in creating predictive and adaptive models for speech and gaze interaction. There are two hypotheses tested in this experiment:

- Combining speech and gaze provides higher performance in human machine interaction than a speech-only system.
- It is possible to use gaze behavior around the onset of speech to predict the user's attention on the screen

3.3 Task Description

In each trial of this experiment, a cross-hair "+" (or marker) appears on the screen at a random location and when the subject looks at it, it disappears. Immediately after the cross-hair disappears, a word appears on the screen at another random location. The subject has been instructed to read the word. The subject does not know what word will be displayed. The system then recognizes the subject's speech and registers the word as recognized if spoken correctly or as if not spoken correctly. Cross-hairs and words are displayed randomly at different locations on the screen to insure that the subject will not be able to expect a particular display pattern which might potentially influence speech and gaze interaction. Each interaction, where the cross-hair and the word are displayed, constitutes a *trial* in the experiment. The experimental design separates one *trial* from another by having the subject look at the cross-hairs (*i.e.*, '+') before reading the word. Thus, each *trial* in this experiment is independent of any other *trial*. A fixed number of *trials* constitute a *run.* Experiment 1 consists of a series of *runs, at least 6*, with each *run* containing 20 trials for a total of at least 120 trials. Subjects are asked to perform more *runs* at the end of 6 *runs* on a screen *if they are comfortable*. This additional data helps in analyzing the fatigue levels in prolonged human machine interaction. Some subjects were able to participate in more than 6 *runs* which resulted in a different number of trials for these subjects. Also, not every *trial* in a *run* is useful in data analysis due to various errors e.g., missing responses from the speech recognizer. This causes the number of trials for each subject to be different in Experiment 1. However, only the first 100 trials of each subject are used in data analysis. Each *run*, of reading words, takes about 1-2 minutes with the total number of *runs* taking *at least* 10 minutes. The word font size is fixed at 36 and is known to be easily visible to the subjects in the age range used (*i.e.,* they all own drivers licenses and had acceptable vision corrected or not needing correction) without any strain on the eye. All the letters in the word are *lower case*. There are a total of 531 words [Appendix B] in Experiment 1 chosen to contain 1 to 5 syllables to analyze the effect of the number of syllables on speech/gaze interaction patterns. Figure 1 shows a single word trial as it appears on a large 20'' screen. Words are randomly chosen from the 531 words for display in this experiment.

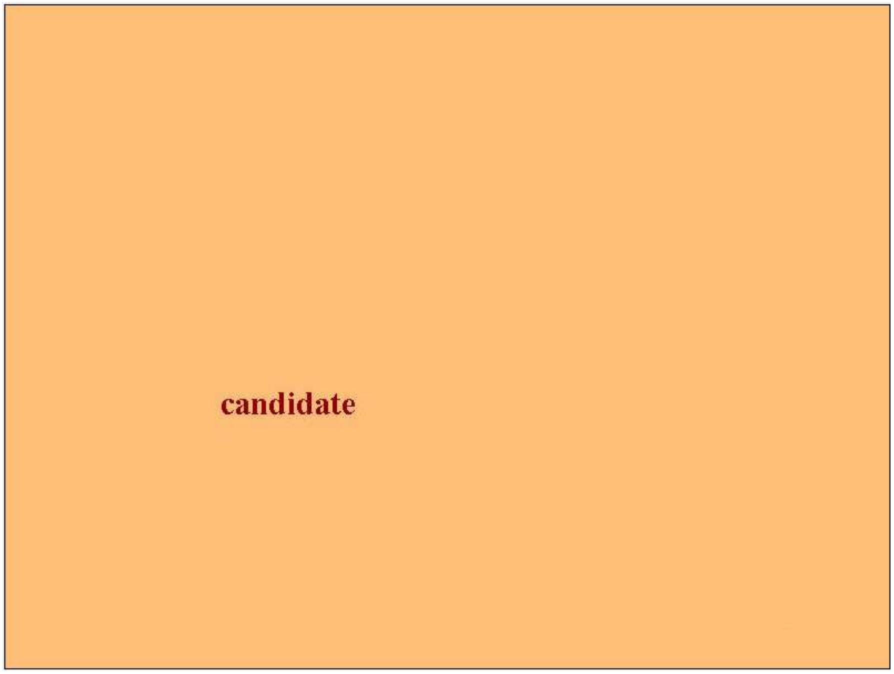


Figure 1 No Interference Word Display

Each trial is independent of other trials because the subject dismisses the cross-hair before performing the word reading task in each trial. The cross-hair ensures that the subject's eye is coming to the word from an arbitrary location in each trial instead of having to come from a previous word utterance. Both the display location of the cross-hair and the word is completely random. The screen space is used uniformly to display the cross-hair and the word instead of any specific affinity to any area on the screen. The random display pattern eliminates the subject's predictability of cross/word location. Each word stays on the screen for 3 seconds within which the subject is supposed to speak the word. If the word is not spoken or is not recognized by the recognition engine within 3 seconds, the word automatically disappears and the cross-hair appears, initiating the next trial. The 3 second limit is chosen because after 3 seconds the subject's eye is likely to be moving to another location on the screen invalidating the trial.

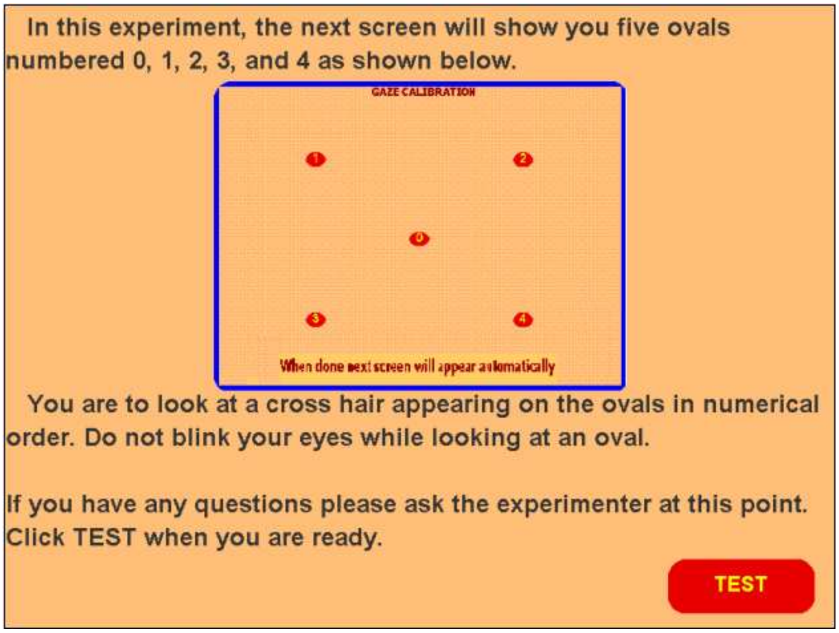
3.4 Experimental Procedure

3.4.1 Speech Training and Gaze Calibration

Each subject is debriefed about the speech/gaze experiments and has signed a consent form before starting the experiments. The subject is seated about 2 feet away from the screen in a chair that is fixed in front of the machine. A microphone array located under the screen is focused on the user's mouth. Although the eye tracker can track two eyes of the subject, its physical position is fixed and it is focused manually to the left or right eye by operating the eye tracker's controls.

After the subject is seated properly, the subject needs to train the speech recognition system to their voice before performing the experiments because the system uses the voice model developed during training to recognize the subject's utterances during experiments. The experiment is carried out using a Via Voice speech recognition system. Subjects read text to the speech machine to train the Via Voice system. This training session is used for the following experiment (Experiment 2) as well as for this one. The speech training process takes about 10-30 minutes for each subject to read about 57 sentences and is identical for both the experiments.

After the speech training session, gaze calibration instructions are displayed. Gaze needs to be calibrated before tracking the eye movements in speech/gaze interactions during experiments. Figure 2 shows the calibration instructions screen in Experiment 1.



In this experiment, the next screen will show you five ovals numbered 0, 1, 2, 3, and 4 as shown below.

You are to look at a cross hair appearing on the ovals in numerical order. Do not blink your eyes while looking at an oval.

If you have any questions please ask the experimenter at this point. Click TEST when you are ready.

Figure 2 Gaze Calibration Instructions

The gaze calibration screen contains 5 points as shown in Figure 3 which are marked 0 (center), 1(top-left), 2(top-right), 3(bottom-left), and 4(bottom-right). A cross-hair appears as shown in Figure 3 on ‘0’ (in the center) and then moves from 0 through to 4 and back to 0. The subject needs to focus and follow the cross-hair as it moves through this sequence.

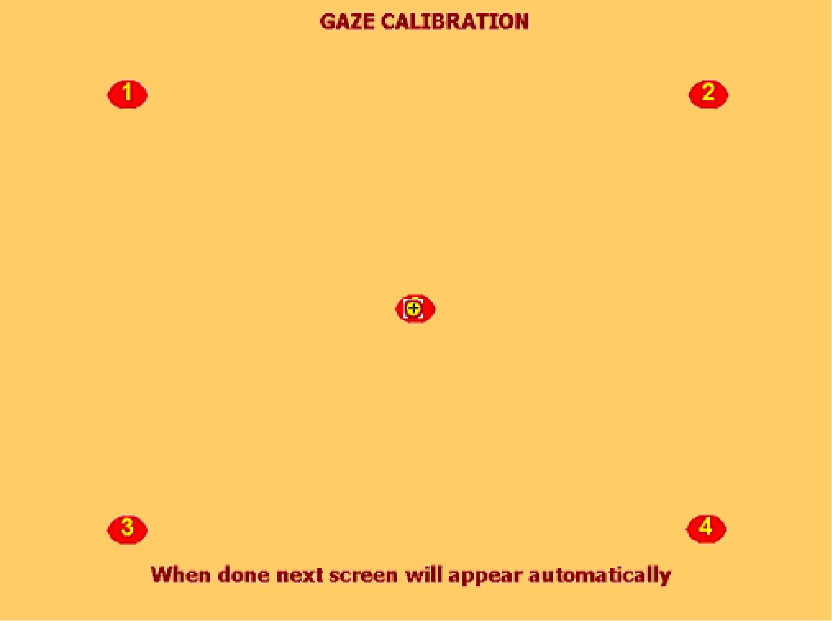
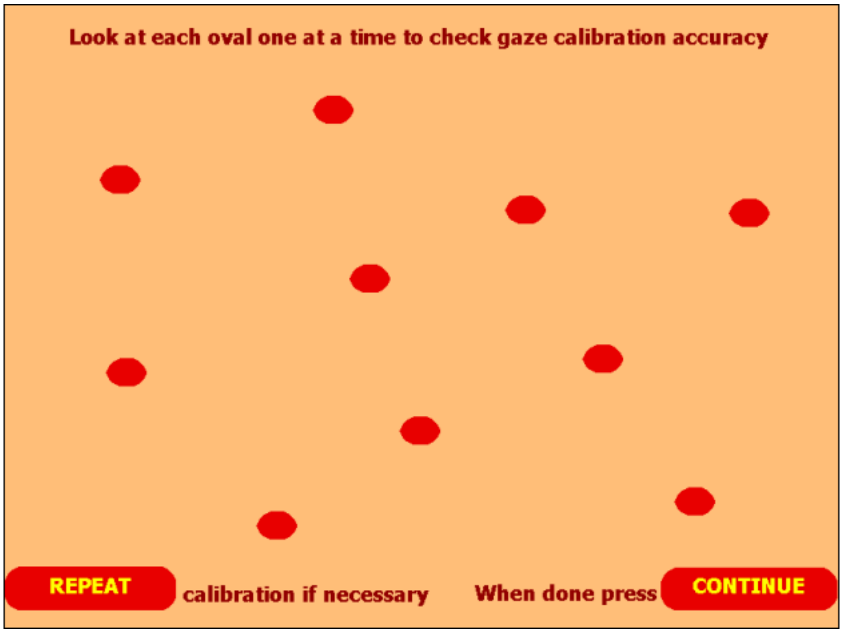


Figure 3 Gaze Calibration Panel

After the gaze calibration is complete, the subject performs a calibration accuracy check. Figure 4 shows the calibration accuracy verification screen in Experiment 1. The subject is instructed to look at the 10 points and the points on the screen change color (from red to blue) when they are looked at successfully (i.e., gaze falls close to the point target). The experimenter verifies the calibration accuracy by noting whether all of the points turn blue. If this is the case, the experimenter tells the subject to begin the experiment. Otherwise, the subject repeats the calibration. The experimenter's visual verification is deemed sufficient to check the accuracy. After the speech training, gaze calibration and calibration accuracy verification, the subject proceeds to perform the task of the experiment. See Appendix D for the system installation/setup for running the experiments and Appendix E for a thorough treatment of the data capture and validation.



Look at each oval one at a time to check gaze calibration accuracy

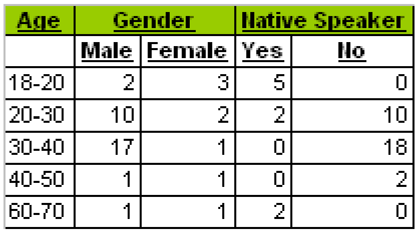
REPEAT calibration if necessary

When done press CONTINUE

Figure 4 Gaze Calibration Accuracy Panel

3.4.2 Subject Population

This experiment, and experiment 2 as well, is designed to be independent of the native language of the subjects or anything related to their origin. These experiments are designed to extract the fundamental behavior of users when they interact with the machine using speech and gaze. It is hypothesized that the speech and gaze interaction relationship is independent of the native language of the subjects. The speech recognizer's performance used here is known to be biased towards English speakers. Hence, the experiments are carried out with both native and non-native speakers to eliminate any bias in results. However, because the experiments are designed to extract the fundamental relationship between speech and gaze, the subject's native language is not a major concern. Table 1 describes the population of the 39 subjects that took part in Experiment 1. The subject population is a mix of gender, age, and native/non-native speakers. It includes some subjects wearing glasses or contact lens and some subjects that had a cataract operation.



Age	Gender		Native Speaker	
	Male	Female	Yes	No
18-20	2	3	5	0
20-30	10	2	2	10
30-40	17	1	0	18
40-50	1	1	0	2
60-70	1	1	2	0

Table 1 Experiment 1 Subject Population Characteristics

3.5 Summary

In this chapter, a speech/gaze experiment was described whose purpose was to extract empirical knowledge that contributes to the development to a speech/gaze interaction model. The design of the experiments, the procedures employed, and the subjects' characteristics were described. The issues/limitations involved in carrying out the experiments were addressed to ensure that the data collected is valid. The next chapter analyzes the data to establish parameters that are important in the development of a speech/gaze integration model.

4. Experiment 1: Results

4.1 Introduction

In this chapter, the data collected from the first of two experiments is analyzed to illustrate the differences in gaze patterns across subjects and across time for a single subject. Then, linear and adaptive prediction models of subject's behavior are compared. The intent of this work is to demonstrate that adaptive prediction is necessary and improves the overall prediction accuracy of the user's intent. To do this, the typical approaches to prediction are calculated and compared to an adaptive prediction approach. To this end, five different approaches to extracting the user's intent (speech only, gaze only, linear prediction, adaptive prediction, combined speech and adaptive prediction) are evaluated and compared.

Before beginning with the comparisons, it is necessary to lay out the data that is being analyzed and the analysis process that is taking place. In particular, this work uses what are known as fixations and saccades, characteristic eye movements. The algorithms used to aggregate the eye movement data into fixations and the list of particular fixations from which we select those most relevant to this experiment are defined in Section 4.2. In Section 4.3, dominant gaze features (fixations from the list developed in 4.2) are identified. Furthermore, it is shown that the dominant gaze features for a particular subject can vary over time. Section 4.4 discusses the need for adaptation. Section 4.5 introduces the Linear Prediction Model as a means for identifying the focus of a subject's attention based on gaze behavior. An adaptive prediction model of gaze behavior, based on the Row Action Projection algorithm, is introduced in Section 4.6 and its performance, in comparison with Linear Prediction is discussed in Section 4.7 Section 4.8 compares the performance of all of the approaches to target detection considered and their overall performance in interface applications. Section 4.9 summarizes the results of the first experiment.

4.2 Fixation Features

Each trial consists of a scanpath, i.e., a sequence of gaze samples. The gaze samples are then separated into saccades and fixations, where the saccades are those gaze samples associated with rapid eye movements across the display associated with the search for a target, and fixations are clusters of close-by samples associated with attention and focus on a found target (see [Appendix E] for details).. There are a number of algorithms for extracting the fixations from a collection of gaze samples [123]. Three widely used algorithms for detecting fixations (a cluster of samples given by their (X, Y) coordinates) use the following criteria:

Dispersion: (maxX – minX + maxY – minY) < dispersion threshold (DT) and cluster size > # of points (NP)

Velocity : v (i.e., the spacing between samples at a given sample rate) < velocity threshold (VT), and cluster size > # of points (NP)

Area : sum (i.e., the point to point distance) < area threshold (AT) and cluster size > # of points (NP)

It is known that the choice of fixation algorithm affects the data analysis [136]. The dispersion-based fixation algorithm is chosen in analyzing the gaze data in Experiment 1 and 2 because of its robustness in calculating the fixations accurately [123]. Also, the dispersion-based fixation algorithm's complexity and computational requirements are low enough for a real-time adaptive multimodal algorithm. In our experiments, we are concerned with the relation between the fixations occurring during visual search and the onset of speech. We define the following parameters:

(x_i ,y_i,t_i) = position on a display and time of occurrence of a gaze sample (sample rate for our gaze tracker = 60/second)

S_s = speech onset time stamp

fst = fixation start time stamp, (the time of the first gaze sample in a fixation) fet = fixation end time (the time of the last gaze sample in a fixation)

T = S_s - (fst+fet)/2 (fixation offset time w.r.t. speech onset time)

DT = maximum permissible spread in the location of samples for a valid fixation

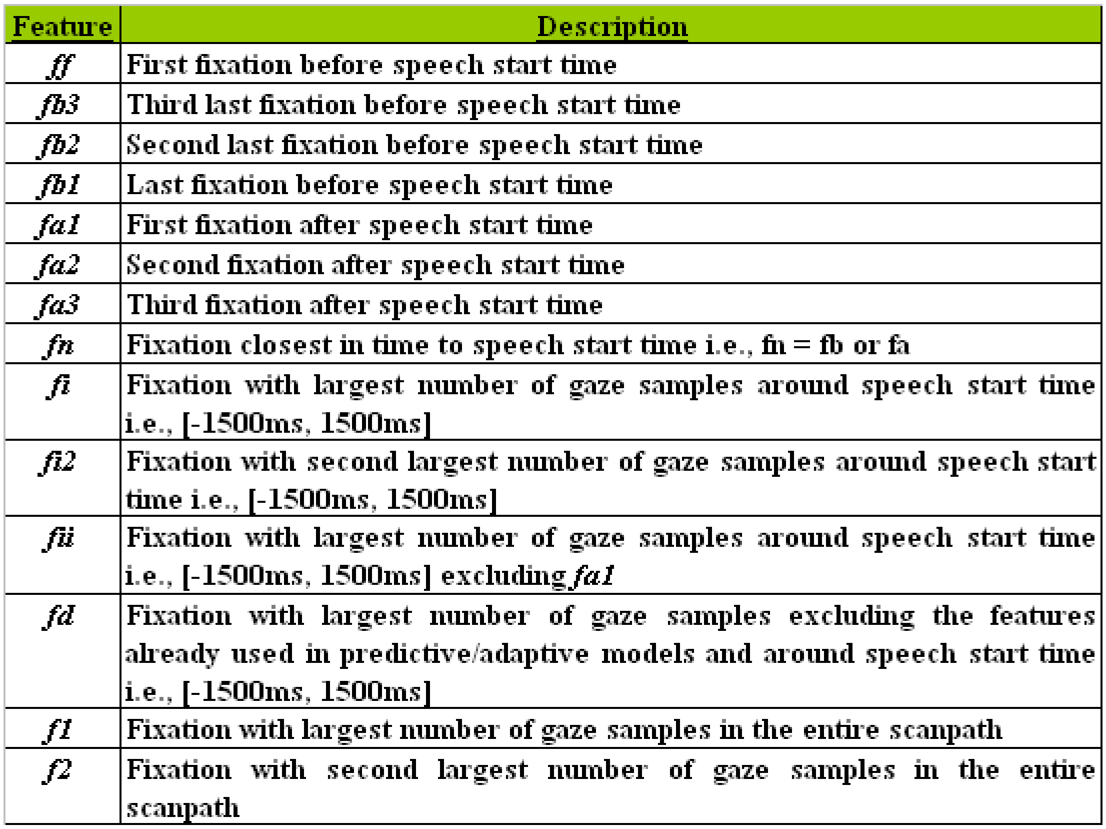
NP = minimum number of samples for a valid fixation

The following equation defines the fixation computation for the dispersion algorithm in our application. For all gaze samples (x_i, y_i), such that [max(x_i) - min(x_i)] + [max(y_i) - min(y_i)] < DT and N > NP, the location and time (w.r.t. speech onset time) is given by:

$$\mathbf{f(x, y, t) =} \left(\frac{\sum_{i=1}^{N} x_i}{N}, \frac{\sum_{i=1}^{N} y_i}{N}, \mathbf{T} \right)$$

Prasov *et. al* [106] showed that *fixation intensity*, *i.e.,* the number of gaze samples in a fixation, is one of the important features in understanding the gaze attention on an object on a screen. They found 1500ms is a sufficient time window around speech onset time, to look for fixations of importance to determine the user's intent. Fixation intensity is also a measure of fixation duration whose distribution has a positive skew [137]. In the current speech/gaze experiments, each fixation in [-1500ms, 1500ms] is given an index w.r.t. to speech onset time. Fixations that occur **before** the speech onset time are denoted by ***fb1***, the last fixation before speech onset time, ***fb2***,the second to last fixation before a speech onset time, ***fb3,*** the third to last and so on, (see Table 2), while the fixations occurring **after** speech onset time are denoted by ***fa1,*** the first fixation after speech onset time, ***fa2*** the second fixation after speech onset time, and so on, (see Table 2). Additionally, each scanpath has one fixation ***fi*** that has the largest number of gaze samples in [-1500ms, 1500ms].

The features to be used in speech/gaze integration model are selected based on their ability to track the interaction in all possible conditions *i.e.*, tasks, users *etc.* Speech/gaze based features include speech attributes (*e.g.*, speech onset time) and gaze attributes (*e.g.*, raw gaze samples' based features and fixations – clusters of gaze samples in space and time). Speech onset time categorizes the dispersion-based fixations, as shown in Table 2, for further data analysis.



Feature	Description
ff	First fixation before speech start time
fb3	Third last fixation before speech start time
fb2	Second last fixation before speech start time
fb1	Last fixation before speech start time
fa1	First fixation after speech start time
fa2	Second fixation after speech start time
fa3	Third fixation after speech start time
fn	Fixation closest in time to speech start time i.e., fn = fb or fa
fi	Fixation with largest number of gaze samples around speech start time i.e., [-1500ms, 1500ms]
fi2	Fixation with second largest number of gaze samples around speech start time i.e., [-1500ms, 1500ms]
fii	Fixation with largest number of gaze samples around speech start time i.e., [-1500ms, 1500ms] excluding *fa1*
fd	Fixation with largest number of gaze samples excluding the features already used in predictive/adaptive models and around speech start time i.e., [-1500ms, 1500ms]
f1	Fixation with largest number of gaze samples in the entire scanpath
f2	Fixation with second largest number of gaze samples in the entire scanpath

Table 2 Fixation Features

Each of the above features has six parameters ***(x, y, t, n, m, v)*** where ***(x, y)*** are the feature's location, ***t*** is the time difference w.r.t. speech onset time, ***n*** is the number gaze samples in the fixation, ***m*** is the mean pupil diameter of the fixation, and ***v*** is the variance of the pupil diameter of the samples in the fixation. Note that out of the fixation features outlined in Table 2, only a few will be selected based on the modeling process.

4.3 Dominant Gaze Features

A dominant gaze feature is defined as the single fixation feature that can detect the target with the highest probability. Each subject takes part in N trials (i.e, scanpaths) in each experiment. Each trial produces several fixations and the fixations are analyzed w.r.t. to the speech onset time using different search areas around the target [Appendix E.7]. The probability of hits in a search region,

P, i.e., target detection probability, is computed as shown below. Note that all model parameters (DT, NP, VD, W – VD and W defined in section 4.6) are optimized in computing the target detection probabilities.

P = S/N

Where

P = probability of detecting the target by a speech and/or gaze criteria

S = number of trials in which a target was detected successfully by a speech and/or gaze criteria

N = total number of trials by a subject

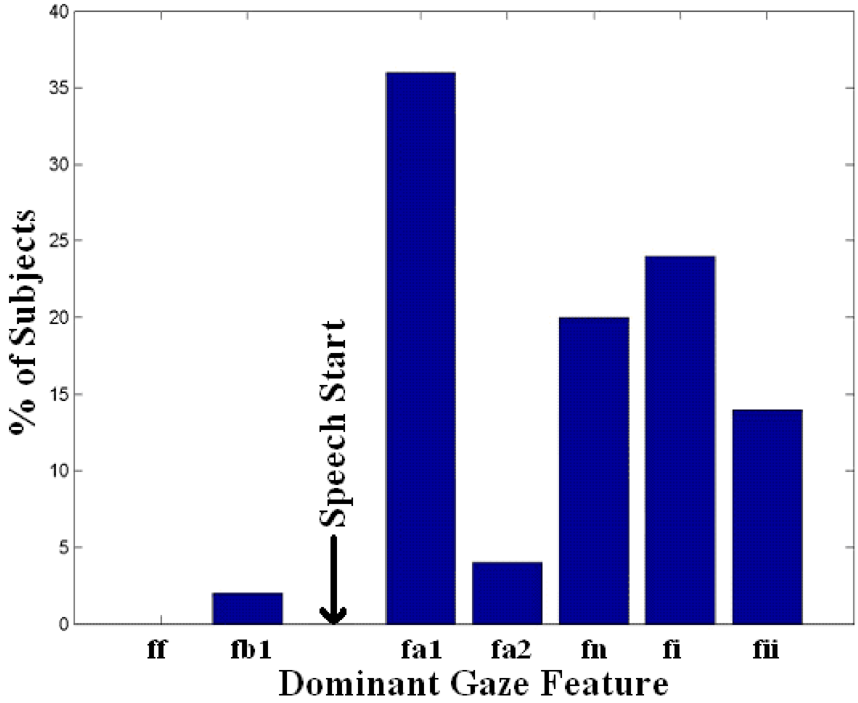


Figure 5 Dominant Gaze Features for all Subjects in One Word Task

Figure 5 shows that the dominant gaze features for all subjects in the simple one word experiment. It illustrates that not all subjects have the same dominant gaze feature in their speech/gaze interaction patterns. ***fa1*** can be seen as the dominant gaze feature for majority of the subjects. The features ***ff, fb1, fa1***, and ***fa2*** are independent features and the features ***fn, fi,*** and ***fii*** are dependent features. Although ***fi*** and ***fn*** seem to be higher than ***fii,*** they are not selected as the second major feature because they are dependent on the ***fa1*** feature (i.e. many of the successful trials included in ***fi*** and ***fn*** are also examples of ***fa1***). Consequently, the two highly effective independent features ***fa1*** and ***fii*** are selected in the data analysis for all subjects in Experiment 1.

Next, Figure 6 illustrates (for 5 subjects) that the dominant gaze feature for a particular user changes over time. Each subject performed 100 consecutive trials, and the dominant feature was found and plotted for each of 10 consecutive subsets of 10 trials. The plots show that, for these 5 subjects, the dominant gaze feature was not constant over time. This result was observed to be consistent with the behavior of many subjects.

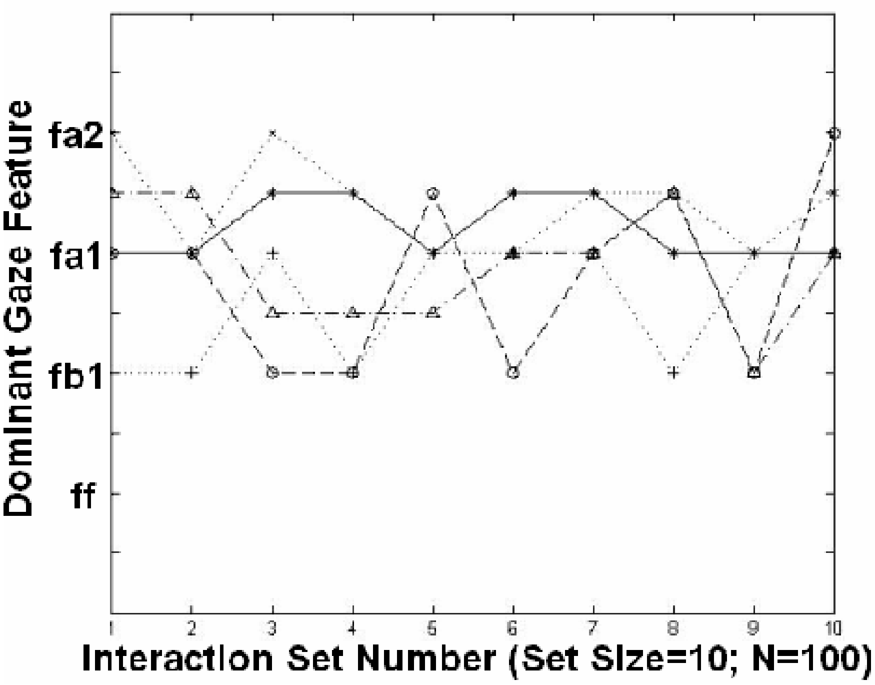


Figure 6 Dominant Gaze Variations Over Time

4.4 Need for Adaptation

Since it has been shown that the dominant gaze features are not constant across users and also not constant for a particular user over time, an adaptive integration model is required. What follows is an analysis of results of experiment 1 using both non-adaptive and adaptive prediction models. A detailed analysis of the data for five subjects was conducted using both linear and non-linear models and a number of different tuning constants. The analysis showed that a linear system performed better than robust fit multi-linear regression model with 10 different weighting functions. Consequently, the linear system model was adopted for further analysis.

4.5 Linear Prediction Model

A linear, time varying system **y = Ax** was constructed using the fixation-based features. Each of these features has six parameters associated with it namely *(x, y, t, n, m, v)* where *(x, y)* is the feature's predicted location of gaze, *t* is the time of the fixation relative to speech onset time (+ or **-**, depending on whether the fixation occurred before or after speech onset time), *n* is the number of gaze samples in the fixation measuring the intensity of the fixation, and *(m, v)* are the mean and variance of the pupil diameter of gaze samples in the fixation. No raw gaze samples-based features were used in constructing the linear system because these features were noisy and didn't produce results comparable to those achieved with fixation-based features. A Fisher's distance measure was applied to measure each feature's effectiveness in classifying the scanpaths, i.e., whether they result in a hit (fall within the target region) or a miss (do not fall within the target region). Principal Component Analysis did not help in reducing the dimensionality of the system significantly to lower the computational complexity.

Based on the analysis of all features, individually, and in various combinations, in Experiment 1, the combination of *fa1/fii*, the first fixation after speech onset time and the fixation with the largest number of gaze samples within the time window [-1500ms, 1500ms] excluding the *fa1*, has proven to be the optimal features for the Experiment 1 prediction model. This feature combination forms a linear system of 6 variables by using the *(x, y, t)* feature parameters of the two selected fixations. The influence of the *(n, m, v)* parameters on the speech/gaze integration model needs further research. The number of gaze samples in the fixation, *n* is not used for each feature because the fixation with the largest number of gaze samples other than *fa1* already accounts for this information. The fixations (e.g., fb1, fa1, fa2, fii) around the speech start time,

S_s are shown in Figure 7. The time window around S_s is [-TM1, TM2]. More fixations, before and after the speech onset time, can be included in the model depending on the task complexity.

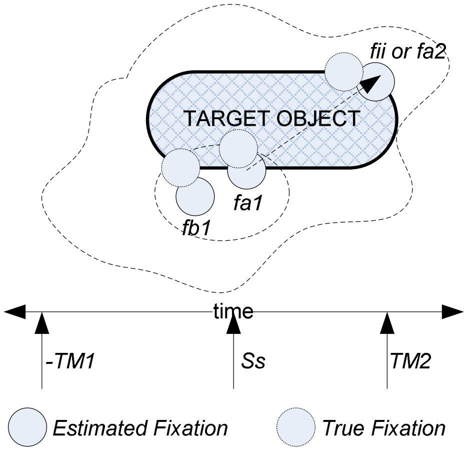


Figure 7 Speech and Gaze Interaction Model, S_s = speech start time, -TM1 and TM2 bound the time window (+/- 1500msec) around S_s.

The time window is not necessarily symmetric around the speech onset time, because the nature of the task being performed can influence the time window limits. Also, these time window limits *TM1* and *TM2* for searching the fixations around the speech start time S_s are random variables. Their relative magnitudes are also a function of task complexity. For more complex tasks involving searching/preprocessing, our experiments show that more time is required before the speech command is uttered and $|TM1| > |TM2|$. Currently the model assumes that the gaze will always be associated with the speech onset time. In this basic model, it is assumed that the two variables *TM1* and *TM2* are constant and model all tasks' fixations to be within these two time limits .For simple tasks *fb1* is not included in the model for computing the predicted gaze location. Only for complex tasks which involve searching is *fb1* included in the model along with *fa1/fii* and any other pertinent features. In addition, the fixations are only an estimation of the *true* fixations because the gaze samples may suffer from equipment errors, calibration errors, *etc.*

Although, *fa1/fii* has proven to be the optimal feature set for Experiment 1 (Figure 5), the models do not strictly depend on any specific feature combination and can be extended to include any feature sets. It only indicates that a particular feature combination is efficient for the task to be used in prediction/adaptation models.

The distribution of fixations around speech onset time varies with the task complexity and hence the choice of fixations to be used in adaptive prediction (the RAP algorithm introduced in Section 4.6) depends significantly on the task complexity. However, it is assumed that a majority of the tasks can be captured with the use of a few fixations around speech onset time. The linear prediction model for gaze is now described, in detail.

Each interaction's gaze attention location is predicted as:

$$x = [fa1_x\, W_{x1x} + fa1_y\, W_{y1x} + fa1_t\, W_{t1x}] + [fii_x\, W_{x2x} + fii_y\, W_{y2x} + fii_t\, W_{t2x}]$$
$$y = [fa1_x\, W_{x1y} + fa1_y\, W_{y1y} + fa1_t\, W_{t1y}] + [fii_x\, W_{x2y} + fii_y\, W_{y2y} + fii_t\, W_{t2y}]$$

Where two features *(fa1, fii)* each having *(x, y, t)* parameters are used

fa1 – first fixation after speech start time

fii – fixation with maximum gaze samples excluding *fa1*

$W_{???}$ – coefficient value

e.g., W_{y1x} is the coefficient of first feature's *y* in predicting *x* of gaze attention location

After the first interaction, the following equation describes the system.

$$(x\ y)_1 = [fa1_x\ fa1_y\ fa1_t\ fii_x\ fii_y\ fii_t]_1 \begin{bmatrix} W_{x1x} & W_{x1y} \\ W_{y1x} & W_{y1y} \\ W_{t1x} & W_{t1y} \\ W_{x2x} & W_{x2y} \\ W_{y2x} & W_{y2y} \\ W_{t2x} & W_{t2y} \end{bmatrix}$$

After collecting M interactions and rewriting the features more generally, the following equation describes the system.

$$\begin{bmatrix} x_1\ y_1 \\ x_2\ y_2 \\ \ldots \\ x_M\ y_M \end{bmatrix} = \begin{bmatrix} f_{11} & f_{12} & f_{13} & f_{14} & f_{15} & f_{16} \\ f_{21} & f_{22} & f_{23} & f_{24} & f_{25} & f_{26} \\ & & & \ldots & & \\ f_{M1} & f_{M2} & f_{M3} & f_{M4} & f_{M5} & f_{M6} \end{bmatrix} \begin{bmatrix} W_{x1x} & W_{x1y} \\ W_{y1x} & W_{y1y} \\ W_{t1x} & W_{t1y} \\ W_{x2x} & W_{x2y} \\ W_{y2x} & W_{y2y} \\ W_{t2x} & W_{t2y} \end{bmatrix}$$

The coefficients W can be obtained by solving for the inverse of the feature matrix and then the M+1th interaction can be estimated using the following equations.

$$x_{M+1} = [f_{M+1,1}\ W_{x1x} + f_{M+1,2}\ W_{y1x} + f_{M+1,3}\ W_{t1x} + f_{M+1,4}\ W_{x2x} + f_{M+1,5}\ W_{y2x} + f_{M+1,6}\ W_{t2x}]$$
$$y_{M+1} = [f_{M+1,1}\ W_{x1y} + f_{M+1,2}\ W_{y1y} + f_{M+1,3}\ W_{t1y} + f_{M+1,4}\ W_{x2y} + f_{M+1,5}\ W_{y2y} + f_{M+1,6}\ W_{t2y}]$$

Each subject produces a sample set of about 100-200 trials/samples/scanpaths in the one word task. To compute each subject's linear prediction model coefficient matrix, half of the sample set is selected randomly to train the model and the other half is tested. This process is repeated 500 times and an average coefficient matrix for each subject is calculated. The results based on the average subject specific coefficient matrix are denoted as *LPh* while the results based on a universal coefficient matrix are denoted as *LPu***.** The universal coefficient matrix for each subject is the average of all other subjects' coefficient matrices excluding that of the subject. The universal coefficient matrix helps as the initial condition for new users of the system. Once the system starts with initial condition *LPu* for that subject, the system can start adapting to the user over the course of user interactions. The universal coefficient matrix is expected to converge to the subject specific coefficient matrix through the adaptation process. Moreover, the adaptive algorithm requires an initial condition to converge to a solution quickly and the *LPu* has proven to be a good initial condition for the adaptation instead of a zero initial condition. A non-zero initial condition also reduces the computational complexity of the adaptation process.

Solving for the inverse may not be feasible due to inconsistencies in the system of equations due to measurement noise. So an error term is added to the above system of equations which then results in the following equation.

$$\begin{bmatrix} x_1\ y_1 \\ x_2\ y_2 \\ \cdots \\ x_M\ y_M \end{bmatrix} = \begin{bmatrix} f_{11} & f_{12} & f_{13} & f_{14} & f_{15} & f_{16} \\ f_{21} & f_{22} & f_{23} & f_{24} & f_{25} & f_{26} \\ & & & \cdots & & \\ f_{M1} & f_{M2} & f_{M3} & f_{M4} & f_{M5} & f_{M6} \end{bmatrix} \begin{bmatrix} W_{x1x} & W_{x1y} \\ W_{y1x} & W_{y1y} \\ W_{t1x} & W_{t1y} \\ W_{x2x} & W_{x2y} \\ W_{y2x} & W_{y2y} \\ W_{t2x} & W_{t2y} \end{bmatrix} + \begin{bmatrix} e_{x1} & e_{y1} \\ e_{x2} & e_{y2} \\ \ldots & \cdots \\ e_{xM} & e_{yM} \end{bmatrix}$$

The above equation can be written in a simpler form as below.

$$\begin{aligned} &\textit{Minimize: } e1^2 + e2^2 + e3^2 \\ &y1 = h_{11}\, x_1 + h_{12}\, x_2 + h_{13}\, x_3 + e1 \\ &y2 = h_{21}\, x_1 + h_{22}\, x_2 + h_{23}\, x_3 + e2 \\ &y3 = h_{31}\, x_1 + h_{32}\, x_2 + h_{33}\, x_3 + e3 \end{aligned}$$

Solving for the least squares solution yields the following set of equations.

$$\textit{minimize} \sum_{i=1}^{M} ei^2$$

$$\textit{subject to } \mathbf{y} = \mathbf{Hx} + \mathbf{e}$$

$$\textit{Least Squares Solution: } \mathbf{x}_{LS} = [\mathbf{H}^T\mathbf{H}]^{-1}\mathbf{H}^T\mathbf{y}$$

$$\textit{For Speech/Gaze Experiments: } \mathbf{w}_{LS} = [\mathbf{F}^T\mathbf{F}]^{-1}\mathbf{F}^T\mathbf{y}$$

So, for Experiment 1, based on the two features *fa1* and *fii* the following linear system is constructed where (x_{ij}, y_{ij}, t_{ij}) are the parameters of the feature *i* in the design matrix, ***x(t)*** is the coefficient matrix, ***H(t)*** is the design matrix of features, and ***y(t)*** is output matrix.

$$\begin{array}{ccccc} \begin{bmatrix} x_1, y_1 \\ x_2, y_2 \\ \ldots \\ x_n, y_n \end{bmatrix} & = & \begin{bmatrix} x_{11}, y_{11}, t_{11}, x_{12}, y_{12}, t_{12} \\ x_{21}, y_{21}, t_{21}, x_{22}, y_{22}, t_{22} \\ \ldots \\ x_{n1}, y_{n1}, t_{n1}, x_{n2}, y_{n2}, t_{n2} \end{bmatrix} & & \begin{bmatrix} a_{11}, a_{12} \\ a_{21}, a_{22} \\ a_{31}, a_{32} \\ a_{41}, a_{42} \\ a_{51}, a_{52} \\ a_{61}, a_{62} \end{bmatrix} \\ \boldsymbol{y(t)} & = & \boldsymbol{H(t)} & * & \boldsymbol{x(t)} \end{array}$$

4.6 Adaptive Prediction Model

The linear prediction model described above may not be computationally feasible so Row Action Projection (RAP), a sample based iterative technique, is chosen to solve the system of equations. In this section the RAP technique will be covered before delving into the adaptive speech/gaze integration model. Refer to [61] for a thorough treatment of the RAP and associated techniques.

All physical processes are continuous time systems and the majority of these are non-linear in nature. However, to keep the modeling process simpler, almost all systems start off with modeling the physical process as a linear system. It may be a time invariant or time varying system depending on the nature of the underlying physical process governing the system of equations. Consider the system of linear equations $\boldsymbol{y} = \boldsymbol{Hx}$ where $\boldsymbol{H}$ is the design matrix of the system, $\boldsymbol{x}$ is the observed input entering the system, and $\boldsymbol{y}$ is the estimated output. The standard least squares solution can be written as $\boldsymbol{x_{LS} = [H^T H]^{-1} H^T y = H_{LS} y}$ where $\boldsymbol{H_{LS} = [H^T H]^{-1} H^T}$. The linear system representation may sometimes yield a rank deficient matrix giving ill conditioned system matrix $\boldsymbol{H}$**.** Then the solution $\boldsymbol{H_{LS}}$ is not possible to evaluate directly. The design matrix needs to be expressed in diagonal form using singular value decomposition as $\boldsymbol{H = U\, D\, V^{T}}$ for finding a pseudo-inverse solution.

The techniques described thus far are good for systems of deterministic variables. If the physical process is a random process (*e.g.*, speech/gaze integration model), then a statistical least squares solution needs to be applied in place of the standard least squares solution. The following set of equations describes the statistical least squares method.

$\boldsymbol{y = Hx + e}$ minimizing $E[\sum \boldsymbol{e}^2]$

whose solution is given as $\boldsymbol{x_{ALS}} = E\boldsymbol{[H^T H]^{-1}}\, E\boldsymbol{[H^T y]}$

The standard linear system discussed thus far requires a block of data to train the linear system before estimating the next sample (*i.e.*, scanpath's predicted gaze location). A real-time implementation of the system will not always have a block of data available. Even after collecting a few samples when the block of data is available, the system doesn't always need all the data to predict the next sample (*i.e.*, scanpath). In some cases, such block processing may not be computationally feasible at times. Thus, a single sample of data needs to be processed one at a time for any real-time and adaptive data processing. Among several methods that exist for sample based processing, row based data processing is better because each row of the design matrix corresponds to one data sample. Row action methods (e.g., ART algebraic reconstruction technique) are suitable for such sample-based processing and are preferred because of their ability to work for rank deficient or ill conditioned systems. Any continuous time system needs to be represented as a discrete time system for sample based processing. Consider a discrete time system representation as shown below (Note that all the description of RAP in this Section can be found in Computational Methods of Signal Recovery by Mammone [61]).

$$y_1 = h_{11}x_1 + h_{12}x_2 + h_{13}x_3$$

$$y_2 = h_{21}x_1 + h_{22}x_2 + h_{23}x_3$$

$$\dots$$

$$y_M = h_{M1}x_1 + h_{M2}x_2 + h_{M3}x_3$$

The above equations can be generalized using $y_i = <h_i . x>$ and each equation is a hyperplane in N dimensional space. Here there are M equations in a 3 dimensional space. Normally the value of M is always greater than N to ensure that the system is not ill conditioned. Even if M<N (the number of equations is smaller than the number of variable), the RAP technique can still converge to a solution bounded by M hyperplanes, as described below. These M hyperplanes given by M equations form a convex set in the hyperspace. The RAP technique starts off assuming a solution or initial condition and iterates over the set of hyperplanes by repeatedly projecting onto them. Given an initial solution of $\boldsymbol{x_0}$, project it onto the first hyperplane given by the first equation. This basically means that a new vector $\boldsymbol{x_1}$ will be calculated by moving a distance of *d* from $\boldsymbol{x_0}$. The direction of movement is given by the *unit normal* vector of the hyperplane being projected. Similarly, projection onto a hyperplane $\boldsymbol{h_k}$ comes from a point $\boldsymbol{x_{k-1}}$

in hyperspace.

So at any stage k,

$$\boldsymbol{x_k} = \boldsymbol{x_{k-1}} + d\,(\boldsymbol{h_k} / |\boldsymbol{h_k}|)$$

where $d = (\boldsymbol{x_k} - \boldsymbol{x_{k-1}}) \,.\, (\boldsymbol{h_k} / |\boldsymbol{h_k}|)$ after projecting $\boldsymbol{x_k} - \boldsymbol{x_{k-1}}$ along unit normal of $\boldsymbol{h_k}$

$$= (\boldsymbol{x_k\, h_k} - \boldsymbol{x_{k-1} h_k}) \;/\; |\boldsymbol{h_k}|$$

$$= (\boldsymbol{y_k} - \boldsymbol{x_{k-1} h_k}) / |\boldsymbol{h_k}|$$

$= \boldsymbol{e_k} / |\boldsymbol{h_k}|$ after generalizing the $\boldsymbol{x_{k-1} h_k}$ term

Then, $\boldsymbol{x_k} = \boldsymbol{x_{k-1}} + (\boldsymbol{e_k} / |\boldsymbol{h_k}|) \,.(\boldsymbol{h_k} / |\boldsymbol{h_k}|)$

$$= \boldsymbol{x_{k-1}} + (\boldsymbol{e_k} / |\boldsymbol{h_k}|^2)\, \boldsymbol{h_k}$$

Adding a convergence factor into the equation the main RAP equation for iterative hyperplane projection becomes $\boldsymbol{x_k} = \boldsymbol{x_{k-1}} + \boldsymbol{\mu}\, (\boldsymbol{e_k} / |\boldsymbol{h_k}|^2)\, \boldsymbol{h_k}$. Figure 8 illustrates how an initial solution $\boldsymbol{x_0}$ will converge to the solution bounded by the hyperplanes in the hyperspace defined by N dimensions. The value of $\boldsymbol{\mu}$ will determine how fast the solution will converged. Very small values of $\boldsymbol{\mu}$ will take a long time to converge to a solution but it will give a more accurate solution. However, a large value may or may not converge at times to a solution because it might be skipping the convex set altogether and may be oscillating around the solution. In order to ensure the solution convergence, the RAP needs to be terminated after either the solution converges to within certain error or a maximum number of iterations are reached in trying to converge to a solution.

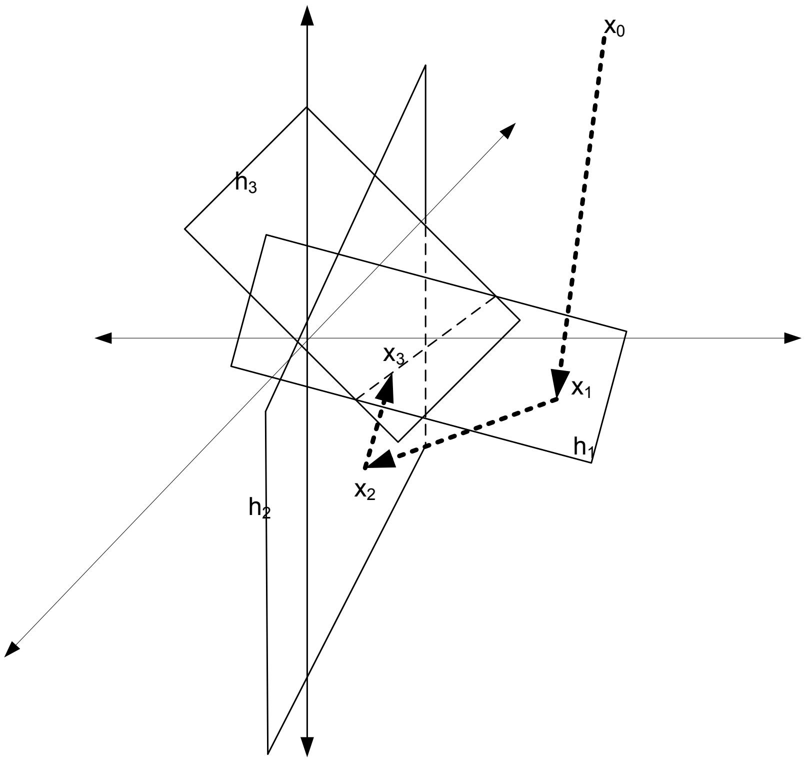
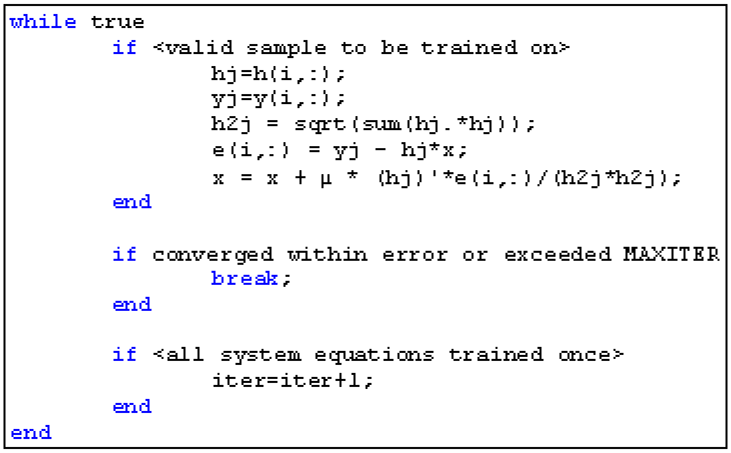


Figure 8 RAP Algorithm – An x_0 converging to a solution x_3 after 3 projections (1 iteration)

In order to build a real-time and/or an adaptive algorithm, single samples *i.e.*, scanpaths need to be processed, one at a time, with *a priori* information from the last few samples *i.e.*, scanpaths/interactions. The RAP technique is a sample-based technique that doesn't require a block of data. If a block of data, *i.e.*, the past few interactions, is available, the algorithm converges more rapidly, to the solution space for the sample being estimated. Like any other adaptation techniques, the underlying process is assumed to be a slow non-stationary one. In order to improve the prediction of the next sample/scanpath, a window of samples from the past is used to compute the adaptation coefficient values. The current interaction may not always depend on too many past interaction samples as the user's behavior is determined by several factors like training and fatigue. The moving window tracks the user's behavior while providing adaptation. The initial condition required for computing the first scanpath's features comes from the *LPu* coefficients for the subject.



```
while true
        if <valid sample to be trained on>
               hj=h(i,:);
               yj=y(i,:);
               h2j = sqrt(sum(hj.*hj));
               e(i,:) = yj - hj*x;
               x = x + μ * (hj)'*e(i,:)/(h2j*h2j);
        end

        if converged within error or exceeded MAXITER
               break;
        end

        if <all system equations trained once>
               iter=iter+1;
        end
end
```

Figure 9 Pseudo code of RAP Algorithm

Figure 9 illustrates the pseudo code for the RAP algorithm in the adaptation model. The RAP technique uses the basic concepts of a linear system, but it is nonlinear in nature to converge to the solution for a system of equations. Consider a linear system $\boldsymbol{y} = \boldsymbol{Hx}$ where $\boldsymbol{y}$ is the output, $\boldsymbol{x}$ is the coefficient matrix of the linear system, and $\boldsymbol{H}$ is the design matrix constituted by the feature space. For each scanpath, the RAP algorithm looks back in time for a window of samples *W* and estimates the current sample. In the window of *W* samples, whether a sample is included in the RAP algorithm or not depends on its proximity to the target object's center to the predicted location. It is determined by *valid distance, VD*. With large *VD* values, noisy scanpaths would also be included in the determination of the adaptive coefficients potentially yielding estimation errors for the next scanpath. If the *VD* values are too small then it would throw away too many scanpaths giving less data for the model parameters to converge. This in turn would yield large estimation errors for the next scanpath. Thus, it is very critical to choose appropriate values for *W/VD*. Given appropriate *W/VD*, the *Dispersion Threshold DT*, and *Number of Points NP*, the RAP algorithm iterates over the *W* samples to calculate the adaptation coefficients of the linear system for estimating the next scanpath. With 2000 iterations and step size $\mu = 0.015$ and *LPu* as the initial condition, the RAP algorithm converges to next scanpath's estimated location with reasonable accuracy for a majority of subjects. The RAP equation can be written as the following equation:

$$W^{i+1} = w^{i} + \lambda_{k} \frac{y_{k} - \langle F_{k}, w^{i} \rangle}{|F_{k}|^{2}} F_{k}$$

Where

$Fk = \langle fa1\mathrm{x}, fa1\mathrm{y}, fa1\mathrm{t}, fii\mathrm{x}, fii\mathrm{y}, fii\mathrm{t} \rangle$ (Expt 1)

$\quad = \langle fb1\mathrm{x}, fb1\mathrm{y}, fb1\mathrm{t}, fa1\mathrm{x}, fa1\mathrm{y}, fa1\mathrm{t}, fa2\mathrm{x}, fa2\mathrm{y}, fa2\mathrm{t} \rangle$ (Expt 2)

λ_{k} = *relaxation parameter*

4.7 Comparison of Linear and Adaptive Prediction

Figure 10 shows the comparison of target detection probability for linear and adaptive prediction for all subjects. The difference between the linear and adaptive prediction is more clearly depicted in the following section (4.8), where the interface performance comparisons are presented.

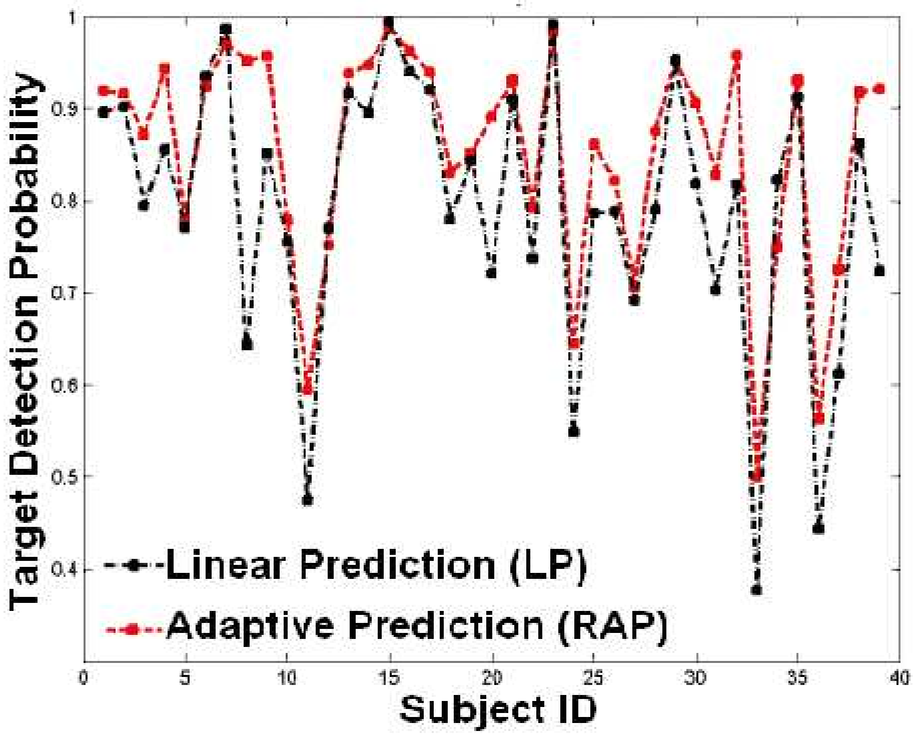


Figure 10 Comparison of Linear and Adaptive Prediction

Figure 11 shows the L2 distance of predicted gaze location from the center of the target for linear and adaptive prediction models for a subject. It can be seen that the adaptive prediction model's error is lower than that of the linear prediction model. The error can be further minimized through the use of projection operators onto convex sets. Using these operators, RAP guarantees the convergence to a solution of the system of equations and reduces the prediction error. The modified RAP equation using the convex sets is shown below. Several constraints like positivity, band limiting, time limiting, etc., can be used. But only the positivity constraint is employed in the Experiment 1 analysis.

$$w^{j+1} = P_c\{w^j\} + \lambda_k \frac{y_k - \langle F_k, P_c\{w^j\}\rangle}{|F_k|^2} F_k$$

Where

$P_c = P_{c1}\, P_{c2} \ldots P_{cr}$ defines the convex sets defined by constraints

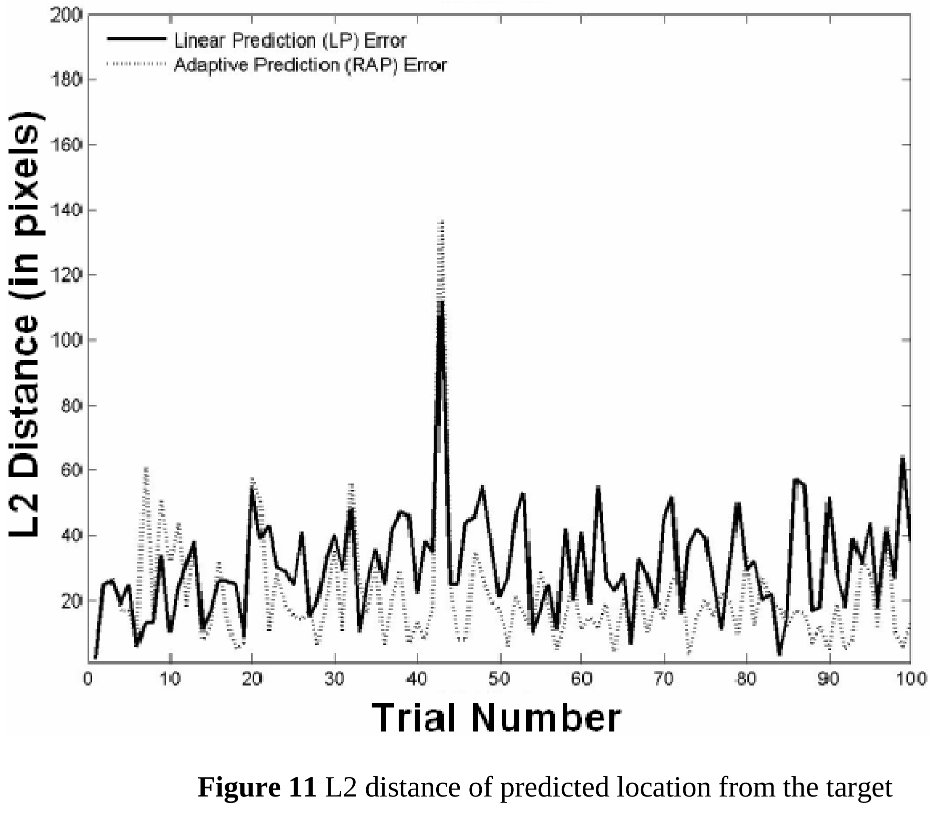
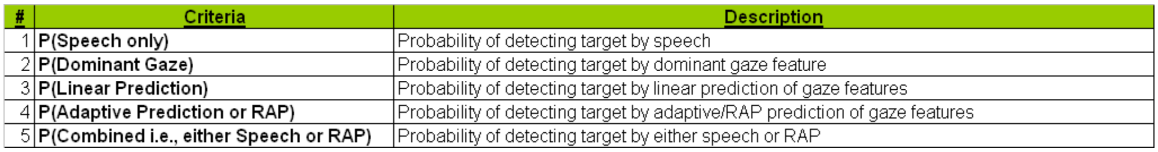


Figure 11 L2 distance of predicted location from the target

4.8 Performance Comparisons of Speech, Gaze, Adaptive and Non-Adaptive Approaches to Target Detection

4.8.1 Target Detection Approaches Compared

Five different methods of calculating target detection probabilities, shown in Table 3, are compared to evaluate the performance of the interface.



#	Criteria	Description
1	**P(Speech only)**	Probability of detecting target by speech
2	**P(Dominant Gaze)**	Probability of detecting target by dominant gaze feature
3	**P(Linear Prediction)**	Probability of detecting target by linear prediction of gaze features
4	**P(Adaptive Prediction or RAP)**	Probability of detecting target by adaptive/RAP prediction of gaze features
5	**P(Combined i.e., either Speech or RAP)**	Probability of detecting target by either speech or RAP

Table 3 Different Target Detection Probabilities

Figure 12 shows the target detection probabilities for all subjects, calculated by the five criteria listed in Table 3. It should be noted that “Dominant Gaze”, the result for the best single feature found for each subject is presented as a standard against which the other gaze approaches are to be measured, but does not represent a practical system implementation (the parameters *DT*, *NP*, *W*, and *VD* are optimized for each subject giving *maximum RAP probability*). Basically the optimal algorithmic parameters that can maximize the RAP probability are extracted for each subject and the maximum RAP probabilities are plotted. This is the best the RAP algorithm can do given the optimal parameters for the algorithm. It can be seen that the combined speech/gaze system performs better than a speech-only and gaze-only system.

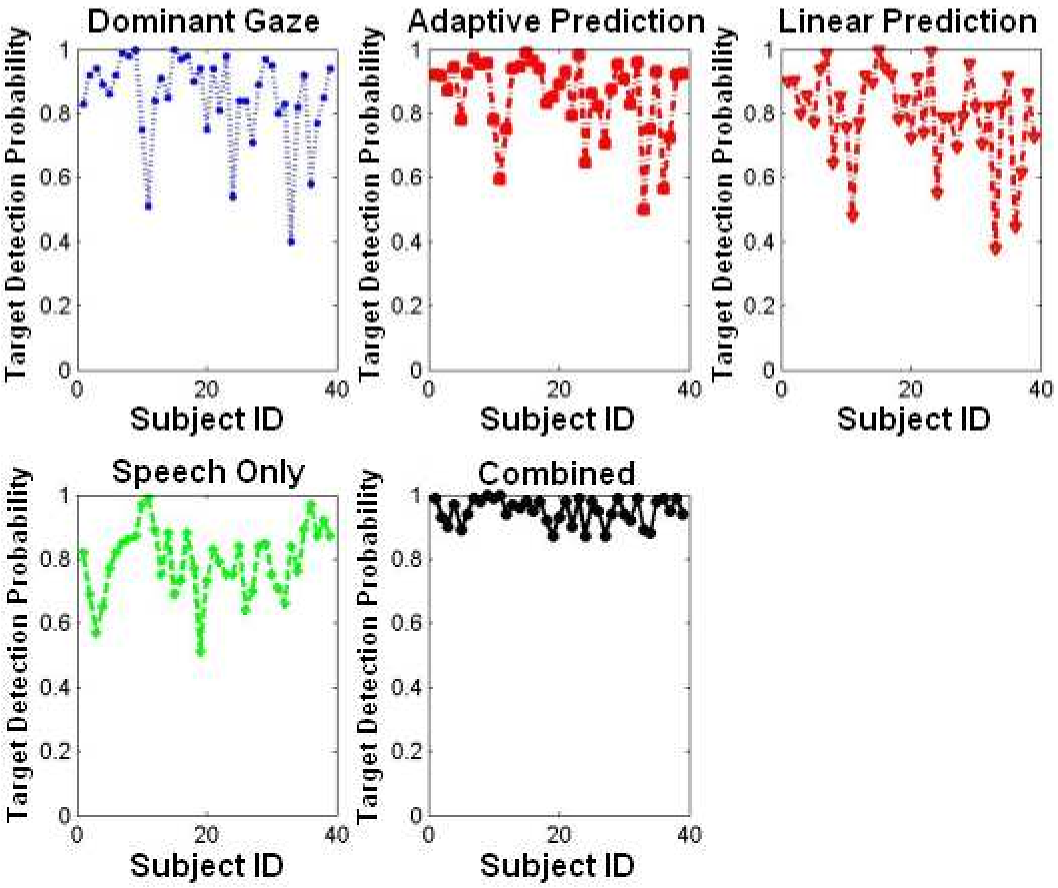
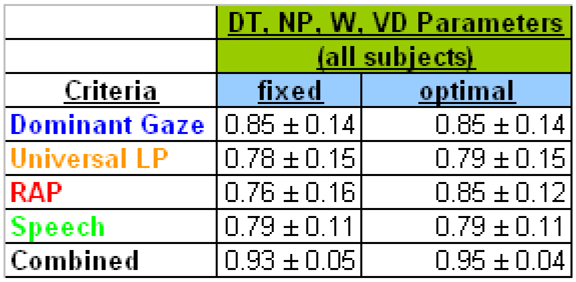


Figure 12 Target Detection Probabilities in One Word Task

4.8.2 Statistical Analysis: Paired Samples t-Tests

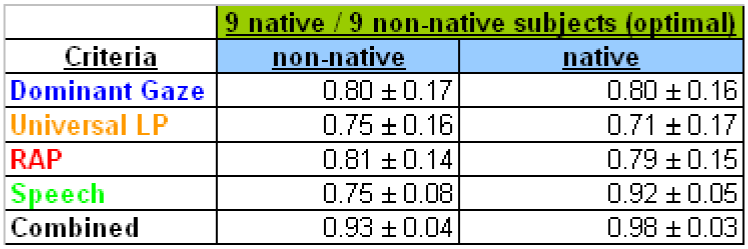
There are different kinds of statistical tests one can perform on the data to extrapolate the results observed with a finite number of subjects to deduce a general conclusion. The statistical test of choice depends on the nature of the variables governing the physical process. The two types of statistical tests that can be employed differ in the assumption of whether the underlying variables follow a normal distribution or not. *Parametric* tests assume that the underlying variables follow a normal distribution while the *non-parametric* tests do not make any assumption about the distribution of the underlying variables.

Different methods of calculating target detection probabilities (i.e., *Dominant Gaze*, *LPu*, *RAP*, *Speech*, and *Combined*) are compared using the t-Tests. The *sample* values for each of these detection criterions are *target detection probability* values. These probability values are obtained after processing the fundamental variable values using different processing modules (e.g., *Speech Recognizer*, *LPu*, and *RAP*). Although these probability numbers are not direct variable measurements, these probability numbers are demonstrated to be normally distributed using Q-Q plots. Table 4 shows the mean and standard deviation of target detection probabilities of all 39 subjects in Experiment 1 and Table 5 shows the mean and standard deviation of target detection probabilities for 9 optimally performing native and non-native subjects. There are two columns of results shown in these two tables, i.e., "fixed" and "optimal". The "fixed" column gives the results when the parameters DT, NP, W, and VD are the same for all subjects. The "optimum" gives the results when the parameters are optimized for each subject (note that the window parameter "W" is only in use when RAP is involved, i.e. in "RAP" and "combined"). From these two tables, it can be inferred that that the adaptive model (*i.e.*, RAP) works better than the non adaptive model (*i.e.*, Universal LP) when the DT, NP, W, and VD parameters are optimized. It is worth noting that, even in the "fixed" case, "Combined" still works significantly better than "Speech", supporting the main contention of this dissertation that speech and gaze working together outperform speech alone. However, the strength of the relationship is not clear from the means/standard deviations alone.



	DT, NP, W, VD Parameters	
	(all subjects)	
Criteria	fixed	optimal
Dominant Gaze	0.85 ± 0.14	0.85 ± 0.14
Universal LP	0.78 ± 0.15	0.79 ± 0.15
RAP	0.76 ± 0.16	0.85 ± 0.12
Speech	0.79 ± 0.11	0.79 ± 0.11
Combined	0.93 ± 0.05	0.95 ± 0.04

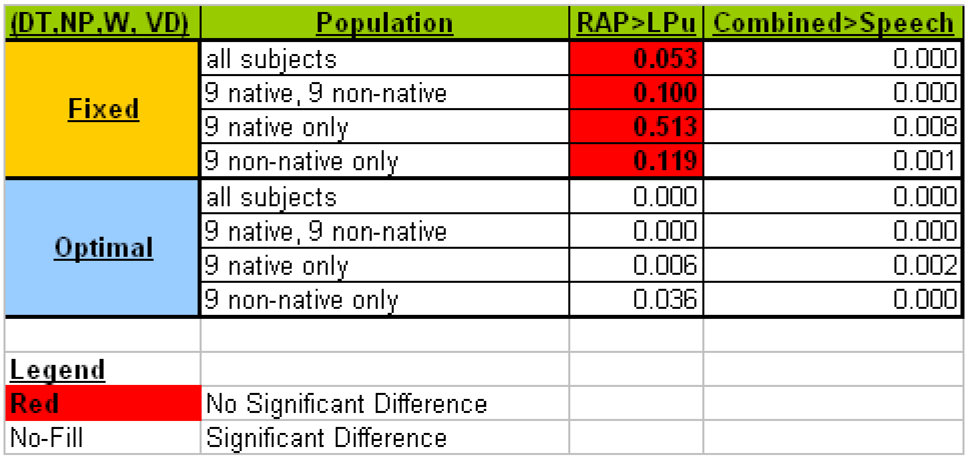
Table 4 μ / σ of Target Detection Probabilities in Experiment 1 (all subjects)



	9 native / 9 non-native subjects (optimal)	
Criteria	non-native	native
Dominant Gaze	0.80 ± 0.17	0.80 ± 0.16
Universal LP	0.75 ± 0.16	0.71 ± 0.17
RAP	0.81 ± 0.14	0.79 ± 0.15
Speech	0.75 ± 0.08	0.92 ± 0.05
Combined	0.93 ± 0.04	0.98 ± 0.03

Table 5 μ ⁄ σ of Target Detection Probabilities in Experiment 1 (only native subjects)

To investigate the significance of the results, the paired samples t-Test is used and the *p*-values are shown in Table 6 for all population categories. The values indicate the statistical significance of the samples' comparisons. A value less than 0.05, indicates that there is a significant difference between the two samples being compared. From the Table 6, it can be seen that the RAP adaptation is significantly better than non-adaptive universal linear prediction when the parameters DT, NP, W, and VD are optimized for each subject. The combined speech/gaze system is significantly better than a speech-only system regardless of the population category and whether the parameters are optimized or not. All values in the table are rounded to 3 digits precision. A value of 0.000 implies that the *p*-value is less than 0.0005.



(DT,NP,W, VD)	Population	RAP>LPu	Combined>Speech
Fixed	all subjects	0.053	0.000
	9 native, 9 non-native	0.100	0.000
	9 native only	0.513	0.008
	9 non-native only	0.119	0.001
Optimal	all subjects	0.000	0.000
	9 native, 9 non-native	0.000	0.000
	9 native only	0.006	0.002
	9 non-native only	0.036	0.000
Legend			
Red	No Significant Difference		
No-Fill	Significant Difference		

Table 6 *p*-values of paired samples t-Test in Experiment 1

Summarizing the most significant t-Test results for Optimized Adaptive vs. Non-adaptive, Combined Speech/Optimized Adaptive vs. Speech alone, Combined Speech/Non-Optimized Adaptive vs. Speech alone

All Subjects

- Paired Samples t-Test using target detection probability
  - Null Hypothesis H0: $\mu 1 = \mu 2$
    - If $t(df) > t_{critical}$ reject H0 where *df* is degrees of freedom
  - Optimized Adaptive (RAP) vs. Non-Adaptive (Linear Prediction)
    - Adaptive model ($\mu = 0.85$, $\sigma = 0.12$) **performed significantly better** than non-adaptive model [$\mu = 0.79$, $\sigma = 0.15$, $t(38) = 5.284$, $p < 0.0005$]
  - Combined Speech/Optimized Adaptive (RAP) vs. Speech alone
    - Combined ($\mu = 0.95$, $\sigma = 0.04$) **performed significantly better** than Speech alone [$\mu = 0.79$, $\sigma = 0.11$, $t(38) = 10.525$, $p < 0.0005$]
  - Combined Speech/Non-Optimized Adaptive (RAP) vs. Speech alone
    - Combined ($\mu = 0.93$, $\sigma = 0.05$) **performed significantly better** than Speech alone [$\mu = 0.79$, $\sigma = 0.11$, $t(38) = 9.636$, $p < 0.0005$]

Native Subjects

- Paired Samples t-Test using target detection probability
  - Optimized Adaptive (RAP) vs. Non-Adaptive (Linear Prediction)
    - Adaptive model ($\mu = 0.79$, $\sigma = 0.15$) **performed significantly better** than non-adaptive model [$\mu = 0.71$, $\sigma = 0.17$, $t(8) = 3.667$, $p < 0.006$]
  - Combined Speech/Optimized Adaptive (RAP) vs. Speech alone
    - Combined ($\mu = 0.98$, $\sigma = 0.03$) **performed significantly better** than Speech alone [$\mu = 0.92$, $\sigma = 0.05$, $t(8) = 4.609$, $p < 0.002$]
  - Combined Speech/Non-Optimized Adaptive (RAP) vs. Speech alone
    - Combined ($\mu = 0.96$, $\sigma = 0.03$) **performed significantly better** than Speech alone [$\mu = 0.92$, $\sigma = 0.05$, $t(8) = 3.496$, $p < 0.008$]

It should be pointed out that, for this analysis, the optimization of the internal parameters was carried out using a semi-automated process. This process used an algorithm that computed results for a range of values of DT, NP, W, and VD (W = 5, 10, 20, 30; VD = 100, 150, 200; DT = 20, 30, 40, 50; NP = 8; for a total of 4 x 3 x 4 x 1 = 48 cases) for all subjects for all cases and automatically picked the optimal case (i.e., maximizing the RAP) for each subject. For a real-time application, the parameters need to be updated for each scanpath/interaction for each user. Initial experiments and analyses for real-time adaptation of these parameters are under way and preliminary results are being examined. The integration of the real-time optimization will be an important aspect of future work.

4.8.3 Interface Usability Performance

Any human/machine interface with one or more modalities is normally effective in a well controlled environment. The interface effectiveness is not a well defined term and needs a more rigorous definition to evaluate the performance of human machine interactions. Moreover, the interface evaluation shouldn’t depend on the modalities or any other factor in the interface. It should only reflect how well the system/interface was able to understand the user’s intent. In this dissertation, *Interface performance* is defined to evaluate the performance of an interface during a set of interactions in a session and also to evaluate the reliability with which the results can be reproducible by the system/interface.

An interaction is considered as a task within a session (or sitting). In a session, a user can issue N number of commands/interactions to the system/interface using one or more modalities. *Target detection probability* measures the number of successful commands/interactions out of the total number of commands issued. A single session provides the detection probability for that session alone. There is no guarantee that the same performance can be delivered each time the user uses the system. So the interface performance needs to be evaluated over a series of sessions. Thus, *Interface performance* denotes the effectiveness in terms of the percentage of the times the target detection probability is higher than an acceptable target detection probability. All target detection probabilities can be evaluated independently to evaluate the modalities independently. The calculation of interface performance is defined as follows.

N – Total number of interactions in a session/sitting

Ns – Total number of interactions where the speech and/or gaze successfully recognized the target

P – Target detection probability = **Ns** / **N** for a particular criteria e.g.., Speech Only, Dominant Gaze, Linear prediction, Adaptive Prediction (RAP), Combined Speech and Adaptive Prediction

S – Total number of sessions for all users who used the system. A session corresponds to a single set of closely spaced (in time) interactions. Note that Experiment 1 has only one session/subject, and 39 total sessions for 39 subjects (whereas Experiment 2, described in the next chapter has 26 subjects, 8 sessions/subject and a total of 26x8 = 208 total sessions for 26 subjects). Sessions can be separated by long time spans anywhere between a few minutes to days. There is no specific minimum or maximum time span limit between two sessions.

Sp – Number of sessions whose probability P is higher than a given probability p

Interface Performance (IP) – For a given P, **IP = Sp / S**

Figure 13 shows the interface performance in the constraint-free interaction task, where **S** = **39** and if, e.g., for a 60% speech recognition rate (*i.e.*, session probability), the interface performance is about 90% when speech is the only modality in the system. That means that for **P= 0.60,** about **35 out 39** subjects (i.e. **Sp = 35**) had a correct recognition rate of 60%. In other words, 9 out of 10 times that the system; is used, the user can expect to see a 60% recognition rates with respect to speech alone.

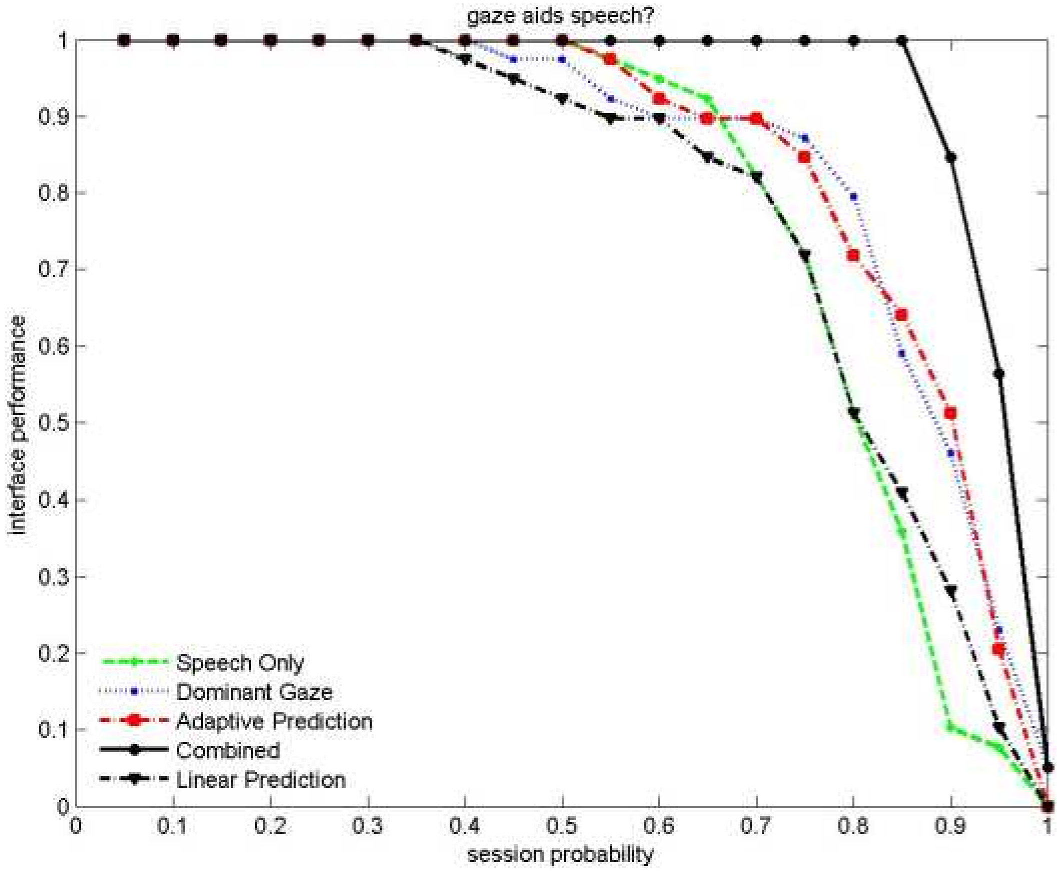


Figure 13 Interface Performance for all 39 subjects in One Word Task

Next, the target detection probability for native and non-native speakers is compared to evaluate any differences in the speech/gaze interface performance for native and non-native speakers. Although both the experiments in this dissertation are not expected to show any bias towards native speakers *with respect to speech/gaze interaction,* it is not entirely clear if a speech/gaze interface has any bias towards the native speakers. It is reasonable to expect that the speech recognition performance is higher for native speakers and with a better recognizer it will be better for all users. But speech recognition performance is not high all the time for any user regardless of the speech recognizer performance. There may be several factors like ambient noise, improper pronunciation, and different accents which can potentially influence the speech recognition performance.

Out of 39 subjects in Experiment 1, 9 are native speakers and 30 are non native speakers. Figure 14 shows results for an equal number of native and non-native speakers selected such that all the subjects are closely spaced in time. When the optimal variables are chosen to extract the maximum RAP performance for each subject, Figure 14 shows that the gap between RAP and speech narrows down. It indicates that RAP is on par with Speech performance with optimal parameters for each subject. The combined speech/gaze performance is higher than either modality acting alone.

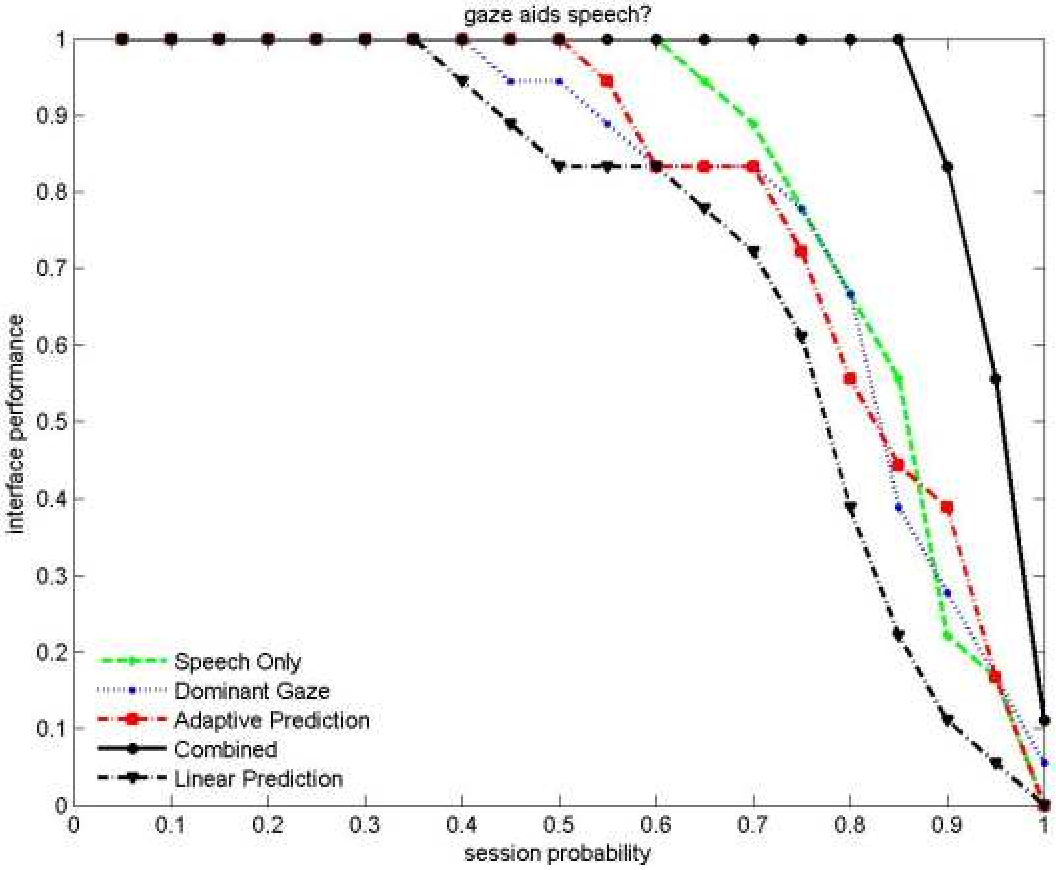


Figure 14 Interface Performance for (9 native and 9 non-native) Speakers in One Word Task

4.9 Summary

In this chapter, a speech/gaze experiment was conducted that collected data on subjects' abilities to identify isolated words on a monitor. The results were analyzed to illustrate the differences in gaze patterns across subjects and across time for a single subject. Linear and adaptive prediction models of a subject's behavior were compared and the results demonstrated that adaptive prediction is necessary and improves the overall accuracy in discerning the user's intent. The improvement with the adaptive prediction approach in speech/gaze-based interactions, for the one-word task, is summarized below.

- Target Detection Probability Results
    - Combined Speech/Adaptive Gaze vs. Speech alone
        - 0.95±0.04 vs. 0.79±0.11 (all subjects)
        - 0.98±0.03 vs. 0.92±0.05 (native subjects)

The result for native subjects (albeit for a small sample) is particularly important because it shows that when the speech recognition percentage is in the low 90's, typical of some of the better speech recognizers under noisy conditions, the combination of speech and gaze raises the percentage into the upper 90's, a region in where many practical systems need to operate.

5. Experiment 2: Menu Task

5.1 Introduction

Speech/gaze based menu systems have been studied to understand menu systems (*e.g.*, the hierarchical menu systems [79]), but not with a focus on an adaptive speech/gaze integration model. This experiment is designed to extract the speech and gaze interaction when the screen contains distractions to gaze. A simple word reading interaction may not fully provide all the design parameters of a predictive/adaptive model for speech/gaze integration. As the task in the human/machine interaction changes the speech/gaze interrelationship may also change. Experiment 2 is designed to understand the impact of task complexity on the speech/gaze interaction and the relevance of speech/gaze interaction in a real world application. So, a menu interaction task is selected in Experiment 2 where the subject interacts with a set of menu items displayed as an array on the screen and the subject speaks a command from the array.

5.2 Hypotheses

The second experiment is identical to the first experiment but the task complexity is slightly increased. It is not expected that the integration algorithm developed in Experiment 1 will perform as well because of the distraction of other elements on the screen. The subject may need to carry out a more involved search causing the scanpath (i.e., a sequence of gaze samples on the screen) to be different, so that arrival time, location of the fixations and speech onset time may be different. It is likely that the key parameters forming the model will be different for this more complex task. The hypotheses to be tested in Experiment 2 include:

- combined speech/gaze systems performs better than a speech-only systems
- increasing spacing improves the gaze prediction
- increasing the font-size will improve the gaze prediction

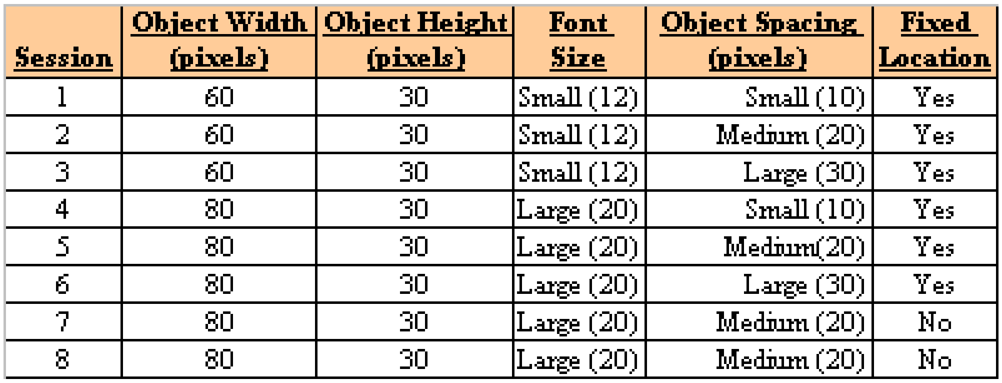
- screen location will affect behavior with less central locations generating less accurate predictions.

The values for the independent variables were chosen to represent reasonably large differences in speech/gaze interaction performance.

5.3 Task Description

The second experiment was designed to represent an individual giving spoken commands in a menu-based system when there is distraction to the gaze from surrounding menu items. Another design criterion for this experiment is to reflect a real-world application more closely than Experiment 1. Each trial in experiment 2 consists of a letter display followed by a 6x6 array of buttons, each one containing a word. The subject is to find the word that starts with the displayed letter and speak the word. By experiment design, a trial is separated from another trial by displaying a letter between the two. This helps in minimizing the correlation between two successive interactions in the experiment.

It is known that optimal letter spacing exists for best reading performance [77] but how object-spacing impacts gaze prediction in speech/gaze interaction is not known. In a typical user interface, command buttons are arranged as a rectangular array of buttons with minimal spacing between them to save real estate space on the screen. It enables the screen designers to present more information to the user and provides a smaller set of workflow management steps (i.e., number of application screens to navigate). When the buttons or objects on the screen are placed very close together, the eye tracker can not accurately identify the user’s desired object. On the other hand, if the object spacing becomes very large, then the amount of information presentable decreases which increases the number of workflow steps in performing a task. Thus, optimizing performance requires a trade off between the two competing requirements of workflow management and accuracy when designing multimodal interfaces that include gaze. The design parameters for Experiment 2 include spacing, font-size, and array location with each session having different values for these parameters. Three spacings 10, 20 and 30 pixel distances (edge-to-edge) are used when displaying the command buttons/icons on the screen for small, medium and large spacings respectively. The spacing of the buttons is identical in the horizontal and vertical directions. Two font sizes 12dpi and 20dpi are used with button sizes 60x30 and 80x30 pixels respectively. In Sessions 1-6, the entire array of 6x6 buttons is centered vertically and horizontally. In session 7, 20 trials are positioned in UL (Upper Left), 10 trials are positioned in the center, and 20 trials are positioned in LR (Lower Right). In session 8, 20 trials are positioned in UR (Upper Right), 10 trials are positioned in the center, and 20 trials are positioned in LL (Lower Left). Sessions 1-8 are summarized in Table 7. Sessions 7 and 8 differ from the other sessions because their goal is to measure the effectiveness of the interface when buttons are in screen corners.



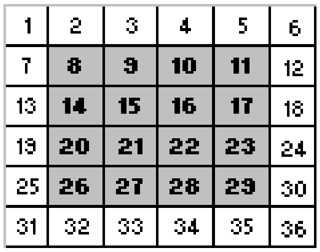
Session	Object Width (pixels)	Object Height (pixels)	Font Size	Object Spacing (pixels)	Fixed Location
1	60	30	Small (12)	Small (10)	Yes
2	60	30	Small (12)	Medium (20)	Yes
3	60	30	Small (12)	Large (30)	Yes
4	80	30	Large (20)	Small (10)	Yes
5	80	30	Large (20)	Medium(20)	Yes
6	80	30	Large (20)	Large (30)	Yes
7	80	30	Large (20)	Medium (20)	No
8	80	30	Large (20)	Medium (20)	No

Table 7 Menu System Experiment Sessions

In each of the 8 sessions of Experiment 2, a subject performs 50 trials of reading menu items from the display of an array of 6x6 menu items. The order of the sequence of sessions is varied across subjects to eliminate *order effects*. In each trial, the subject looks at a letter as shown in Figure 15, which disappears after the subject looks at it.

Figure 15 Menu Item Letter

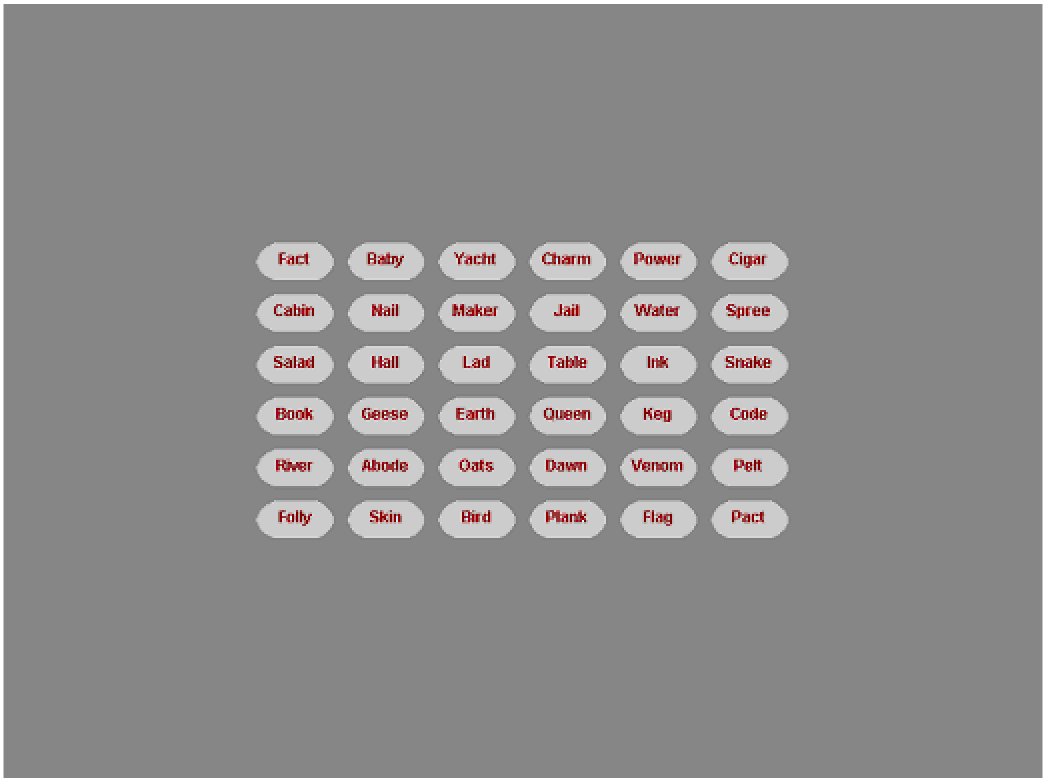
The subject is instructed to look for the word beginning with that letter in the array of buttons and to speak that word. Although the display of menu items is an array of 6x6 buttons, the word the subject speaks always appears in the internal 4x4 array (as in Figure 16), thereby eliminating the boundary conditions by maintaining 8-connectivity around each target word. Eliminating boundary conditions provides uniform treatment of all trials and renders the sample size significant enough to deduce the subject's behavior. The inner 4x4 array is highlighted in Figure 16 only to indicate that this is the area in which the words the subjects are asked to speak are found. In the experiment screens, nothing in the 6x6 array is highlighted in any manner and all menu items appear uniform to the subject.



1	2	3	4	5	6
7	**8**	**9**	**10**	**11**	12
13	**14**	**15**	**16**	**17**	18
19	**20**	**21**	**22**	**23**	24
25	**26**	**27**	**28**	**29**	30
31	32	33	34	35	36

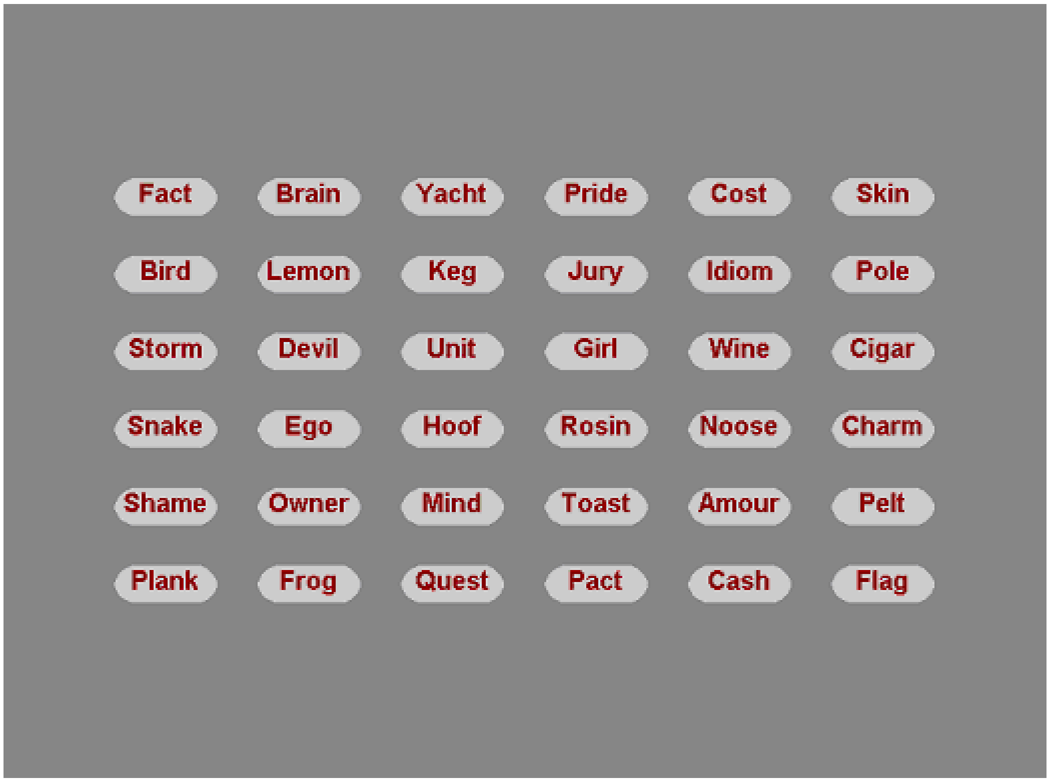
Figure 16 Inner 4x4 Array (the inner array is NOT highlighted in the actual experiment screen)

Each session uses its own set of 36 words to display in the 6x6 array (see Appendix C for word lists of each session) and the word list of a session doesn't change from trial to trail. In each of the 50 trials of any session, the subject reads only 5 out of 36 words/targets repeatedly. The spacing, button width, and font-size do not change within a session for sessions 1-6 (Figure 17 and Figure 18). In sessions 7-8, the location of the word array changes and 3 positions are used for 20, 10, and 20 trials respectively as shown in Figure 19/Figure 20 (session 7) and Figure 21/Figure 22 (session 8). The first letter of all words is capitalized and the words are center justified vertically/horizontally when displayed on the buttons. Some of the target words end with letter “z”. If the word ends with a ‘z’, then the subject is to utter ‘zero’ instead of the actual word displayed on the screen. For example, if the word is displayed as ‘Lakez’ then the subject is to utter ‘zero’ instead of ‘lake’. This ensures that the subject is using gaze when uttering the word.



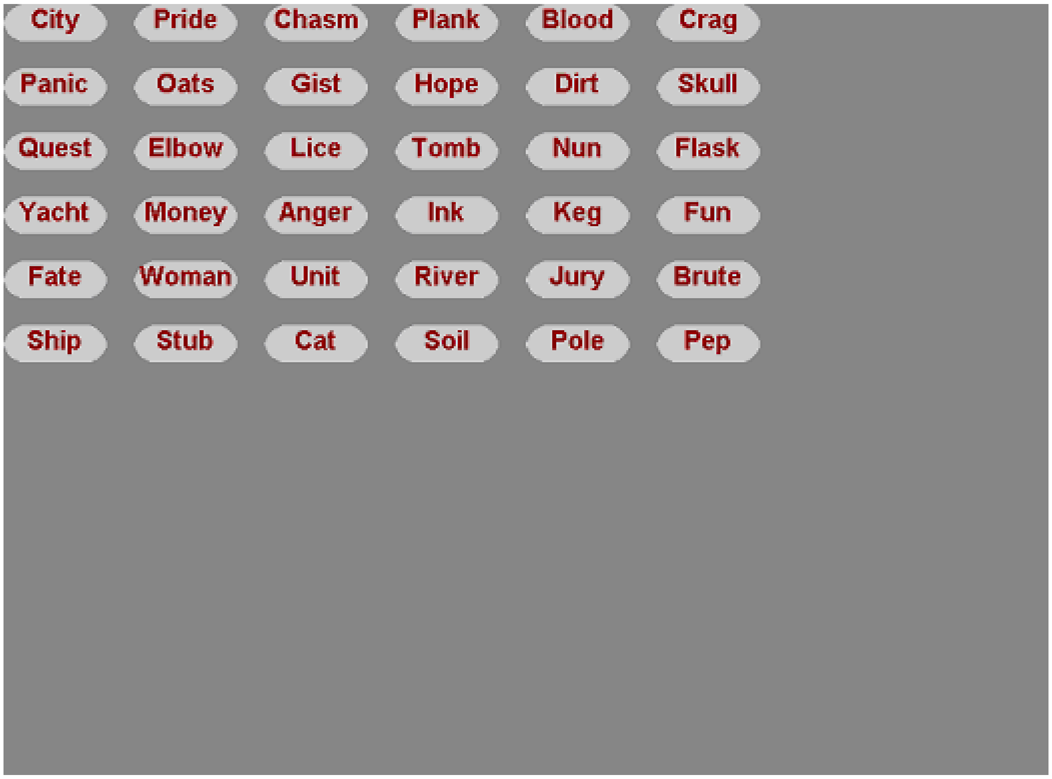
Fact	Baby	Yacht	Charm	Power	Cigar
Cabin	Nail	Maker	Jail	Water	Spree
Salad	Hall	Lad	Table	Ink	Snake
Book	Geese	Earth	Queen	Keg	Code
River	Abode	Oats	Dawn	Venom	Pelt
Folly	Skin	Bird	Plank	Flag	Pact

Figure 17 Small-Font Small-Spacing



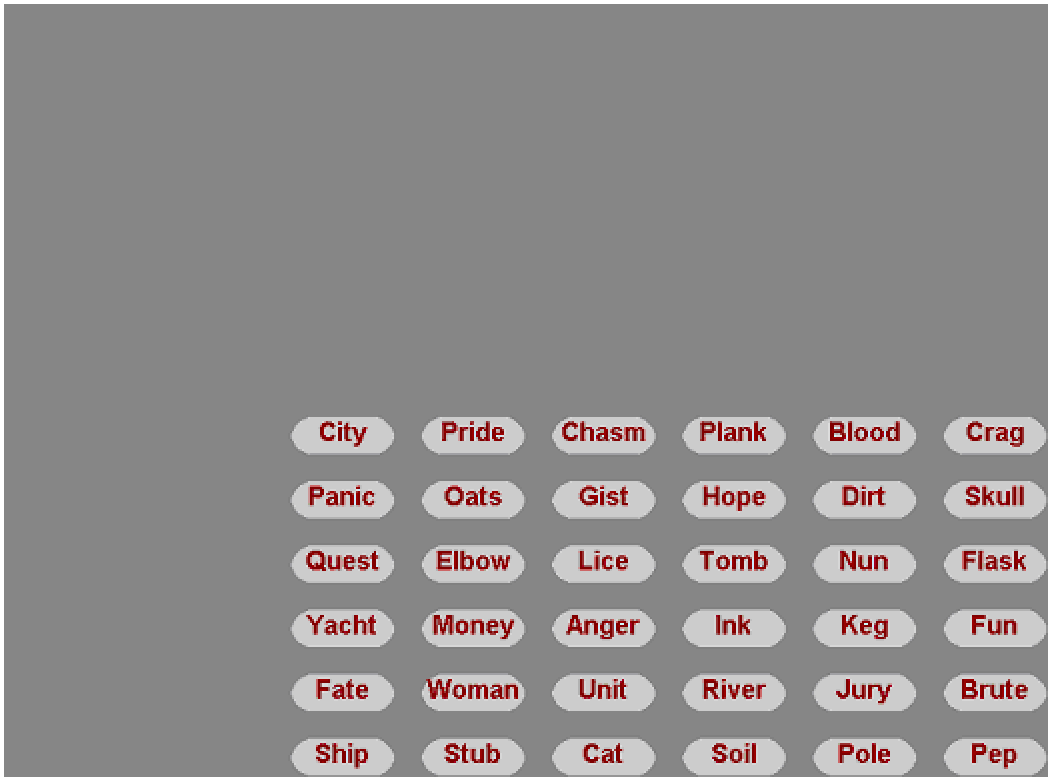
Fact	Brain	Yacht	Pride	Cost	Skin
Bird	Lemon	Keg	Jury	Idiom	Pole
Storm	Devil	Unit	Girl	Wine	Cigar
Snake	Ego	Hoof	Rosin	Noose	Charm
Shame	Owner	Mind	Toast	Amour	Pelt
Plank	Frog	Quest	Pact	Cash	Flag

Figure 18 Large-Font Large-Spacing



City	Pride	Chasm	Plank	Blood	Crag
Panic	Oats	Gist	Hope	Dirt	Skull
Quest	Elbow	Lice	Tomb	Nun	Flask
Yacht	Money	Anger	Ink	Keg	Fun
Fate	Woman	Unit	River	Jury	Brute
Ship	Stub	Cat	Soil	Pole	Pep

Figure 19 Upper-Left Screen



City	Pride	Chasm	Plank	Blood	Crag
Panic	Oats	Gist	Hope	Dirt	Skull
Quest	Elbow	Lice	Tomb	Nun	Flask
Yacht	Money	Anger	Ink	Keg	Fun
Fate	Woman	Unit	River	Jury	Brute
Ship	Stub	Cat	Soil	Pole	Pep

Figure 20 Lower-Right Screen

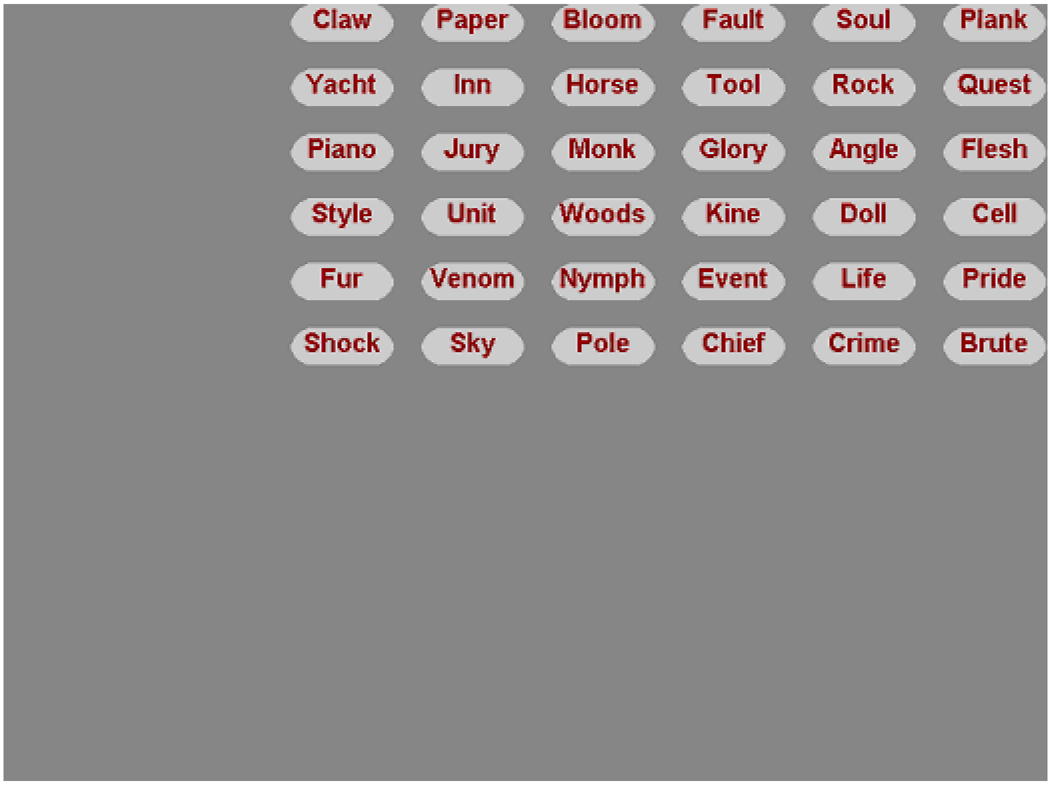


Figure 21 Upper-Right Screen

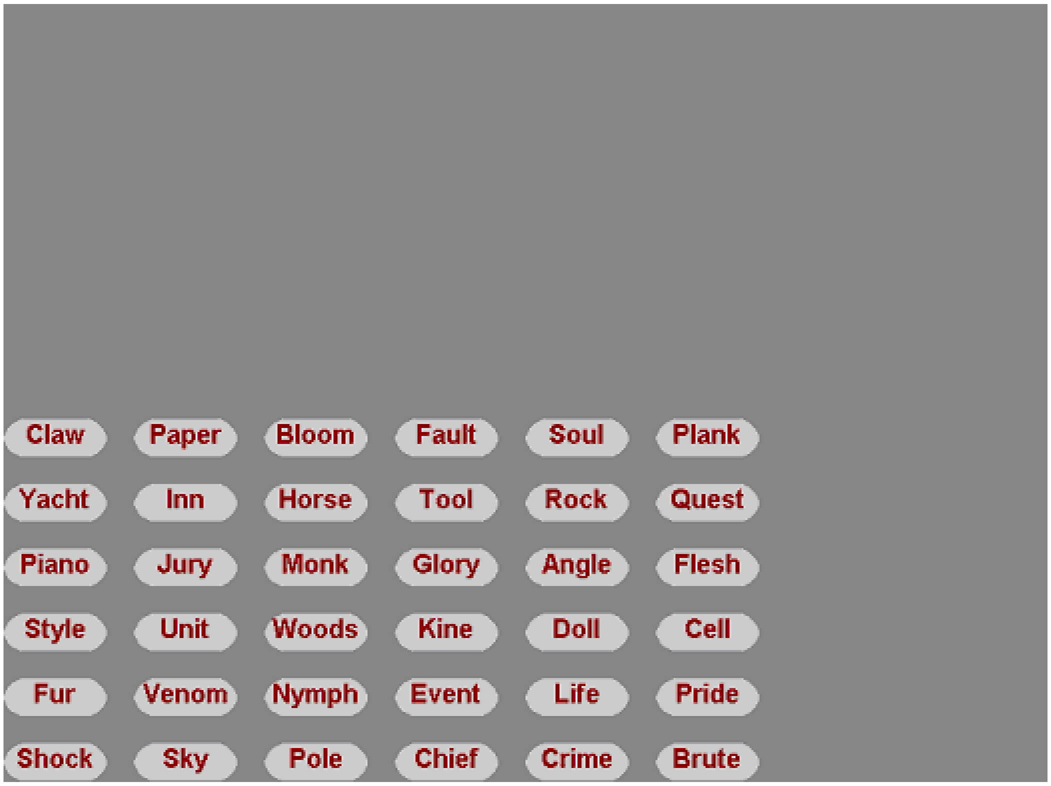
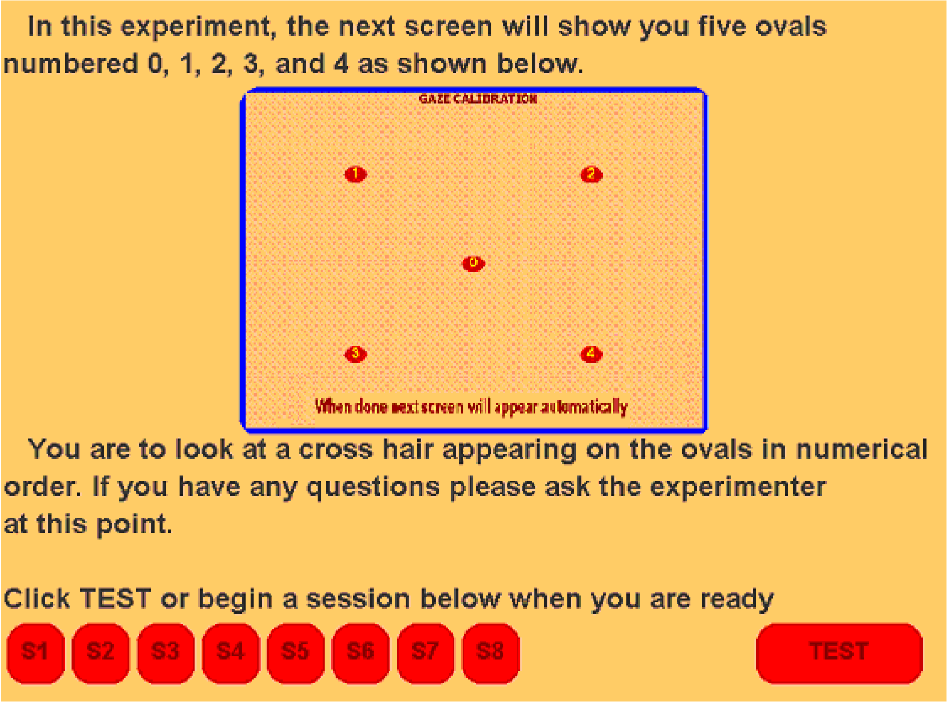


Figure 22 Lower-Left Screen

5.4 *Experimental Procedure*

5.4.1 Gaze Calibration

In this experiment, the subject doesn't need to go through the speech training again. Figure 23 shows the calibration instructions screen in Experiment 2. Functionally, the calibration instruction screens in both experiments are identical. They differ only in the number of buttons at the bottom because Experiment 1 contains only one session whereas Experiment 2 contains 8 sessions. The calibration instructions, i.e., *screen layout*, are different in both experiments to better manage the presentation of the experiment to the subject in a uniform manner.



In this experiment, the next screen will show you five ovals numbered 0, 1, 2, 3, and 4 as shown below.

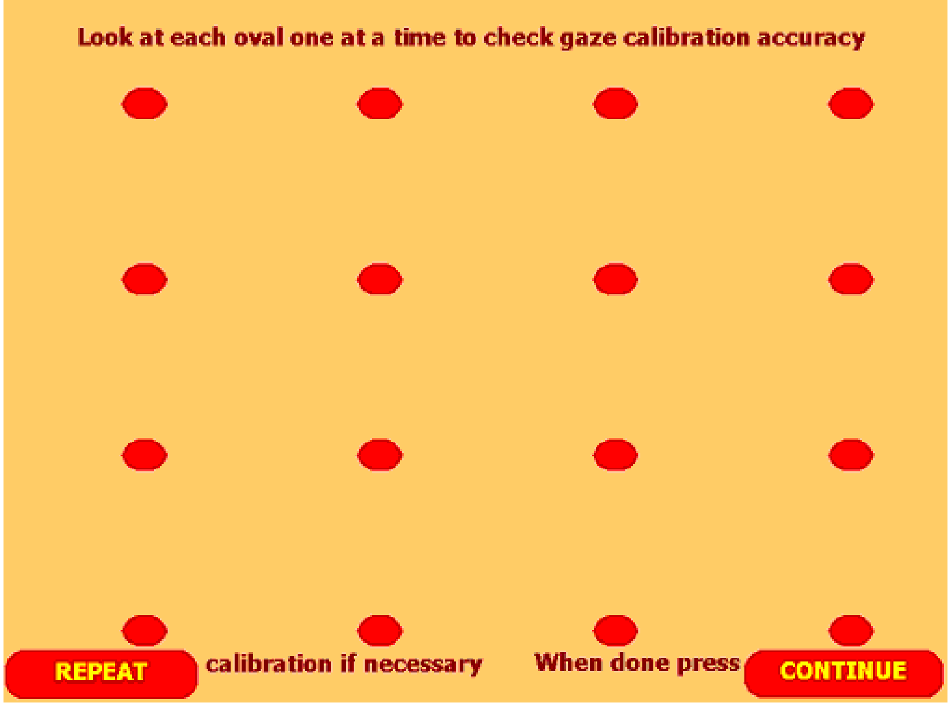
When done next screen will appear automatically

You are to look at a cross hair appearing on the ovals in numerical order. If you have any questions please ask the experimenter at this point.

Click TEST or begin a session below when you are ready

Figure 23 Gaze Calibration Instructions

The gaze calibration screen contains 5 points as shown in Experiment 1. After the gaze calibration is complete, the subject performs a calibration accuracy check. Figure 24 shows the calibration accuracy verification screen in Experiment 2. The subject is instructed to look at the 16 points and the points on the screen change color (from red to blue) when they are looked at successfully (i.e., gaze falls close to the point target). The experimenter verifies the calibration accuracy by noting whether all of the points turn blue. If this is the case, the experimenter tells the subject to begin the experiment. Otherwise, the subject repeats the calibration. The calibration accuracy screen in Experiment 2 is different from the calibration accuracy screen in Experiment 1. In Experiment 1, randomly selected 10 point targets are chosen to cover the screen area to verify the calibration accuracy. In Experiment 2, 16 point targets in a well-defined layout are used. The calibration accuracy doesn't depend on the number of point targets and location of point targets. So, the calibration accuracy screen differences in these experiments can be safely ignored. As the task in Experiment 2 is more complicated than the task in Experiment 1, a well-defined layout of a larger number of point targets helps in *measuring* calibration accuracy. The experimenter's visual verification is deemed enough to check the accuracy. After the speech training, gaze calibration and calibration accuracy verification, the subject proceeds to perform the task of the experiment.



Look at each oval one at a time to check gaze calibration accuracy

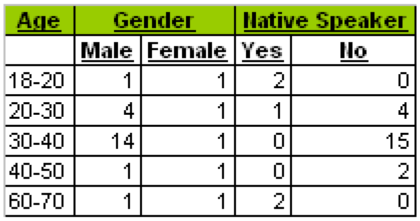
REPEAT calibration if necessary

When done press CONTINUE

Figure 24 Gaze Calibration Accuracy Panel

5.4.2 Subject Population

Table 8 describes the attributes of the 26 the subjects participating in Experiment 2. As in Experiment 1, the subject population is a mix of gender, age, and native/non-native speakers.



Age	Gender		Native Speaker	
	Male	Female	Yes	No
18-20	1	1	2	0
20-30	4	1	1	4
30-40	14	1	0	15
40-50	1	1	0	2
60-70	1	1	2	0

Table 8 Experiment 2 Subject Population Characteristics

Note that a few subjects couldn’t perform the experiment (not included in population characteristics) because the gaze calibration didn’t work for them as they were wearing either eye glasses or contact lenses for corrected vision.

5.5 Summary

In this chapter, a speech/gaze experiment was described whose purpose was to evaluate speech/gaze interactions in a menu selection task typical of many human/computer applications. The design of the experiment, the procedures employed, and the subjects’ characteristics were described. The issues/limitations involved in carrying out the experiments were addressed to ensure that the data collected is valid. The next chapter analyzes the data to establish parameters that are important to the development of practical human/computer interface applications.

6. Experiment 2: Results

6.1 Introduction

In this chapter, the data collected from the second of the two experiments is analyzed to illustrate the differences in gaze patterns across subjects and across time for a single subject. Then, linear and adaptive prediction models of the subject's behavior are compared. The intent of this work is to demonstrate that adaptive prediction is necessary and improves the overall accuracy in assessing the user's intent. To do this, the typical approaches to prediction are calculated and compared to an adaptive prediction approach. Toward this end, five different approaches to extracting the user's intent (speech only, gaze only, linear prediction, adaptive prediction, combined speech and adaptive prediction) are evaluated and compared.

6.2 Dominant Gaze Features in Menu Task

Figure 25 and Figure 26 indicate dominant gaze features for the menu task experiment where the task is more complicated than in the one word task. Figure 25 shows that the gaze features *fb1, fa1,* and *fa2* are the dominant gaze features among those computed. Although there are some differences in dominant gaze patterns across different font sizes and spacings, the (*fb1, fa1, fa2*) combination seems to be the dominant gaze feature combination for the menu interaction, as can be observed in the eight conditions illustrated in Figure 26 (note that although Figure 26b, shows significant values for fi, the fixation around speech start time with the largest number of samples, it is not included since, as mentioned previously, it includes fixations already found in (*fb1, fa1, fb2*) and is therefore not an independent feature). Figure 25 and Figure 26 also show that the dominant gaze features differ not only across users but also differ depending on the interaction task. In the simple one word task, *fa1/fii* was the dominant gaze feature combination whereas in the more complex menu interaction task <*fb1, fa1, fb2*> was shown to be the dominant combination.

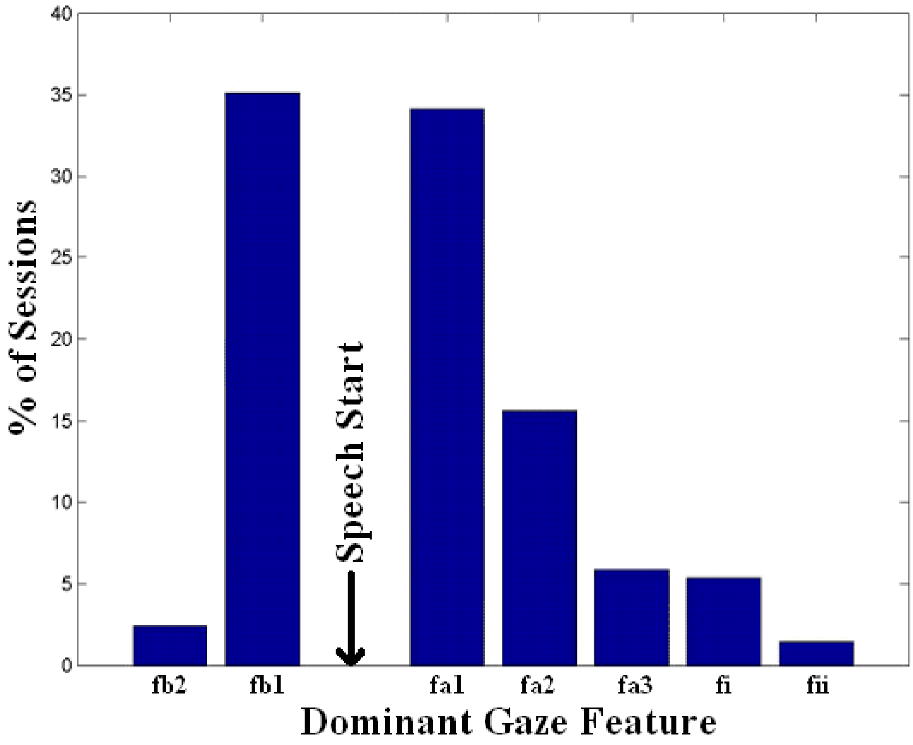
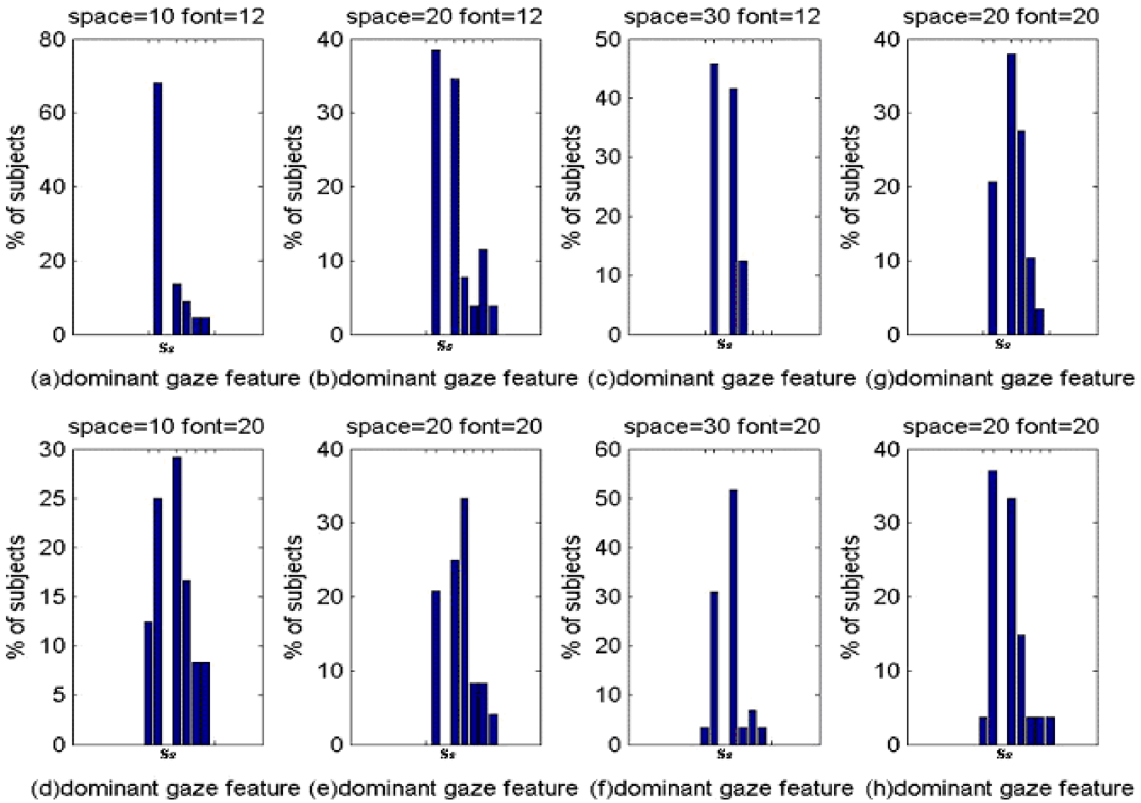


Figure 25 Dominant Gaze Feature for All Subjects in Menu Task



(a)dominant gaze feature (b)dominant gaze feature (c)dominant gaze feature (g)dominant gaze feature

(d)dominant gaze feature (e)dominant gaze feature (f)dominant gaze feature (h)dominant gaze feature

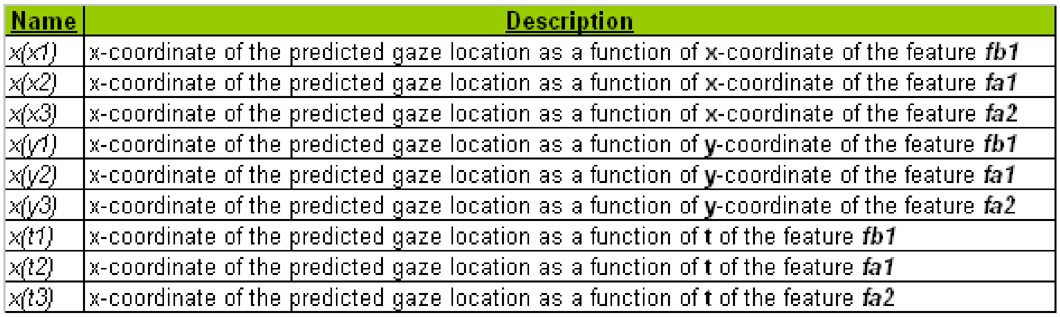
There are seven bars in each graph corresponding to features (fb2, fb1, fa1, fa2, fa3, fi, fii) with Speech Start (Ss) acting as the seperator. (fb2, fb1) are the features before Ss and (fa1, fa2, fa3, fi, fii) are the features after Ss.

Figure 26 Dominant Gaze for each Session in Menu Interaction

6.3 Adaptation Coefficients

In this section the adaptation coefficients are analyzed to observe how efficiently RAP is able to adapt to the changes in speech/gaze interactions. The adaptation coefficients for a single user from Experiment 2 are analyzed. The adaptation coefficients from Experiment 1 contain only 2 features whereas from Experiment 2 they contain 3 features and hence Experiment 2 is selected to show the coefficients.

In Experiment 2, RAP takes input feature set *(fb1, fa1, fa2)* with the parameters *(x, y, t)* for each feature and produces an output vector which is the predicted gaze location *(x, y)* based on the most recent window of interactions. The *x*-coordinate of the predicted gaze location depends on the 9 parameters [3 parameters $(x_i, y_i, t_i,)$ for each feature *i* from (*fb1, fa1, fa2)*] described in Table 9 and plotted for one subject, subject u2, in Figure 27. Each of the graphs shows the coefficient value during the adaptation process for all the scanpaths or interactions for the subject. At each step of the adaptation process, a window of past interactions *W* is chosen to calculate the coefficient values to predict the next scanpath.



Name	Description
x(x1)	x-coordinate of the predicted gaze location as a function of **x**-coordinate of the feature ***fb1***
x(x2)	x-coordinate of the predicted gaze location as a function of **x**-coordinate of the feature ***fa1***
x(x3)	x-coordinate of the predicted gaze location as a function of **x**-coordinate of the feature ***fa2***
x(y1)	x-coordinate of the predicted gaze location as a function of **y**-coordinate of the feature ***fb1***
x(y2)	x-coordinate of the predicted gaze location as a function of **y**-coordinate of the feature ***fa1***
x(y3)	x-coordinate of the predicted gaze location as a function of **y**-coordinate of the feature ***fa2***
x(t1)	x-coordinate of the predicted gaze location as a function of **t** of the feature ***fb1***
x(t2)	x-coordinate of the predicted gaze location as a function of **t** of the feature ***fa1***
x(t3)	x-coordinate of the predicted gaze location as a function of **t** of the feature ***fa2***

Table 9 Adaptation Coefficients

It can be seen from Figure 27 that for the user u2, the influence of *fb1*, *fa1* is less than the influence of *fa2* in predicting the x-coordinate of the gaze location (Figure 27a, b, c). Also, the x coordinate of the predicted location depends only on the x coordinates of the three features and has little relationship to the y-coordinate of the features (Figure 27d, e, f) because the coefficient values of the $x(y_i)$ are much smaller compared to $x(x_i)$. The very small coefficient values of $x(t_i)$

indicate that the predicted gaze location has little or no dependency on the fixation's timestamp (Figure 27g, h, i) with reference to speech onset time. This is because the time is already factored into the fixations' classification with reference to speech onset time.

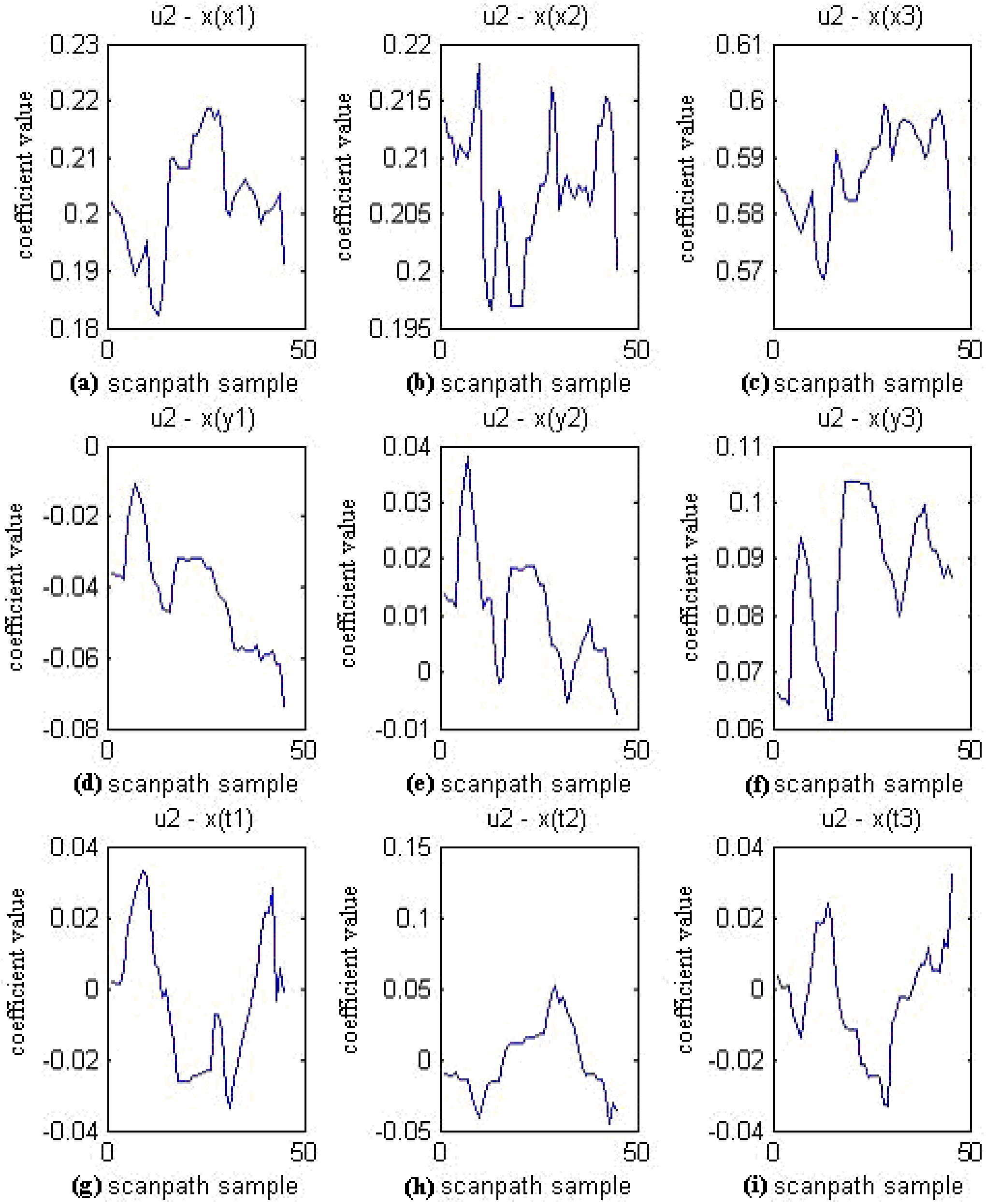


Figure 27 RAP Coefficients for Predicting x coordinate of the gaze location (x, y)

Figure 28 shows the coefficient values of the y-coordinate. Again, it can be seen that the y-coordinate of predicted gaze location depends only on the y-coordinates of the three features (Figure 28d, e, and f) and doesn't depend a lot on the x-coordinates (Figure 28a, b, and c). The time dependency can be attributed to noise similar to the x coordinate prediction (Figure 28g, h, i). The RAP algorithm can be enhanced to converge to the solution space faster by various techniques which are not explored as part of this dissertation and are left for future studies.

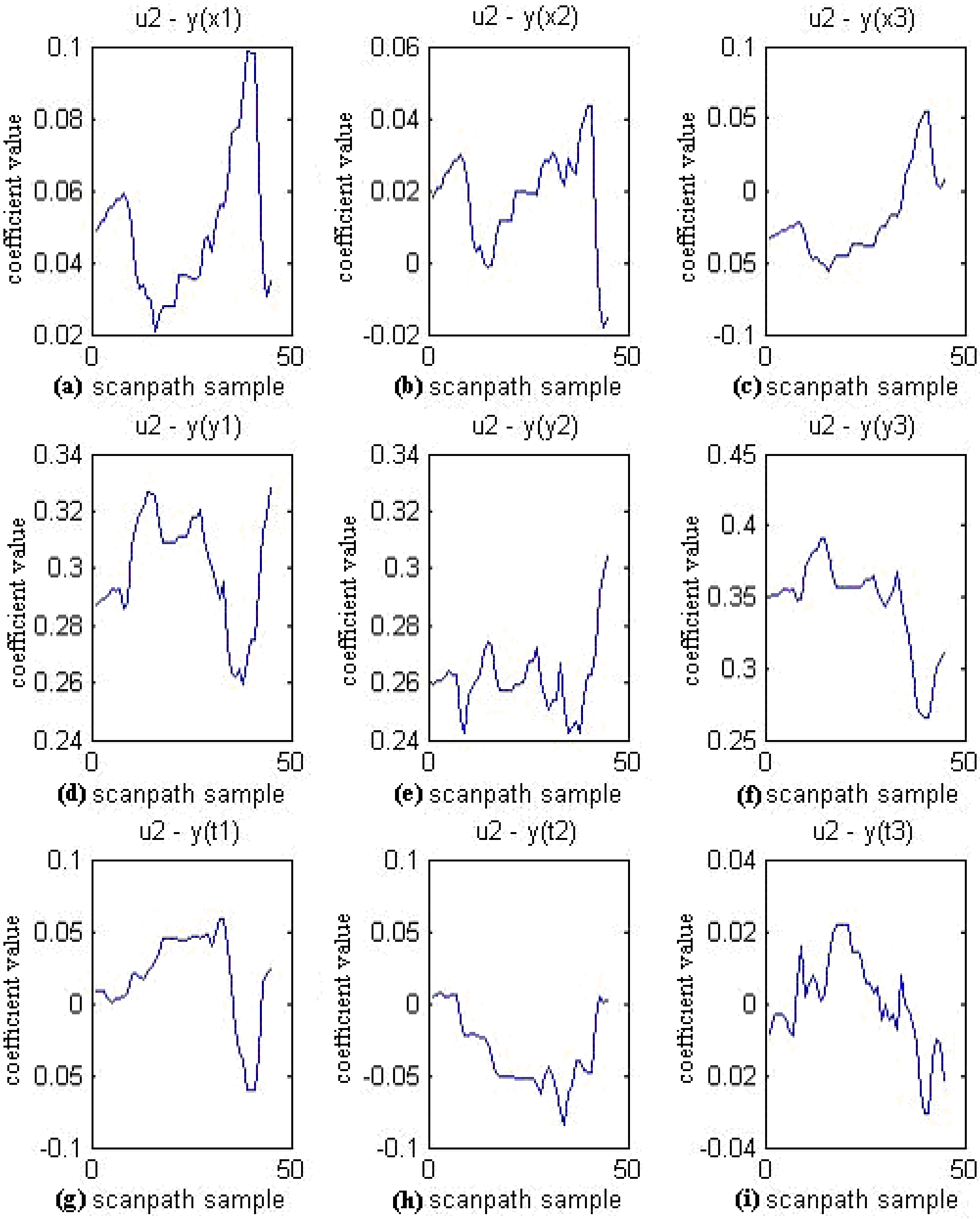


Figure 28 RAP Coefficients for Predicting y coordinate of the gaze location (x, y)

6.4 Experiment 2 Performance

6.4.1 Effect of font size and spacing on performance

Figure 29 and Figure 30 show the Gaze/RAP prediction-based target detection probabilities for all sessions individually in Experiment 2 for both fixed (Figure 29) and optimum (Figure 30) parameters. Specific space/font settings are associated with each graph. Although the data are noisy, they suggest that as the spacing increases the detection probability improves, but there is no significant difference in detection probability associated with changes in font size. A more detailed statistical analysis presented in Section 6.4 below supports these contentions. Note that graphs (g) and (h) of Figure 29 and Figure 30 also show that detection probabilities did not deteriorate significantly in the corner sessions.

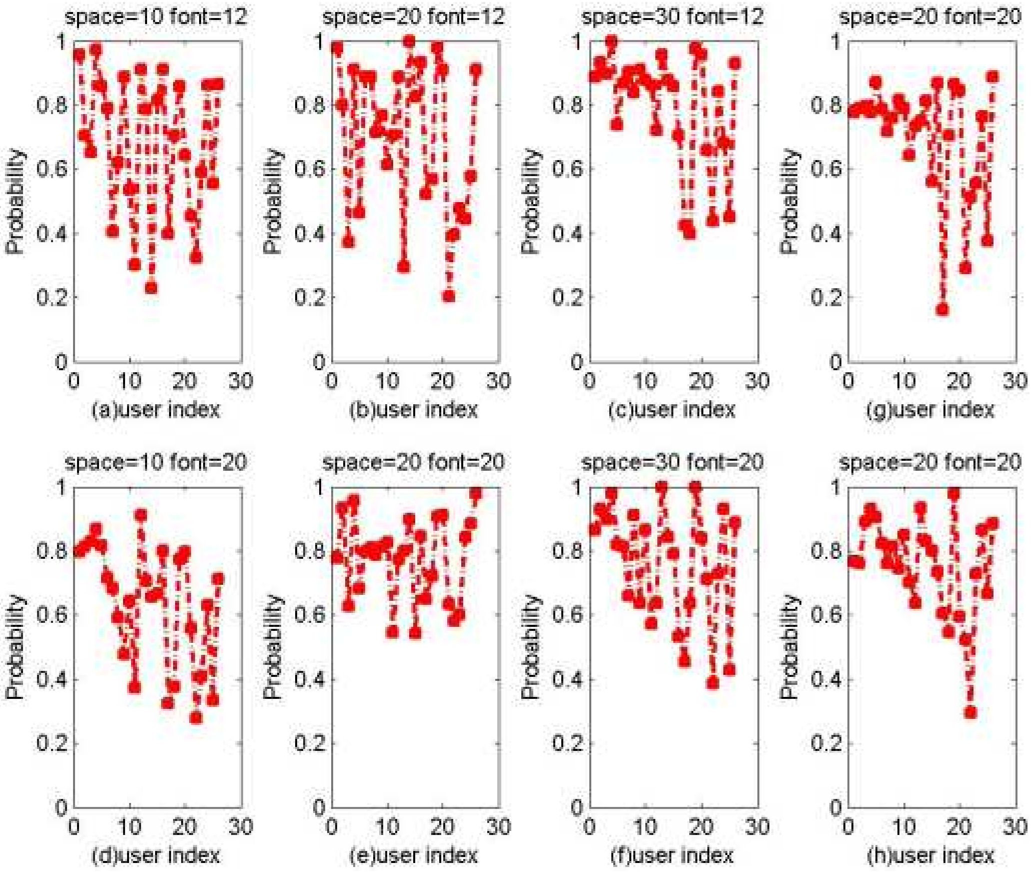


Figure 29 Session Performances (Gaze/RAP) in Menu Interaction (fixed parameters)

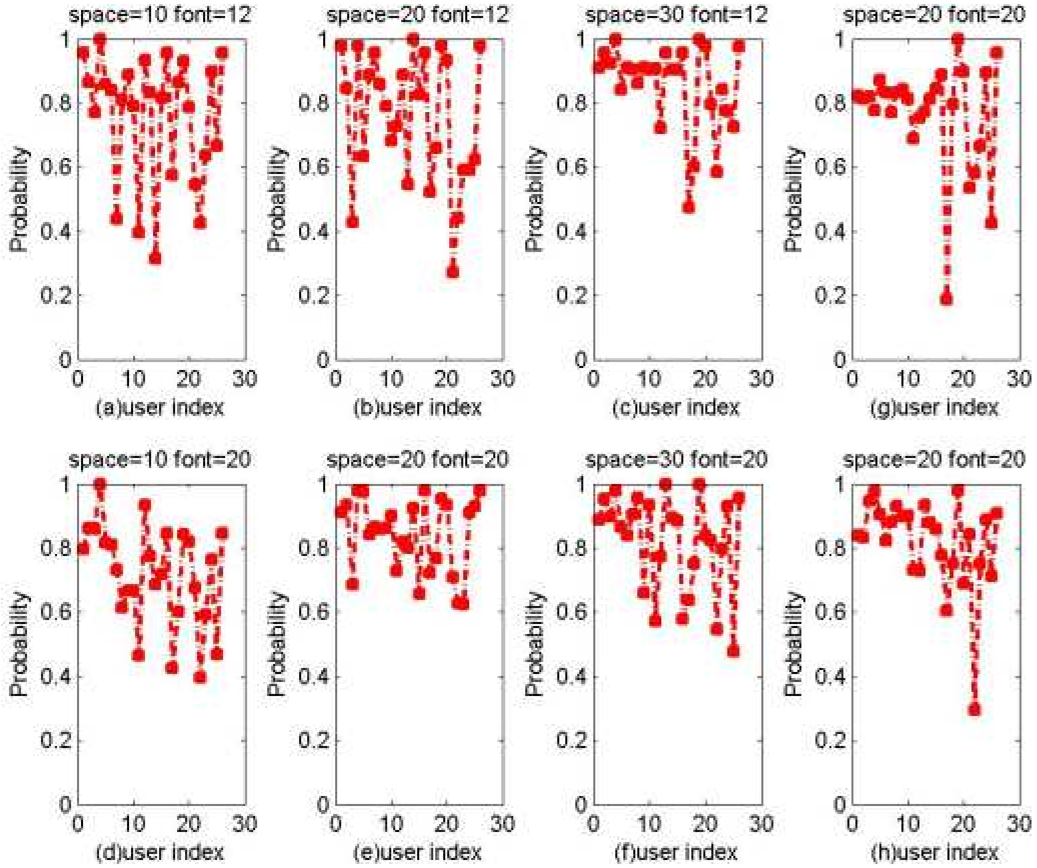
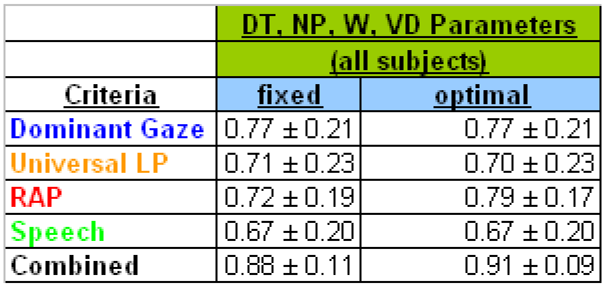


Figure 30 Session Performances (Gaze/RAP) in Menu Interaction (optimal parameters)

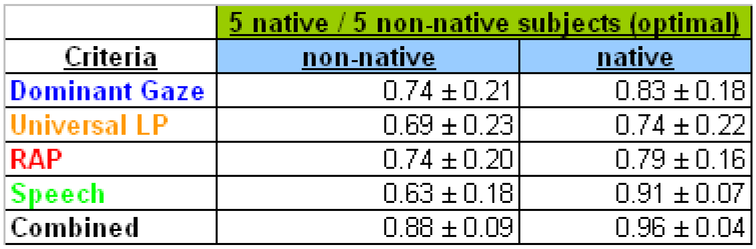
6.4.2 Statistical Analysis: Paired Samples t-Tests

For Experiment 2, Table 10 and Table 11 show the means and standard deviations of the target detection probabilities for all subjects and native/non-native subjects respectively. As in Experiment 1, the same set of internal parameters was used for all subjects in the "fixed" case. For the optimal case, the internal parameters were selected using a semi-automated process. This process used an algorithm that computed results for a range of values of DT, NP, W, and VD (W = 6, 9, 12, VD = 200, DT = 20, 30, 40, NP = 8; for a total of 3 x 1 x 3 x 1 = 9 cases) for all subjects for all cases and automatically picked the optimal case (i.e., maximizing the RAP) for each subject.



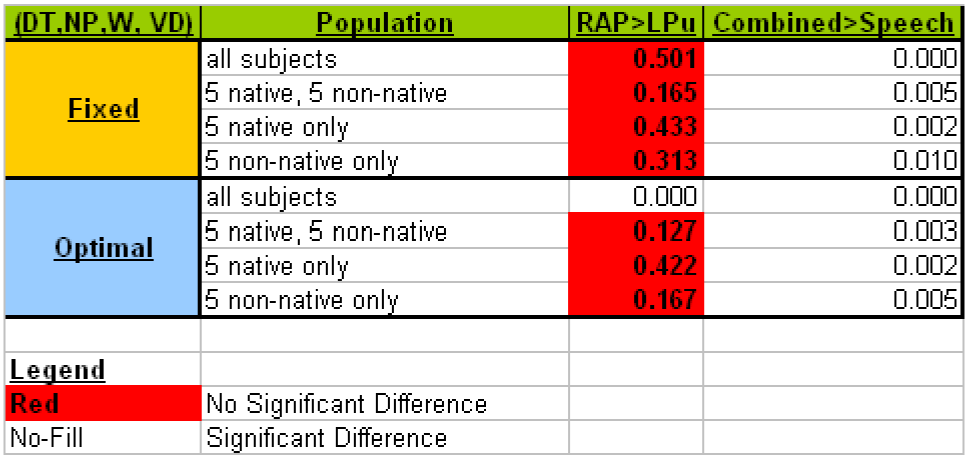
	DT, NP, W, VD Parameters	
	(all subjects)	
Criteria	fixed	optimal
Dominant Gaze	0.77 ± 0.21	0.77 ± 0.21
Universal LP	0.71 ± 0.23	0.70 ± 0.23
RAP	0.72 ± 0.19	0.79 ± 0.17
Speech	0.67 ± 0.20	0.67 ± 0.20
Combined	0.88 ± 0.11	0.91 ± 0.09

Table 10 μ/σ of Target Detection Probabilities in Experiment 2 (all subjects)



	5 native / 5 non-native subjects (optimal)	
Criteria	non-native	native
Dominant Gaze	0.74 ± 0.21	0.83 ± 0.18
Universal LP	0.69 ± 0.23	0.74 ± 0.22
RAP	0.74 ± 0.20	0.79 ± 0.16
Speech	0.63 ± 0.18	0.91 ± 0.07
Combined	0.88 ± 0.09	0.96 ± 0.04

Table 11 μ/σ of Target Detection Probabilities in Experiment 2 (non-native vs. native subjects)



(DT,NP,W, VD)	Population	RAP>LPu	Combined>Speech
Fixed	all subjects	**0.501**	0.000
	5 native, 5 non-native	**0.165**	0.005
	5 native only	**0.433**	0.002
	5 non-native only	**0.313**	0.010
Optimal	all subjects	0.000	0.000
	5 native, 5 non-native	**0.127**	0.003
	5 native only	**0.422**	0.002
	5 non-native only	**0.167**	0.005
Legend			
Red	No Significant Difference		
No-Fill	Significant Difference		

Table 12 *p*-values of paired samples t-Test in Experiment 2

Table 12 shows the *p*-values for various population categories in Experiment 2, where a value less than 0.05, indicates that there is a significant difference between the two samples being compared. Fixed and optimal parameters are considered for each population category to evaluate the significance of the comparison of detection techniques. The table shows that the combined speech/gaze system is significantly better than the speech-only system in all conditions for all population categories. However, although a significant improvement by the adaptation model (RAP) was shown over the non-adaptation model (LPu) when looking across all subjects, using optimal parameters, no significant improvement of RAP over LPu was seen for fixed parameters and for the sub-populations for optimum parameters, particularly for the case of high-performing native-only speakers. The very small subject samples in the sub-populations make drawing conclusions from these results difficult. Further testing with larger populations is required.

Summarizing the main t-Test results for Optimized Adaptive vs. Non-adaptive, Combined Speech/Optimized Adaptive Gaze vs. Speech alone, Combined Speech/Non-Optimized Adaptive vs. Speech alone:

All Subjects

- Paired Samples t-Test using target detection probability
  - Null Hypothesis H0: $\mu 1 = \mu 2$
    - If $t(df) > t_{critical}$ reject H0 where *df* is degrees of freedom
  - Optimized Adaptive (RAP) vs. Non-Adaptive (Linear Prediction)
    - Adaptive model ($\mu = 0.79$, $\sigma = 0.17$) **performed significantly better** than non-adaptive model [$\mu = 0.70$, $\sigma = 0.23$, t(25) = 4.067, $p < 0.0005$]
  - Combined Speech/Optimized Adaptive (RAP) vs. Speech alone
    - Combined ($\mu = 0.91$, $\sigma = 0.09$) **performed significantly better** than Speech alone [$\mu = 0.67$, $\sigma = 0.20$, t(25) = 10.279, $p < 0.0005$]
  - Combined Speech/Non-Optimized Adaptive (RAP) vs. Speech alone
    - Combined ($\mu = 0.88$, $\sigma = 0.11$) **performed significantly better** than Speech alone [μ = 0..67, $\sigma = 0.20$, t(25) = 9.747, $p < 0.0005$]

Native Subjects

- Paired Samples t-Test using target detection probability
  - Optimized Adaptive (RAP) vs. Non-Adaptive (Linear Prediction)
    - Adaptive model ($\mu = 0.79$, $\sigma = 0.16$) **did not perform significantly better** than non-adaptive model [$\mu = 0.74$, $\sigma = 0.22$, t(4) = 0.894, $p < 0.422$]
  - Combined Speech/Optimized Adaptive (RAP) vs. Speech alone
    - Combined ($\mu = 0.96$, $\sigma = 0.04$) **performed significantly better** than Speech alone [$\mu = 0.91$, $\sigma = 0.07$, t(4) = 7.193, $p < 0.002$]

- Combined Speech/Non-Optimized Adaptive (RAP) vs. Speech alone
    - Combined ($\mu = 0.95$, $\sigma = 0.04$) **performed significantly better** than Speech alone [$\mu = 0.91$, $\sigma = 0.07$, t(4) = 7.013, $p < 0.002$]

6.4.3 Interface Usability Performance

The, *Interface usability performance* (described previously in Chapter 4) denotes the effectiveness of an interface in terms of the percentage of the times the target detection probability is higher than an acceptable target detection probability. The interface performance for the menu task for all subjects is shown in Figure 31 (fixed parameters) and Figure 32 (optimum parameters). Now the menu task had 26 subjects, 8 sessions per subject and a total of 26x8 = 208 total sessions for those 26 subjects. The interface performance value for any given target detection probability P is the ratio of the number of sessions out of the 208 that had a probability of successful target detection greater than P. These curves corroborate the findings of the paired sample t-test. Note that the RAP performance tracks the LPu performance for fixed parameters, but for optimum parameters exceeds LPu and closely tracks the ideal Dominant Gaze performance. At the same time, the “Combined” performance exceeds that of all other techniques. Note that the interface performance for this more difficult menu task is not as high as that of the simple, one-word task (Figure 13). It is important to keep in mind that the features used in the one-word task are not the same as those used in the menu task, a further reminder that task complexity does have a strong influence on the design of a speech/gaze application.

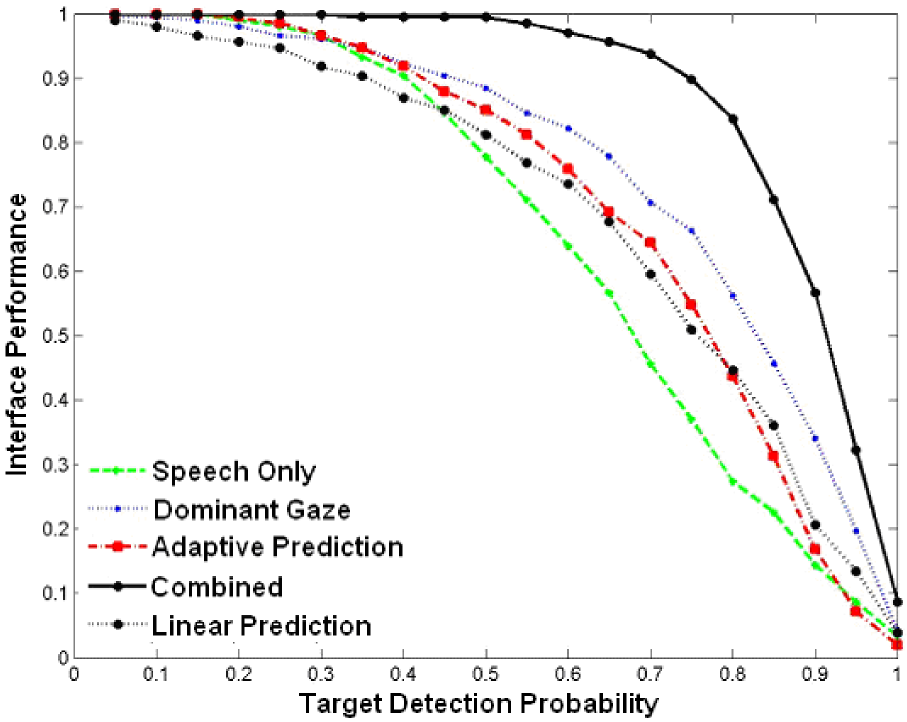


Figure 31 Usability curves (fixed parameters/all subjects)

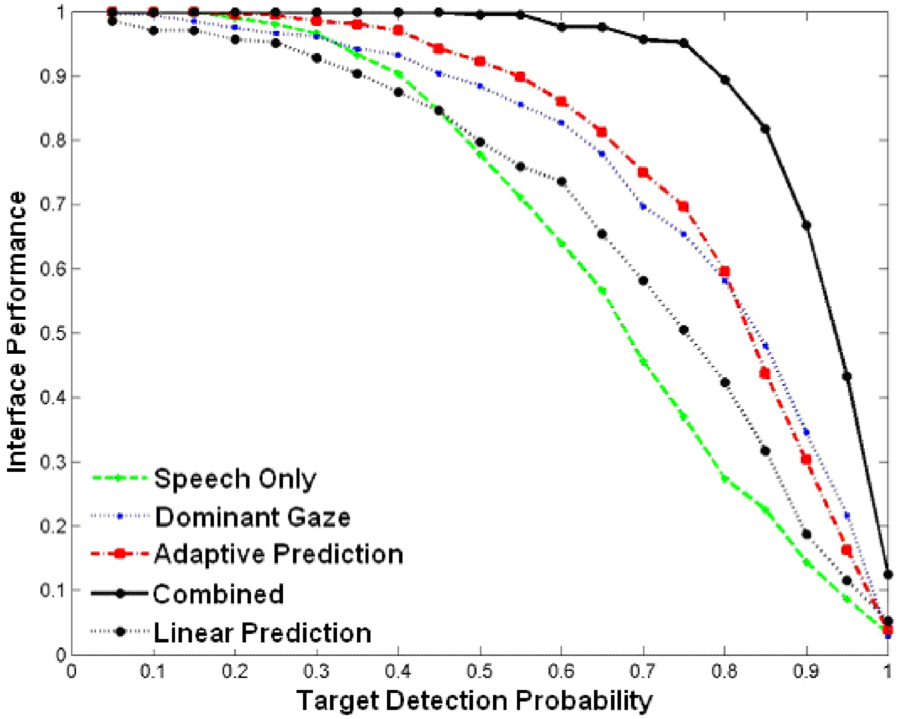


Figure 32 Usability curves (optimal parameters/all subjects)

Out of 26 subjects in Experiment 2, 5 are native speakers and 21 are non-native speakers. Figure 33 lists the main performance probabilities; *Speech Only, Dominant Gaze, Adaptive Prediction (RAP), Linear Prediction* and *Combined* (*Speech/Gaze)* for the same collection of native and non-native speakers drawn from the 26 subjects for different sessions. It shows equal numbers of native and non-native speakers selected such that all the subjects are closely spaced in time in terms of when they ran the experiment. Figure 33 (fixed parameters) and Figure 34 (optimum parameters) show the interface performance for these two combined sub-populations. Note that the plots bear the same basic relations to each other as those for the full population, only with much smaller numbers.

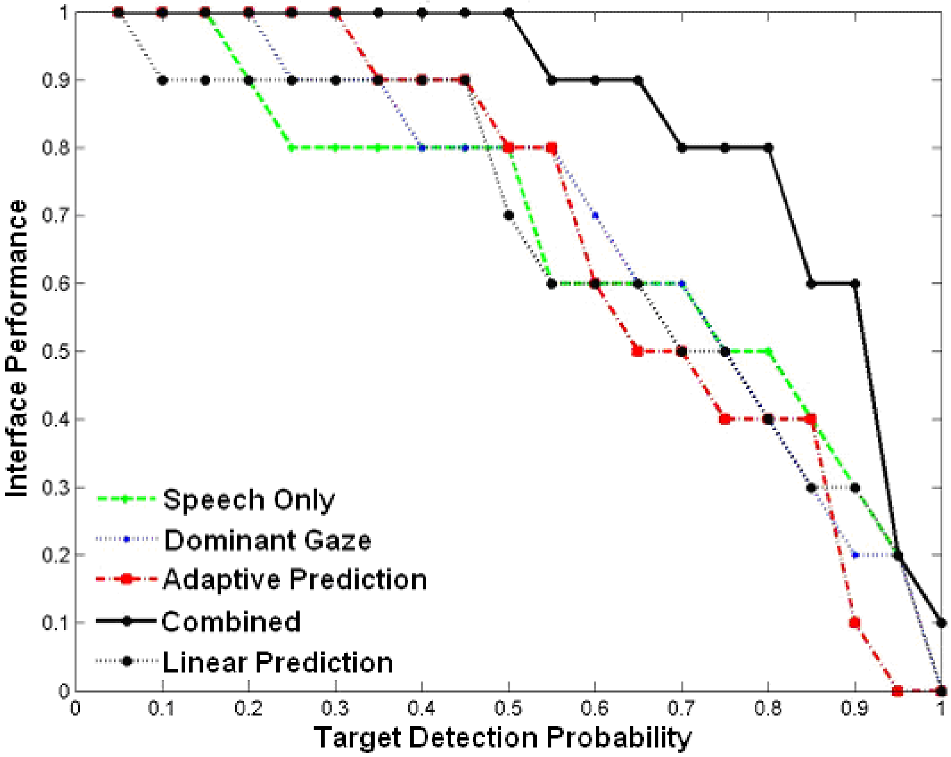


Figure 33 Interface Performance for 5 Native/5 non-Native Speakers (fixed parameters)

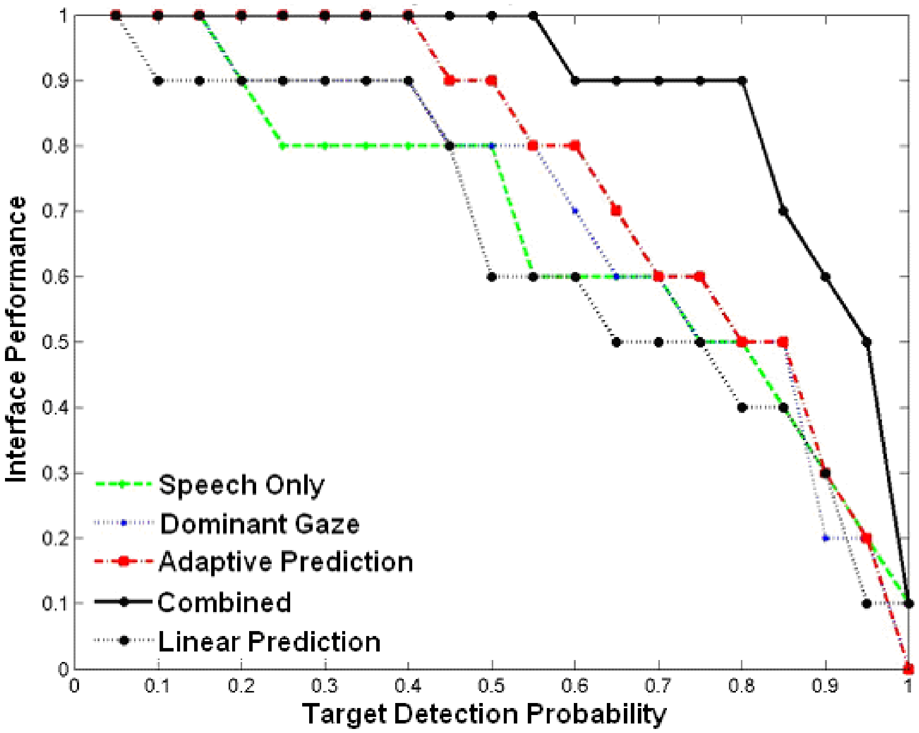
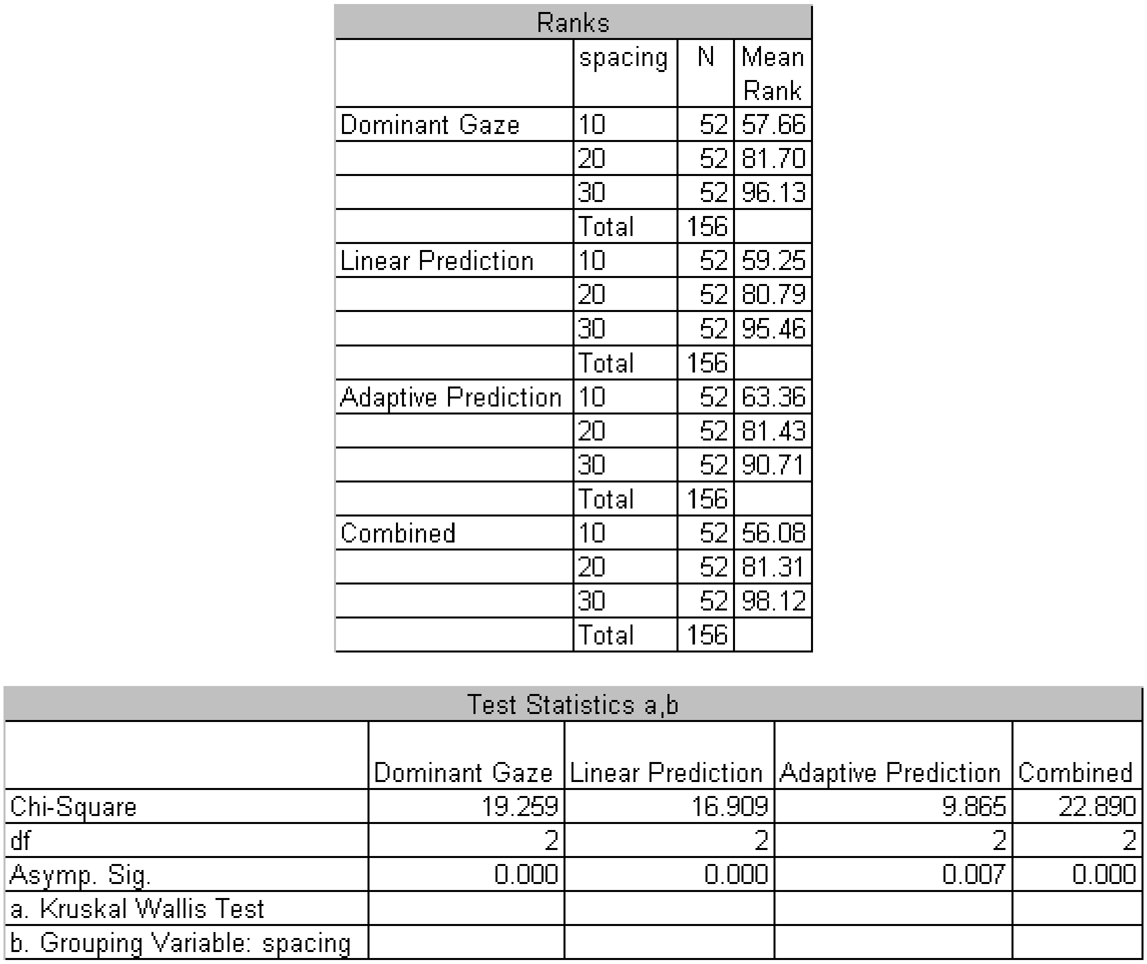


Figure 34 Interface Performance for 5 Native/5 non-Native Speakers (optimum parameters)

6.4.4 Kruskal-Wallis Test for Spacing and Font Size Effects

In the Experiment 2, spacing and font-size are the primary variables of the interface design which govern the target detection probability, apart from other factors like speech recognition and eye tracking performance. In analyzing the data, the Kruskal-Wallis statistical test, a non-parametric test, is employed with independent variables spacing and font-size. Although non parametric, this test is chosen primarily to rank order the spacing and font-size values with respect to a specific target detection probability. The dependent variable chosen is the session target detection probability. The test involves rank ordering, into a single list all of the values of detection probability for all groups being compared (e.g. three spacings, two font sizes) and then re-grouping those rankings within the individual groups. A significance test is then carried out on the mean rankings for each group. The asymptotic significance is given by P(Chi^2 >= 'value') = p-value and p-values less than 0.005 are considered significant, i.e. the rank order differences between groups are considered significant.

The session target detection probability was analyzed for four different detection techniques (speech only doesn’t enter into this) and 3 different spacings (10, 20 and 30 pixels). Table 13 shows the mean ranks for different samples, each with one of the three spacings, for optimal parameters. *df* indicates the degrees of freedom and N indicates the number of values in the sample. The *p-value* significance of the mean ranks for each of the spacings showed significant improvement in target detection probability as spacing increased for Dominant Gaze, Linear Prediction and Combined speech/gaze (adaptive prediction just barely missed).



Ranks

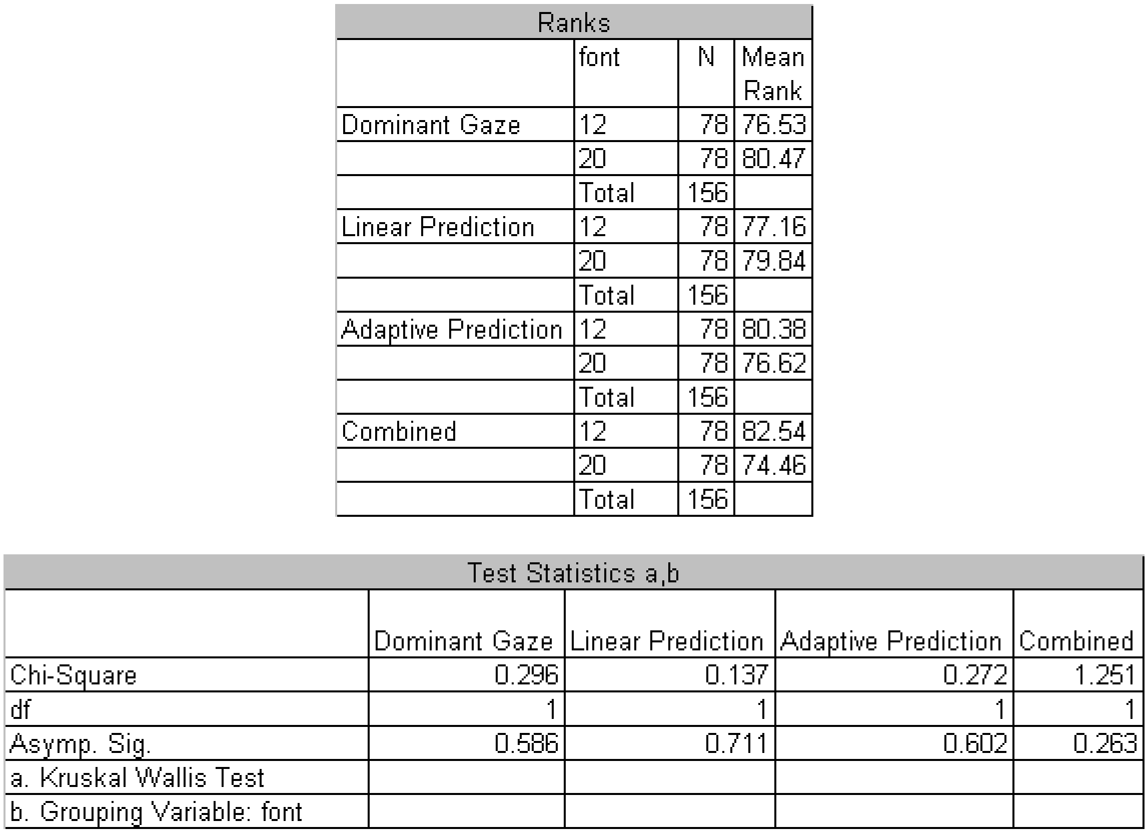
	spacing	N	Mean Rank
Dominant Gaze	10	52	57.66
	20	52	81.70
	30	52	96.13
	Total	156	
Linear Prediction	10	52	59.25
	20	52	80.79
	30	52	95.46
	Total	156	
Adaptive Prediction	10	52	63.36
	20	52	81.43
	30	52	90.71
	Total	156	
Combined	10	52	56.08
	20	52	81.31
	30	52	98.12
	Total	156	

Test Statistics a,b

	Dominant Gaze	Linear Prediction	Adaptive Prediction	Combined
Chi-Square	19.259	16.909	9.865	22.890
df	2	2	2	2
Asymp. Sig.	0.000	0.000	0.007	0.000
a. Kruskal Wallis Test				
b. Grouping Variable: spacing				

Table 13 Spacing (in pixels) effect on target detection probability in Experiment 2

Table 14 shows the mean ranks for different samples each with a different font-size (12 point, 20 point) for optimal parameters. Font-size did not show significant improvement in target detection probability regardless of the detection technique (p-values much greater than 0.005) Thus, it clearly indicates that the spacing has a very significant effect in the interface performance while the font-size has strong influence on the performance but not significant.



Ranks			
	font	N	Mean Rank
Dominant Gaze	12	78	76.53
	20	78	80.47
	Total	156	
Linear Prediction	12	78	77.16
	20	78	79.84
	Total	156	
Adaptive Prediction	12	78	80.38
	20	78	76.62
	Total	156	
Combined	12	78	82.54
	20	78	74.46
	Total	156	

Test Statistics a,b				
	Dominant Gaze	Linear Prediction	Adaptive Prediction	Combined
Chi-Square	0.296	0.137	0.272	1.251
df	1	1	1	1
Asymp. Sig.	0.586	0.711	0.602	0.263
a. Kruskal Wallis Test				
b. Grouping Variable: font				

Table 14 Font-size (in points) effect on target detection probability in Experiment 2

6.4 Summary

This chapter illustrated that a linear, time varying system is adequate to represent an adaptive speech/gaze integration model using a RAP technique.

The main hypotheses under test in Experiment 2, the menu–based task were that for such a task:

- a combined speech/gaze system performs better than a speech-only system
- increasing spacing improves the gaze target detection probability
- increasing the font-size improves the gaze target detection probability

- screen location affects behavior with less central locations generating less accurate target detection probability.

The results of the t-test confirms the **first** hypothesis, i.e., that. a combined speech/gaze system performs better than a speech-only system.

Summarizing the main t-test results for (Optimized Adaptive) vs. (Non-adaptive), (Combined Speech/Optimized Adaptive Gaze) vs. (Speech alone), (Combined Speech/Non-Optimized Adaptive) vs. (Speech alone):

All Subjects

- Paired Samples t-Test using target detection probability
    - Null Hypothesis H0: μ1 = μ2
        - If t(df) > $t_{critical}$ reject H0 where *df* is degrees of freedom
    - Optimized Adaptive (RAP) vs. Non-Adaptive (Linear Prediction)
        - Adaptive model (μ = 0.79, σ = 0.17) **performed significantly better** than non-adaptive model [μ = 0.70, σ = 0.23, t(25) = 4.067, *p* < 0.0005]
    - Combined Speech/Optimized Adaptive (RAP) vs. Speech alone
        - Combined (μ = 0.91, σ = 0.09) **performed significantly better** than Speech alone [μ = 0.67, σ = 0.20, t(25) = 10.279, *p* < 0.0005]
    - Combined Speech/Non-Optimized Adaptive (RAP) vs. Speech alone
        - Combined (μ = 0.88, σ = 0.11) **performed significantly better** than Speech alone [μ = 0..67, σ = 0.20, t(25) = 9.747, *p* < 0.0005]

Native Subjects

- Paired Samples t-Test using target detection probability
    - Optimized Adaptive (RAP) vs. Non-Adaptive (Linear Prediction)
        - Adaptive model (μ = 0.79, σ = 0.16) **did not perform significantly better** than non-adaptive model [μ = 0.74, σ = 0.22, t(4) = 0.894, *p* < 0.422]
    - Combined Speech/Optimized Adaptive (RAP) vs. Speech alone
        - Combined (μ = 0.96, σ = 0.04) **performed significantly better** than Speech alone [μ = 0.91, σ = 0.07, t(4) = 7.193, *p* < 0.002]
    - Combined Speech/Non-Optimized Adaptive (RAP) vs. Speech alone
        - Combined (μ = 0.95, σ = 0.04) **performed significantly better** than Speech alone [μ = 0.91, σ = 0.07, t(4) = 7.013, *p* < 0.002]

The results of the Kruskal-Wallis test confirms that the **second** hypothesis, i.e. that increasing spacing improves the gaze target detection probability, i.e., performance improved significantly with increasing spacing [Asymptotic significance (p-value) <0.005] for Combined Speech/Optimized Adaptive (RAP).

The results of the Kruskal-Wallis test did **not** confirm the **third** hypothesis, i.e., that increasing the font-size improves the gaze target detection probability. There was no evidence that either increasing or decreasing the font size improved the target detection probability, i.e. p-value much greater than 0.005 for all detection techniques.

Analysis of the target detection probability plots for icon arrays in the corners of the displays did **not** confirm the **fourth** hypothesis, i.e., that screen location affects behavior with less central locations generating less accurate target detection probability. Plots (g) and (h) of Figure 29 and Figure 30 did not show a deterioration in performance in the corner sessions.

Other important findings from the Experiment 2 results were:

- The dominant gaze features for the menu task, experiment 2, were different than those found for the simpler one-word task, experiment 1. Although there were some differences in dominant gaze patterns across different fonts/spacings and display locations across the subject pool, the feature combination <*fb1, fa1, fa2*> provided the highest percentage of correct target detection. (the best performing combination for the one-word task in experiment 1 was <*fa1/fii*>).
- The adaptation coefficients were analyzed to understand the efficiency with which the RAP coefficients predict the x and y coordinates of the gaze location (x, y) in adapting to the changes in speech/gaze interactions.
- Performance curves showed that while the improvements of the speech/RAP combination over speech alone were very large when the speech performance was poor (extremely high noise situations), the improvements were still valuable when the speech performance was reasonably good. When the speech-alone recognition performance was between 90

and 95% (native speakers), the combined speech/RAP performance was between 95 and 100%, thereby raising the performance in some practical applications from marginal to acceptable.

- It was shown that a linear, time varying system is adequate to provide an adaptive speech/gaze integration system using a RAP technique.

7. Task Modeling

7.1. Introduction

As described in Chapter 4, Prasov *et. al* ([100] and [106]) showed that *fixation intensity, i.e.,* the duration of a fixation as expressed by the number of gaze samples in a fixation is an important feature contributing to the relationship between gaze and attention to an object on a display. However, it was shown, during the two experiments described in this dissertation, that there are other fixations that make more significant contributions to the prediction of the users attention in a speech/gaze system. In this section, the relationship between *fi,* the fixation with the longest duration in the neighborhood of the onset of speech, and the other fixations in that neighborhood is examined.

7.2 'fi' – better task modeler than attention predictor?

The *fi* fixation can occur anywhere around speech onset time and its time of occurrence is not predictable. Each *fi* fixation in a scanpath has an index associated with it indicating when/where it occurred with respect to speech onset time, e.g. for an index of 2, *fi = fa2*. Figure 35 shows the *fi* index for all subjects for different dispersion thresholds in Experiment 1. Almost all *fi* fixations occur around speech onset time and have indices in the range of [-10 10]. Very few *fi* fixations occur with indices out of this range. All fixations with indices out of this range are summed at the boundary indices -10 and 10. Figure 35 also shows that the dispersion threshold does not have a significant effect on the *fi* index. Since the *fi* fixation can cover any of the fixations *fb2, fb1, fa1,* and *fa2,* another feature *fii* is also considered in experiments 1 and 2 which is similar to *fi* but excludes *fa1,* since *fa1* was shown to make a significant contribution to successful target detection and *fi*'s contribution would be redundant.

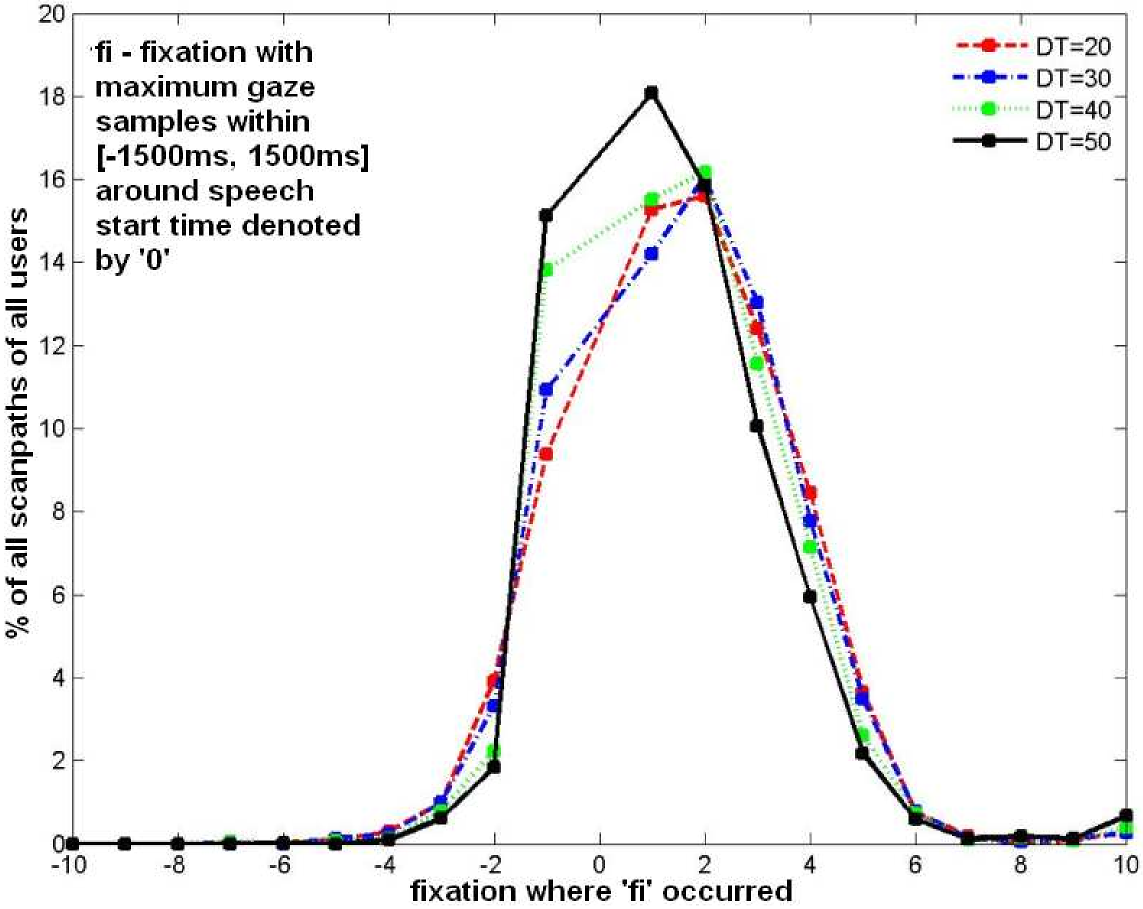


Figure 35 Fixation Intensity Profiles in Experiment 1 (Simple Task) with different dispersion thresholds (DT)

Figure 35 shows clearly that the *fa1* and *fa2* are the dominant fixations and for a large DT *fi* coincides with *fa1* more than with any other fixation. As the task complexity increases *e.g.*, simple word reading to more complex word reading in a display with many distractions, the fixation intensity appears to spread more around the speech onset time., Figure 36 illustrates, for both a simple (experiment 1) and a complex (experiment 2) task, the probability distribution of longest-duration fixations around speech onset time. Notice that the distribution of *fi* samples is more widely and asymmetrically spread for the more complex task. This makes sense, since more searching **prior** to target detection is required when the scene is more complicated.

In addition to task complexity, there are other reasons why *fi* may not be a consistent indicator of the user's focus on the target to be detected. First, *fi* measures the number of gaze samples in a fixation which depends, significantly, on the fixation algorithm. Second, an application may not choose the correct object of attention simply because the user concentrated on an object long enough to give the maximum number of gaze samples. A third reason, related to the second, is that as the user becomes familiar (i.e., trained) with the system, it is not guaranteed that *fi* indicates the user's attention on the object of interest.

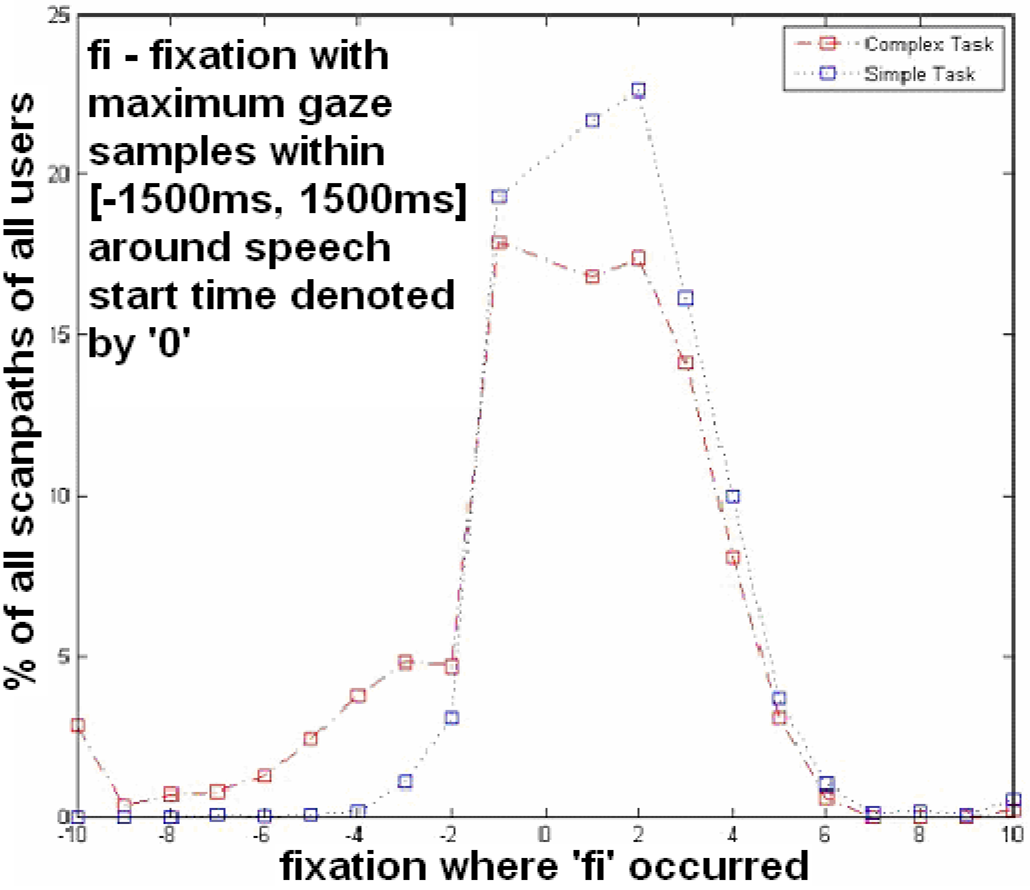


Figure 36 *fi* as a Task Complexity modeler instead of attention predictor

7.3 Summary

In summary, *fi* may not be a very good indicator of the user's attention to objects of interest on a display. The results of experiments 1 and 2 suggest that there are much better ones. However, Figure 36 did show that a complex task had a different distribution (wider and more asymmetric) than a simple one. This leads to a suggestion that the distribution of *fi* relative to speech onset time **may** be useful for modeling task complexity and could provide a useful tool to aid in interface design. Further research is needed to establish that result.

8. Summary of Contributions

8.1. Introduction

The main objective of this research was to explore the interactions between speech and gaze and to determine whether speech and gaze, acting together, can convey the user's intent to a computer-based system more effectively than either modality acting alone. Two experiments were carried out for this exploration. The first experiment involved isolated words on a computer display to be spoken by a subject and recognized by a speech recognition system. The purpose was to shed light on some of the fundamental relations involved in speech/gaze interactions. The second experiment involved a menu selection task, with multiple buttons containing words, one of which was to be selected, spoken and recognized by a speech recognition system. This experiment represents the typical menu-selection usage that would be expected to occur in the envisioned speech-gaze system and was used to gather data on speech/gaze interaction. In both experiments, the ability of a speech/gaze system to adapt to different user's requirements and to adapt to individual user's changes in behavior over time was explored. The main findings are summarized below.

8.2. Contributions

- When a user finds an object (e.g. word) on a display and identifies it verbally, the gaze fixations around the onset of speech are related to the user's attention to the object named. The particular fixations most pertinent to the user's attention vary with the complexity of the task. For the simple task of finding and speaking an isolated word (Experiment 1), the combination of two independent features, ***fa1*** the first fixation after the onset of speech, and ***fii***, the longest fixation around speech onset (speech start time +/- 1500msec), **excluding** ***fa1***, had the greatest ability to predict of the user's attention correctly across a wide variety of subjects. On the other hand, for the more complex menu selection task (Experiment 2), the combination of three independent features, ***fb1,*** the last fixation before the onset of speech, ***fa1,*** and ***fa2,*** the second fixation after the onset of speech, had the greatest ability to predict of the user's attention correctly across a wide variety of subjects.

- Since the experiments have shown that gaze behavior differs from user to user and can vary for an individual user over time, an adaptive technique has been developed for adjusting gaze tracking parameters to provide individualized and efficient gaze performance. The algorithm employs an iterative technique called Row Action Projection (RAP), which has improved target detection performance over non-adaptive techniques.
- Five techniques for conveying the user's intent to a computer system using some combination of speech and gaze were analyzed. They were:
  - **Speech Alone**
  - **Dominant Gaze Fixation**, the particular fixation that best predicts the user's attention to a target on a display. This feature was shown to vary from subject to subject and vary for a particular subject over time. Although it cannot be used by itself in a practical system, it provides a spectrum of features from which a useful subset can be extracted.
  - **Linear Prediction (LPu),** use of a standard linear prediction algorithm on gaze data to assess the user's intent. The analysis showed that a linear, time-varying system, **LPu,** performed better than a robust fit multi-linear regression model with 10 different weighting functions. Consequently, the linear system model was adopted for further analysis.
  - **Adaptive Prediction (RAP)**, use of an adaptive linear prediction algorithm on gaze data to assess the user's intent. The RAP technique uses the basic concepts

of a linear system, but it is nonlinear in nature in that it converges to the solution for a system of equations.

- **Combined Speech and/or RAP**, use of speech and adaptive gaze data to assess the user's intent.

In both experiments, using target detection probability as a measure of success, performance of **LPu**, **RAP**, and **Combined Speech/RAP** probabilities were computed under two conditions; internal gaze and RAP parameters fixed across all subjects and those same parameters optimized for each subject, individually. The various techniques were compared using several different measures. Among them, a paired sample t-test showed that, for both fixed and individually optimized conditions, the **Combined Speech/RAP** technique was **significantly** better than speech alone for both experiments. This result held true for the entire populations of subjects and for various combinations of subsets of native and non-native speakers. As a secondary result, **RAP** was **significantly** better than **LPu** for individually optimized parameters in experiment 1 and **significantly** better than **LPu** for individually optimized parameters for all subjects in experiment 2, but not for the various sub-populations, which had very small populations. Further testing with larger populations is required here.

- Among the other approaches developed for comparing techniques, interface usability curves (the percentage of time the target detection probability is higher than an acceptable threshold) were plotted for the five techniques enumerated above. These are somewhat like Receiver Operating Characteristic (ROC) curves. These curves corroborated the t-test results, and, more importantly, illustrated two important results of those tests. First, under extremely noisy conditions when speech recognition performance is unacceptable, **Combined Speech/RAP provides very large improvements in performance.** Second, when target detection probability is border-line acceptable, e.g. 90-95% (achieved with native-only speakers in our experiments), **Combined**

Speech/RAP improved performance to the 95-100% range. This can make a significant difference in some applications.

- The effect of spacing and font size on target detection probability in the menu selection experiment was evaluated using the Kruskal-Wallis test, a non-parametric test that rank-orders the variables under test. The test showed that for the three different target spacings tested in Experiment 2 (10, 20 & 30 pixels, edge-to-edge), there was a **significant improvement in performance with increased spacing.** On the other hand, for the two font sizes tested in Experiment 2 (12 point, 20 point), **no significant relationship between font size and performance was found.**

Although there is research in the literature stating that the duration of a fixation on a target is a powerful indicator of a user's attention to that target, our experiments, which were, admittedly, quite different from the experiments leading to that conclusion showed that that there are other fixations that make more significant contributions to the prediction of the users attention in a speech/gaze system, i.e., those occurring in the immediate neighborhood of the speech start time. What was found, however, was that the distribution of the time of occurrence of the longest duration fixations, relative to speech start time was much broader for the more complex menu selection task than the simple, isolated word task, especially prior to speech start time. This suggests that **fixation duration might provide useful information about task complexity**.

9. Conclusions and Future Work

This dissertation's key purpose was to investigate the improved efficiency of using speech and gaze together to predict a user's menu selection from a large of screen displayed options. It was found that the combination was better than either modality acting alone. It was also found that to be successful, the system has to be adaptive from user to user and over time for each individual user. It was also found that the best improvement occurred in high noise environments and with individuals having heavily accented speech, but even in nearly perfect environments with high speech recognition performance, gaze provided a performance improvement. The differences in predictive parameters derived from the two experiments also demonstrated that task complexity has to be involved in the design of a speech/gaze interface.

As with any research, more questions end up unanswered at its conclusion. This dissertation thus suggests a set of future studies that are needed to investigate further the feasibility of building a speech/gaze system. These are:

- The system built was not run in real time. There is a significant amount of calculation being done to perform the running adaptation. A higher powered computer and a real time system needs to be set up.
- The study involved only five native English speakers which is too small. Thus, the conclusion of a significant effect from adding gaze to improve system performance is suspect. Thus, more subjects in the categories of native and non-native speakers need to be run.
- The study used an earlier generation speech recognition system. Better systems are available which have higher recognition rates and also adapt to variations in speech better. This work, thus, needs to be redone with a better speech recognition engine.

- Each experiment with a subject involved only one session. It is not clear what will happen over time. Perhaps user's will memorize the menu items and not look at them or perhaps they will change their gaze patterns. In addition, other distractions which might have affected gaze patterns were kept to a minimum. Thus, these studies need to be rerun in a more natural environment.
- To use the particular eye tracker involved in the experiments, it was necessary to spend a few seconds to calibrate the eye tracker both before and at periodic intervals throughout the experiment. This is not a natural situation for a user. What needs to be done is the development of an automatic calibration mechanism that adapts to the user.
- Although the RAP algorithm worked relatively well, it was clear that unexplained situations could readily throw off its prediction. These situations need to be explored further.
- Although the RAP algorithm worked relatively well, it is suggested that the development of this algorithm be explored further leading to possible further improvements in performance.
- Currently the adaptation model requires the task based features to be fed manually to the model. This needs to be enhanced to automatically select the appropriate gaze feature combination required, based on the task complexity
- Font-size has not been shown to have a significant impact on the speech/gaze interactions. Further experiments need to be carried out to fully investigate any font-size effect since other literature suggests its existence.
- Only font size, spacing, and location are studied for the analysis of the speech/gaze integration model for tasks with constraints (*i.e.*, multiple targets). Several other factors could also be explored to enhance the integration model, e.g., size of word, font style, etc.
- It is not clear how the fixation algorithm affects the speech/gaze integration model. Only

a dispersion-based fixation algorithm has been used in the analysis. Other fixation algorithms like velocity and area fixation algorithms are left for future study

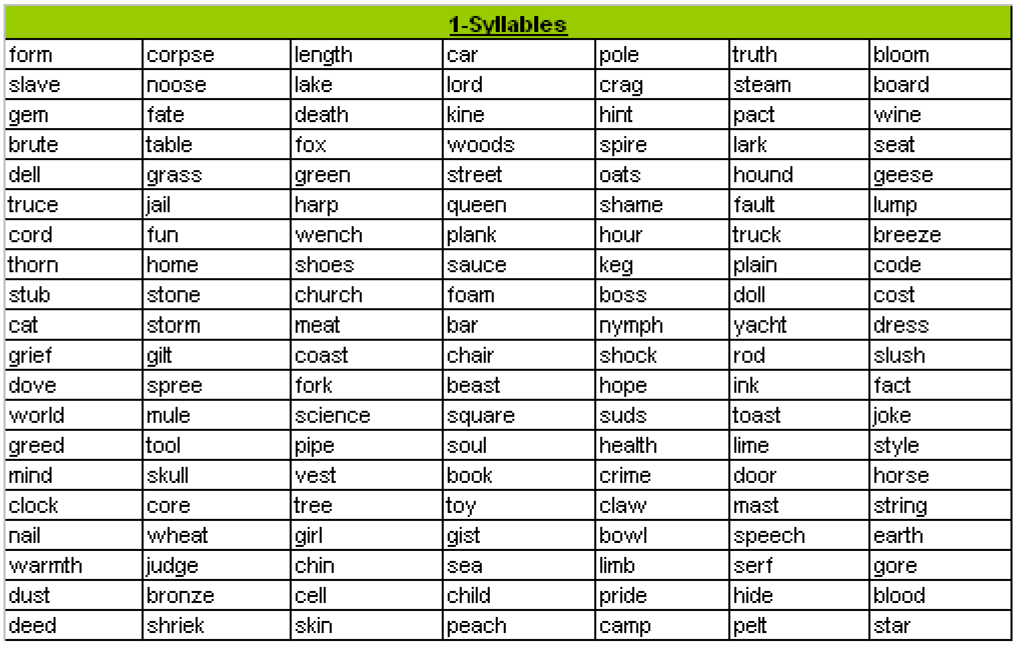
Although the above suggestions for future work will enhance the work of this dissertation, it stands alone as a foundation for developing adaptive speech/gaze integration systems. As gaze tracking technology improves, this work suggests that its inclusion in hands-free gaze/speech interfaces is a definite possibility.

Appendix A. Review of Multimodal Systems

System/Modality	Approach	Speech	Gaze	Touch	Pen	Mouse	Keyboard	Gestures	Sketch	Tactile	Joy Stick	Helmet Visor	Applications	Fusion
Spatial Data Management System (SDMS) [15]	First Multimodal Application	Y	N	N	N	Y	Y	Y	N	N	N	N	Simple Objects on Caribbean Map	Speech As Clutching Mechanism [27]
ICARE [9]	Component Based	Y	N	N	N	Y	Y	N	N	Y	Y	Y	MID, MEMO, FACET	CARE Properties
Multimodal Cell Phone Architecture [13]	Server Side	Y	N	N	Y	Y	Y	N	N	N	N	N	Smart Phone Sony Ericsson P900	Data Manager Synchronization
Disciple Framework [5]	Framework Based	Y	Y	Y	Y	Y	Y	N	N	N	N	N	Flatscape	Command Frame Construction from Parse Tree
VR UI Framework [138]	Framework Based	Y	N	N	N	N	N	Y	N	Y	N	N	Immersive Visualization	tATN Temporal Search
Galaxy Communicator Architecture [141]	Distributed Hub-Spoke, Message Based	Y	N/A	N/A	N/A	Y	Y	N/A	N/A	N/A	N/A	N/A	Spoken Dialogue Systems	Application Dependent
Open Agent Architecture [18]	Distributed Agent Based	Y	N/A	N/A	Y	Y	Y	Y	N/A	Y	N/A	N/A	Distributed Applications	Application Dependent
VIENA System [36]	Timed Agent Based	Y	N	N	N	Y	Y	Y	N	N	N	N	Sample Office Space Application	Rhythm Based Segmentation To Fuse Input
SmartKom [45]	Framework Based	Y	N	Y	Y	Y	Y	Y	Y	N	N	N	PDAs, Public Information System, Desktop Applications	Temporal Hypotheses Merging
Multiplatform [23]	Open Component Architecture	Y	N/A	N/A	N/A	Y	Y	Y	N/A	N/A	N/A	N/A	VERBMOBIL Application	Application Dependent
JEANIE [28]	Framework Based	Y	N	N	Y	Y	Y	N	Y	N	N	N	Calendar Application	Semantic Frame Merging
SEER [47] [65]	Layered Hidden Markov Models (LHMMs)	Y	N	N	N	Y	Y	Y	N	N	N	N	Office Activity Application	LHMMs
PATE/COMIC System [38]	Multi-Blackboard Architecture	Y	N	N	N	Y	Y	Y	N	N	N	N	Bathroom Design Application	Context Based Fusion

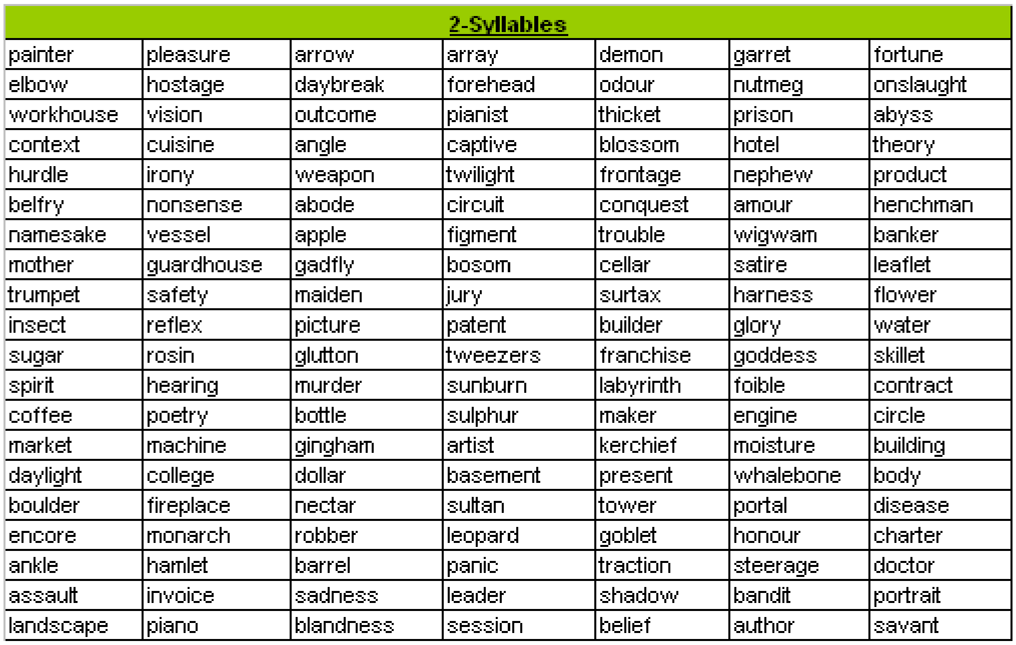
System/Modality	Approach	Speech	Gaze	Touch	Pen	Mouse	Keyboard	Gestures	Sketch	Tactile	Joy Stick	Helmet Visor	Applications	Fusion
EVI3d [31]	Distributed Data Flow Architecture	Y	N	Y	N	Y	Y	Y	N	N	N	N	Virtual Environment Applications	Rule Based Temporal Analysis
Smart Web [140]	Distributed	Y	N	Y	Y	N	Y	N	N	N	Y	N	PDA/Web Applications (Demo’ed in FIFA 2006)	FADE Component w/ temporal analysis
ThreadMill Architecture [22]	Distributed Component Based	Y	N	N	N	Y	Y	Y	N	N	N	N	Sign Language System	Distributed Message Passing Coordination Model
PAC-Amodeus [139]	Agent Based	Y	N	N	N	Y	Y	Y	N	N	N	N	MATIS (Airline Information System)	Rule Based Temporal and Contextual Fusion
QuickSet [12] [33]	Agent (OAA) Based	Y	N	N	Y	Y	Y	Y	Y	N	N	N	Field Medic System, Voice Assistant	Late Symbolic / Statistical Unification Based Fusion
ICO [142]	Model-based	Y	N	Y	N	Y	Y	Y	N	N	Y	Y	Rafale Aircraft System	Temporal Analysis
Georgia Tech Gesture Toolkit [143]	High-level Abstraction on HTK for Gesture Recognition	Y	Y	N	N	Y	Y	Y	N	N	N	N	Gesture Panel, Prescott, Telesign, Workshop Activity Recognition	Application Dependent
MacVisSTA [40]	Modality Visualization Framework	Y	Y	N	N	Y	Y	Y	N	N	N	N	Music-Score Application	Temporal Analysis
IRYS [37]	Virtual Network Computers	N	Y	Y	N	N	N	N	N	N	N	N	Visualization Tool for Online Applications	Application Dependent
VESS [126]	VR Library	N	N	Y	N	N	N	Y	N	Y	N	Y	VR Applications	Application Dependent
EMBASSI Architecture [50]	Message Driven Pipelined Comm. Flow	Y	N	N	N	Y	Y	Y	N	N	N	N	PDA applications and others	PMI Fusion Module

Appendix B. Word List in Experiment 1



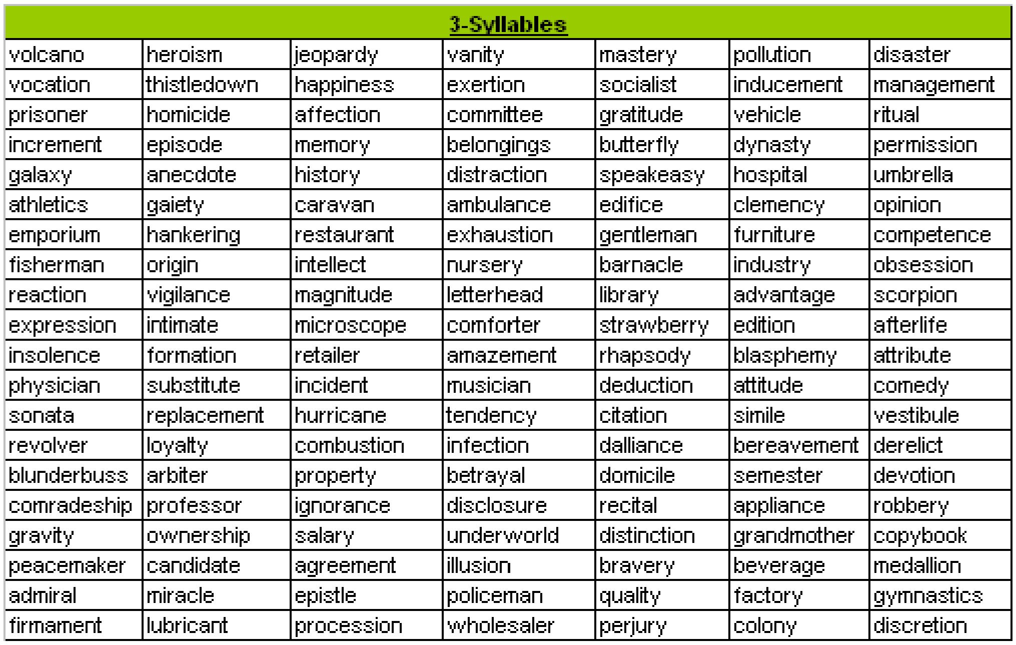
1-Syllables						
form	corpse	length	car	pole	truth	bloom
slave	noose	lake	lord	crag	steam	board
gem	fate	death	kine	hint	pact	wine
brute	table	fox	woods	spire	lark	seat
dell	grass	green	street	oats	hound	geese
truce	jail	harp	queen	shame	fault	lump
cord	fun	wench	plank	hour	truck	breeze
thorn	home	shoes	sauce	keg	plain	code
stub	stone	church	foam	boss	doll	cost
cat	storm	meat	bar	nymph	yacht	dress
grief	gilt	coast	chair	shock	rod	slush
dove	spree	fork	beast	hope	ink	fact
world	mule	science	square	suds	toast	joke
greed	tool	pipe	soul	health	lime	style
mind	skull	vest	book	crime	door	horse
clock	core	tree	toy	claw	mast	string
nail	wheat	girl	gist	bowl	speech	earth
warmth	judge	chin	sea	limb	serf	gore
dust	bronze	cell	child	pride	hide	blood
deed	shriek	skin	peach	camp	pelt	star

Table 15 One syllable words



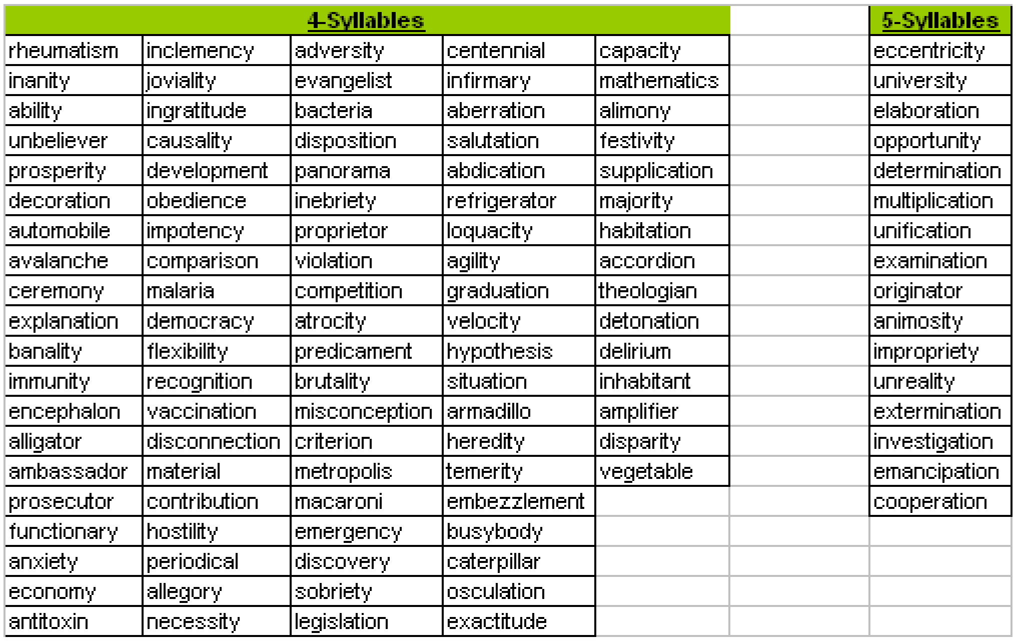
2-Syllables						
painter	pleasure	arrow	array	demon	garret	fortune
elbow	hostage	daybreak	forehead	odour	nutmeg	onslaught
workhouse	vision	outcome	pianist	thicket	prison	abyss
context	cuisine	angle	captive	blossom	hotel	theory
hurdle	irony	weapon	twilight	frontage	nephew	product
belfry	nonsense	abode	circuit	conquest	amour	henchman
namesake	vessel	apple	figment	trouble	wigwam	banker
mother	guardhouse	gadfly	bosom	cellar	satire	leaflet
trumpet	safety	maiden	jury	surtax	harness	flower
insect	reflex	picture	patent	builder	glory	water
sugar	rosin	glutton	tweezers	franchise	goddess	skillet
spirit	hearing	murder	sunburn	labyrinth	foible	contract
coffee	poetry	bottle	sulphur	maker	engine	circle
market	machine	gingham	artist	kerchief	moisture	building
daylight	college	dollar	basement	present	whalebone	body
boulder	fireplace	nectar	sultan	tower	portal	disease
encore	monarch	robber	leopard	goblet	honour	charter
ankle	hamlet	barrel	panic	traction	steerage	doctor
assault	invoice	sadness	leader	shadow	bandit	portrait
landscape	piano	blandness	session	belief	author	savant

Table 16 Two syllable words



3-Syllables						
volcano	heroism	jeopardy	vanity	mastery	pollution	disaster
vocation	thistledown	happiness	exertion	socialist	inducement	management
prisoner	homicide	affection	committee	gratitude	vehicle	ritual
increment	episode	memory	belongings	butterfly	dynasty	permission
galaxy	anecdote	history	distraction	speakeasy	hospital	umbrella
athletics	gaiety	caravan	ambulance	edifice	clemency	opinion
emporium	hankering	restaurant	exhaustion	gentleman	furniture	competence
fisherman	origin	intellect	nursery	barnacle	industry	obsession
reaction	vigilance	magnitude	letterhead	library	advantage	scorpion
expression	intimate	microscope	comforter	strawberry	edition	afterlife
insolence	formation	retailer	amazement	rhapsody	blasphemy	attribute
physician	substitute	incident	musician	deduction	attitude	comedy
sonata	replacement	hurricane	tendency	citation	simile	vestibule
revolver	loyalty	combustion	infection	dalliance	bereavement	derelict
blunderbuss	arbiter	property	betrayal	domicile	semester	devotion
comradeship	professor	ignorance	disclosure	recital	appliance	robbery
gravity	ownership	salary	underworld	distinction	grandmother	copybook
peacemaker	candidate	agreement	illusion	bravery	beverage	medallion
admiral	miracle	epistle	policeman	quality	factory	gymnastics
firmament	lubricant	procession	wholesaler	perjury	colony	discretion

Table 17 Three syllable words



4-Syllables						5-Syllables
rheumatism	inclemency	adversity	centennial	capacity		eccentricity
inanity	joviality	evangelist	infirmary	mathematics		university
ability	ingratitude	bacteria	aberration	alimony		elaboration
unbeliever	causality	disposition	salutation	festivity		opportunity
prosperity	development	panorama	abdication	supplication		determination
decoration	obedience	inebriety	refrigerator	majority		multiplication
automobile	impotency	proprietor	loquacity	habitation		unification
avalanche	comparison	violation	agility	accordion		examination
ceremony	malaria	competition	graduation	theologian		originator
explanation	democracy	atrocity	velocity	detonation		animosity
banality	flexibility	predicament	hypothesis	delirium		impropriety
immunity	recognition	brutality	situation	inhabitant		unreality
encephalon	vaccination	misconception	armadillo	amplifier		extermination
alligator	disconnection	criterion	heredity	disparity		investigation
ambassador	material	metropolis	temerity	vegetable		emancipation
prosecutor	contribution	macaroni	embezzlement			cooperation
functionary	hostility	emergency	busybody			
anxiety	periodical	discovery	caterpillar			
economy	allegory	sobriety	osculation			
antitoxin	necessity	legislation	exactitude			

Table 18 Four and Five syllable words

Appendix C. Word List in Experiment 2

The following list shows all the words used in all sessions in Experiment 2. It lists out the commands that are in inner 4x4 array in bold. The non-bold words are in the edge of the 6x6 array.

Session 1: {**"abode", "dawn", "earth", "geese", "hall", "keg", "lad", "maker", "queen", "table", "water", "ink", "nail", "jail", "oats", "venom",** "river", "yacht", "cabin", "salad", "pact", "fact", "baby", "charm", "skin", "plank", "folly", "book", "code", "spree", "flag", "cigar", "bird", "power", "snake", "pelt"}

Session 2: {**"abyss", "death", "ego", "gem", "harp", "kine", "lake", "mast", "quest", "tank", "wench", "inn", "noose", "jelly", "ocean", "vest",** "river", "yacht", "camp", "sauce", "panic", "fate", "bar", "chasm", "skull", "plank", "fork", "bosom", "coin", "stain", "flask", "city", "blood", "pride", "soil", "pep"}

Session 3: {**"adage", "deed", "elbow", "ghost", "hide", "king", "lark", "meat", "river", "thief", "whale", "iron", "nun", "joke", "odour", "unit",** "quest", "yacht", "candy", "sea", "paper", "fault", "bard", "chief", "sky", "plank", "form", "boss", "cord", "star", "flesh", "claw", "bloom", "pride", "soul", "piano"}

Session 4: {**"agony", "dell", "event", "gift", "hint", "kiss", "law", "mercy", "rock", "thorn", "wheat", "irony", "nymph", "joy", "opium", "unit",** "quest", "yacht", "cane", "seat", "party", "fiord", "baron", "child", "slave", "plank", "fowl", "bowl", "core", "steam", "flood", "clock", "board", "pride", "spire", "pipe"}

Session 5: {**"air", "demon", "idea", "gilt", "home", "nail", "lawn", "metal", "rod", "time", "wife", "earth", "keg", "judge", "oven", "unit",** "quest", "yacht", "car", "serf", "peach", "fire", "beast", "chin", "slush", "plank", "fox", "boy", "corn", "stone", "foam", "coast", "body", "pride", "spray", "plain"}

Session 6: {**"amour", "devil", "idiom", "girl", "hoof", "noose", "lemon", "mind", "rosin", "toast", "wine", "ego", "keg", "jury", "owner", "unit",** "quest", "yacht", "cash", "shame", "pelt", "flag", "bird", "cigar", "snake", "plank", "frog", "brain", "cost", "storm", "fact", "charm", "pole", "pride", "skin", "pact"}

Session 7: {**"anger", "dirt", "ink", "gist", "hope", "nun", "lice", "money", "unit", "tomb", "woman", "elbow", "keg", "jury", "river", "oats",** "quest", "yacht", "cat", "ship", "pep", "flask", "blood", "city", "soil", "plank", "fun", "brute", "crag", "stub", "fate", "chasm", "pole", "pride", "skull", "panic"}

Session 8: {**"angle", "doll", "inn", "glory", "horse", "nymph", "life", "monk", "venom", "tool", "woods", "event", "kine", "jury", "rock", "unit",** "quest", "yacht", "cell", "shock", "piano", "flesh", "bloom", "claw", "soul", "plank", "fur", "brute", "crime", "style", "fault", "chief", "pole", "pride", "sky", "paper"}

Appendix D. System Description

To understand the speech/gaze integration mode empirically, the experimental setup in Figure 37 is (i.e., speech/gaze interface) installed on two computers, one for speech and the other for gaze. The experimental setup (*i.e.*, hardware and software) is exactly identical in both the experiments. An IBM ViaVoice speech recognizer version 8.1 and an ISCAN eye tracker are used for the human factors experiments. User sits in front of the Speech Machine and performs the experiment. The system uses the Java Speech API (JSAPI) to connect to the IBM ViaVoice recognizer / synthesizer. The ISCAN interface (Figure 38) connects to the eye tracker and provides gaze input to the experimental application over a serial port. Speech input is recognized by the *recognizer,* and the *synthesizer* produces the speech output for the speech/gaze interface. Note that the experiments do not use the synthesizer and loud speakers. They are available for use in specific applications where synthesized speech response is required. The speech/gaze interface issues commands via a serial port to the gaze machine to invoke automatic gaze calibration. The eye tracker, upon receiving the commands from the speech machine over the serial port, performs gaze calibration for the subject automatically. Subsequently, the gaze machine provides the *point of regard* (POR) output, i.e., the location on the display where the eye is focused, back to the speech machine over the serial port. The serial port is used for duplex communication between the two machines to process gaze commands from speech machine to gaze machine and to send gaze output from gaze machine to speech machine. The gaze machine produces $<x, y, d>$ tuples to the speech machine where (x, y) represent the gaze location of the subject's eye on the speech machine display and d is the pupil diameter. The speech machine sends commands to the gaze machine to control the eye tracker for calibration, start/stop gaze recording, and to control the camera movements.

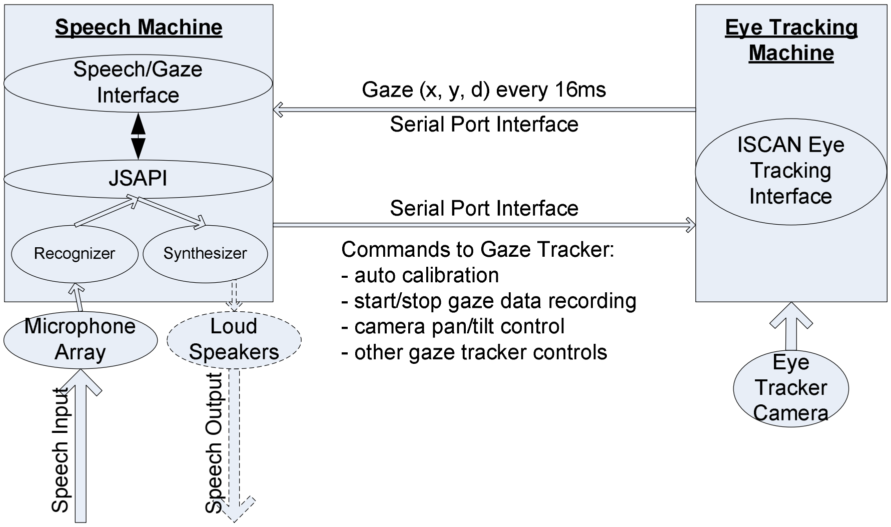


Figure 37 System Installation for Speech/Gaze Interaction Experiments

Figure 38 shows the interface of the ISCAN eye tracker which consists of various graphical resources to control the operation of the eye tracker. These controls handle *eye tracking, POR / Calibration*, *Camera Movement,* and *Scene/Eye Monitor*. The *eye tracking controls* handle the display of cross-hairs, corneal/pupil reflection thresholds and image gate (i.e., the white rectangular border in the “EYE MONITOR – EXPANDED VIEW”). The *POR / Calibration controls* manage the point of regard and calibration procedures. The *PAN / TILT controls* manage the physical camera movements. The two small rectangular areas on the right side of the screen show the scene and eye video streams. These streams can be viewed in an expanded mode as shown in the center of the screen as illustrated by the “EYE MONITOR – EXPANDED VIEW”. Also, there is an “options” control below “EYE MONITOR – EXPANDED VIEW” which manages the gaze data recording. The eye tracker is active when the “Track Active” is checked (above PAN/TILT controls) and continuously provides the eye coordinate information on the object plane. The eye tracker is configured to provide the <x, y, d> tuples to the speech machine over a serial port.

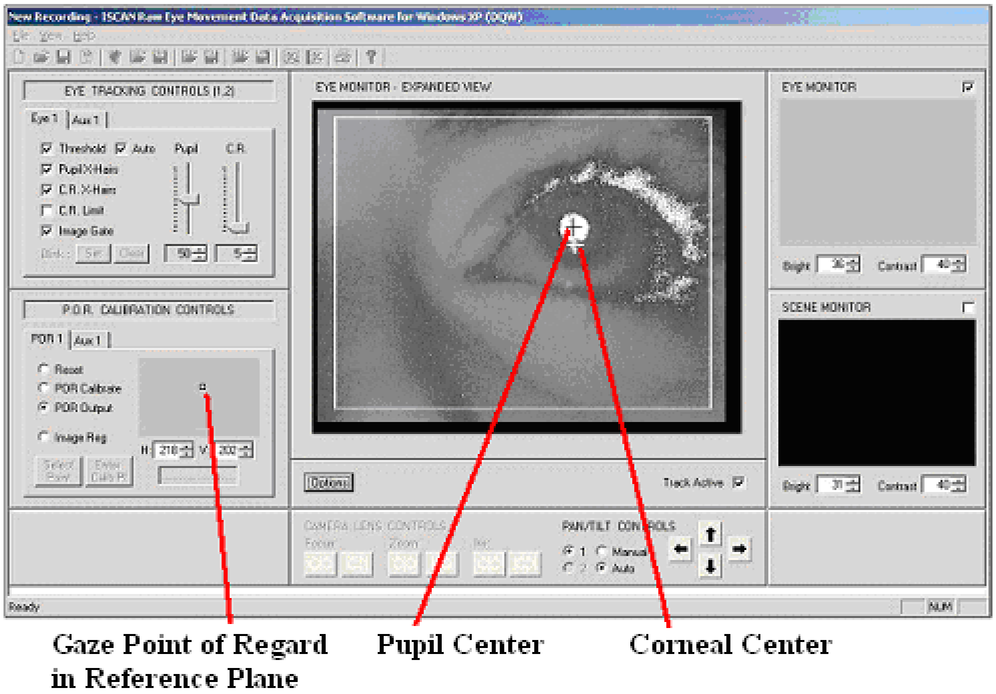


Figure 38 ISCAN Interface

The speech/gaze experiments are programmed in Java and run on the speech machine. Several design choices have to be made in light of carrying out the experiments to collect the data consistently. The system setup is installed on two machines because both the speech recognition and eye tracking are computationally intensive tasks which may compete for the system resources simultaneously. Installing the experimental setup on two machines ensures both modalities running freely on independent machines providing their output. The IBM Via Voice recognizer is less expensive and serves the purpose of simulating problems with real-world speech understanding. The ISCAN eye tracker is chosen because it samples the eye movements at field rates *i.e.*, 60Hz which is representative of an eye tracker that can be used in a practical real-time application. The eye tracker chosen does not include any head mounted device and mimics a real natural human machine interaction.

The current system on two computers often poses challenges in synchronizing clocks between them. Monitoring and correcting for the differences between the clocks is crucial to the data analysis. Section E.5 describes how the data is validated with respect to clock timing differences across two machines.

Apart from system limitations, there are several factors contributing to the complexity of speech and gaze human machine interactions. Although speech recognizers are highly advanced in recent years, recognizers are still having problems in real-world applications, and often require extensive training information to be able to accurately recognize all nuances of various accents. Gaze naturally is highly unpredictable in its nature and is very difficult to track accurately. Moreover, the calibration may not hold long and start producing errors in gaze tracking. In addition to the system and modality limitations, the speech/gaze interaction patterns are highly unpredictable and often yield dynamic and random interaction patterns. All of these factors are carefully considered and appropriate precautionary measures are taken to validate the data capture process which is explained in great detail later in this chapter.

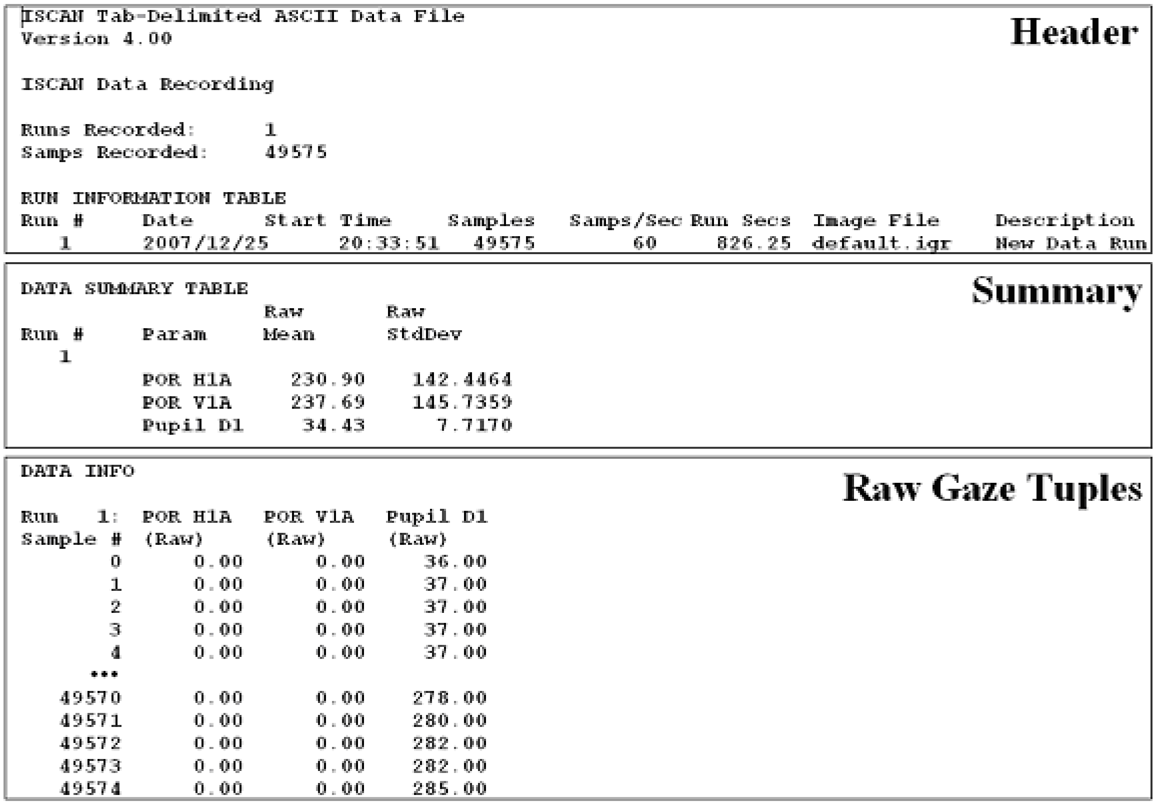
Appendix E. Data Capture and Validation

In this section, the data capture and the validation process is described for the data recorded on the speech and gaze machines. It illustrates an example of the raw gaze data file recorded on the gaze machine and how the same data is obtained on the speech machine over a serial port connection. It also validates that the two machines' setup is not impacting the data analysis in any manner and provides the measures taken to ensure the validity of the data capture process. This section also describes the post processing involved in preparing the data to be suitable to be used by predictive / adaptive models. First, the data collection process is described here because it is complex and so connected to the differences for each of the experiments. Then, the data capture/validation relevant to each experiment is described.

E.1. Raw Gaze Data

Recording scanpaths properly is an important task in obtaining the gaze data consistently. The eye tracker used in both the experiments obtains the data as a sequence of tuples <*n,x,y,d*> (in a file <*subject.tda*>), where *n* is the gaze sample number recorded by the eye tracker, *(x, y)* is the point of regard and *d* is the pupil diameter. A header section followed by the summary information is captured for each run of the experiment. The eye tracker records the raw gaze samples' start date and time and tags them with a sample number *i.e.*, *n*. This gaze data captured on the gaze machine is identical in both experiments.

Figure 39 shows the raw gaze data recorded by the ISCAN eye tracker. It contains three sections *Header, Summary,* and *Raw Gaze Tuples*. The first 4 lines of the *Header* section contain the ISCAN's logo information. After that, it shows the number of runs and the total number of gaze samples recorded. Each *run* is a recording session in the ISCAN system. *Run* information includes the *Run#, Date, StartTime, Samples, Samples/Sec, RunSecs, ImageFile,* and *Description*. The *RunSecs* is the total time span of the recording session. *ImageFile* and *Description* are not used. The *Summary* section shows the statistics of the gaze data for the three parameters *POR H1A* (*i.e.*, x-coordinate), *POR V1A* (*i.e.*, y-coordinate), and *Pupil D1* (*i.e.*, pupil diameter d). The gaze data in section *Raw Gaze Tuples* is preprocessed to create tuples <seqno, timestamp, x, y, d> where the *seqno* is a 0 based index of the gaze sample as generated by the eye tracker. The *seqno* and start time in *Header* are used to compute the *timestamp* of each gaze sample in the gaze data. The *(x, y, d)* indicate x-coordinate, y-coordinate, and pupil diameter.



```
ISCAN Tab-Delimited ASCII Data File
Version 4.00

ISCAN Data Recording

Runs Recorded:      1
Samps Recorded:     49575

RUN INFORMATION TABLE
Run #     Date        Start Time     Samples   Samps/Sec Run Secs  Image File     Description
   1      2007/12/25     20:33:51   49575          60     826.25  default.igr    New Data Run
```

```
DATA SUMMARY TABLE
                    Raw        Raw
Run #     Param     Mean       StdDev
   1
          POR H1A     230.90    142.4464
          POR V1A     237.69    145.7359
          Pupil D1     34.43      7.7170
```

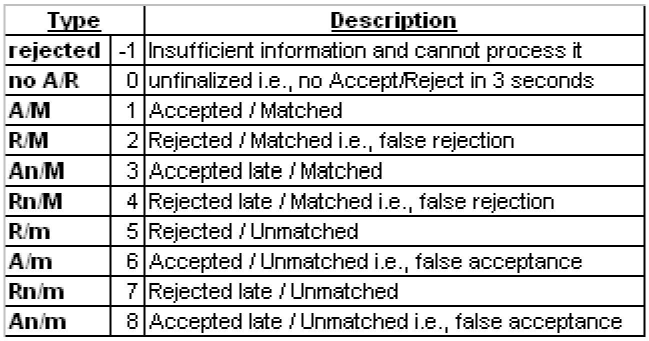
```
DATA INFO

Run    1:  POR H1A   POR V1A   Pupil D1
Sample #  (Raw)     (Raw)     (Raw)
        0      0.00      0.00     36.00
        1      0.00      0.00     37.00
        2      0.00      0.00     37.00
        3      0.00      0.00     37.00
        4      0.00      0.00     37.00
      •••
    49570      0.00      0.00    278.00
    49571      0.00      0.00    280.00
    49572      0.00      0.00    282.00
    49573      0.00      0.00    282.00
    49574      0.00      0.00    285.00
```

Figure 39 Raw Gaze Data

E.2. Speech Machine Data

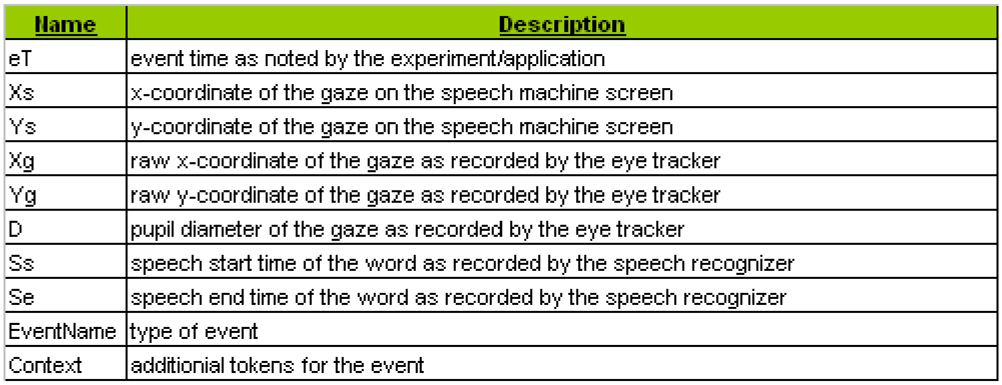
An important factor to be considered in a multimodal (speech/gaze) interaction is the speech recognition accuracy. Regardless of the performance of the speech recognizer, speech recognition can be error prone due to ambient conditions and variations in user pronunciation from time to time. A word uttered by the subject may not always be recognized properly. For any utterance, the current IBM Via Voice speech recognizer issues a *speech-start*, *speech-end*, and *speech-accept/speech-reject* events in that order. It either issues a *speech-accept* or *speech-reject* event but not both. These three events indicate whether an utterance is processed by the recognizer correctly or not. Speech *accept* and *reject* events are indicated by **A** and **R** in Table 19. An accept event means that the audio signal corresponds to and finds a word from among the set of words in the grammar with maximum probability. When the recognizer issues a reject event, it may still produce a word for the audio signal but it only means that the confidence level on the word is below the threshold level set in the recognizer configuration. After the recognizer produces a word corresponding to the audio signal, the word doesn't necessarily have to match what is displayed on the screen. If the word shown on the screen is same as the word the recognizer thinks the audio signal corresponds to, then it is indicated by **M** (*i.e.*, a match). If it is a mismatch then it is indicated by **m**. Since there is a timeout in the experiments (3 seconds in Experiment 1 and 15 seconds in Experiment 2) for each trial, the recognition results of the current trial can come after the next trial starts. This can happen because the subject may speak the word just about when the timeout happens. Hence, the result from the recognizer is recorded during the next trial. Note that the timeout depends on the task complexity. An '**n'** is added to **A** or **R** indicating that it is during the next trial when the result is obtained. Table 19 illustrates various scenarios that exist in the speech/gaze correlation experiments. If the recognizer produces accept or reject event, then the audio signal is considered to be finalized or it is considered to be un-finalized from the recognizer point of view. Few trials are **rejected** because the information is not sufficient in those trials to determine whether they fall into any one of the acceptable categories defined below. These categories are used in analyzing the data captured from Experiment 1 and 2 in the next couple of sections.



Type		Description
rejected	-1	Insufficient information and cannot process it
no A/R	0	unfinalized i.e., no Accept/Reject in 3 seconds
A/M	1	Accepted / Matched
R/M	2	Rejected / Matched i.e., false rejection
An/M	3	Accepted late / Matched
Rn/M	4	Rejected late / Matched i.e., false rejection
R/m	5	Rejected / Unmatched
A/m	6	Accepted / Unmatched i.e., false acceptance
Rn/m	7	Rejected late / Unmatched
An/m	8	Accepted late / Unmatched i.e., false acceptance

Table 19 Speech Recognition Categories

The speech machine records the data (in a file <*subject.txt*>) from both the eye tracker and the speech recognizer as a set of events when subjects are taking part in the experiment. This data recording on the speech machine is exactly identical in both experiments. Each <*subject.txt*> file contains all trials of the experiment for that subject. Each trial in the data sequence (in <*subject.txt*>) obtained on the speech machine can be illustrated as 4 event segments: *Marker Display / Dismissal* segment, *Word Display / Speech Start* segment, *Speech Start / Stop* segment, and *Recognizer Finalization* segment. Conceptually, these data segments are applicable to both experiments. Each trial in Experiment 1 or Experiment 2 collects all these data segments. Each row in these data segments represents an *event record*. Each event record consists of 10 columns as described in Table 20. Examples of event records are illustrated later in Table 23 through Table 26 for Experiment 1and in Table 29 through Table 32 for Experiment 2.



Name	Description
eT	event time as noted by the experiment/application
Xs	x-coordinate of the gaze on the speech machine screen
Ys	y-coordinate of the gaze on the speech machine screen
Xg	raw x-coordinate of the gaze as recorded by the eye tracker
Yg	raw y-coordinate of the gaze as recorded by the eye tracker
D	pupil diameter of the gaze as recorded by the eye tracker
Ss	speech start time of the word as recorded by the speech recognizer
Se	speech end time of the word as recorded by the speech recognizer
EventName	type of event
Context	additionial tokens for the event

Table 20 Speech Machine Event Tuple

The *eT* values are timestamps in milliseconds for events recorded by the speech machine.

These values are all consistently positive and monotonically increasing in every trial in both experiments. For all events of the type *GazeCaptured*, (*Xg*, *Yg*) is the gaze location in eye tracker's reference plane (512x512) which is converted as (*Xs*, *Ys*) onto speech machine's display coordinate system. Sometimes (*Xg*, *Yg*) is recorded as (0, 0) because the eye tracker may not be able to track the subject's eye due to subject's movements, blinks *etc.* (*Xs*, *Ys*) will be recorded as (0, 0) whenever (*Xg*, *Yg*) is (0, 0). Whenever the event is not *GazeCaptured*, the (*Xs*, *Ys*) and (*Xg*, *Yg*) are identical in all data tables. For example, *GazeIn* indicates an internal experiment event denoting that the gaze is inside the target object and *GazeFire* indicates that the application event has been generated. The values in column D indicate the pupil diameter and when the data is not recordable or applicable it is recorded as -999 (only to indicate it's not a good recorded value). The *Ss* and *Se* columns are present to extract the speech recognizer's start and end timestamps of utterances. However, these can be recorded only when the recognizer issues the utterance recognition results. Most of the time, these values are recorded as –1 indicating invalid values and can't be interpreted. However, in Experiment 1 and 2 these columns are overloaded to record additional data for keeping the data structures consistent. There are several event types that are tracked in both the experiments. *CorrelationMarkerShow* is an event indicating that the marker is being displayed and *CorrelationAppOnCommand* indicates that the experiment has taken an action on the event generated by the system. Several events like CorrelationWordShow, CorrelationWordCenter, SpeechStarted, SpeechStopped, SpeechTag, SpeechToken, SpeechAdjustedToken, SpeechAccepted, SpeechRejected *etc.* are recorded by both experiments. The *Context* column contains additional information recorded by experiments which can be used in data pre-processing. For example, if the context of *CorrelationWordShow* and context of *CorrelationAppOnCommand* are equal, then it means that that word uttered by the subject has been recognized accurately by the speech recognizer. All these events mean the same and their processing is identical in both the experiments. In the next two sections two scanpaths, one for each experiment, are described in detail to illustrate the data segments pertinent to the respective experiments. The data segments of each trial in either Experiment 1 or 2 illustrate the detailed nature of the user behavior in terms of the events happening in the interaction for respective tasks. This helps in understanding the integration model parameters involved in the speech/gaze interaction process.

E.3. Experiment 1 Data Capture

Each subject runs the experiment to produce N number of scanpaths/interaction-samples. The subject runs the experiment for as long as there is no discomfort and hence the samples collected varied among the subjects in the constraint-free experiment. Experiment 1 data can be used to illustrate the fatigue levels that can be tolerated by various people in using the interface. But there is no clear relation that can be established between the number of samples/scanpaths and the user fatigue levels. However, not all the samples can be utilized in the analysis because not all samples/scanpaths are usable.

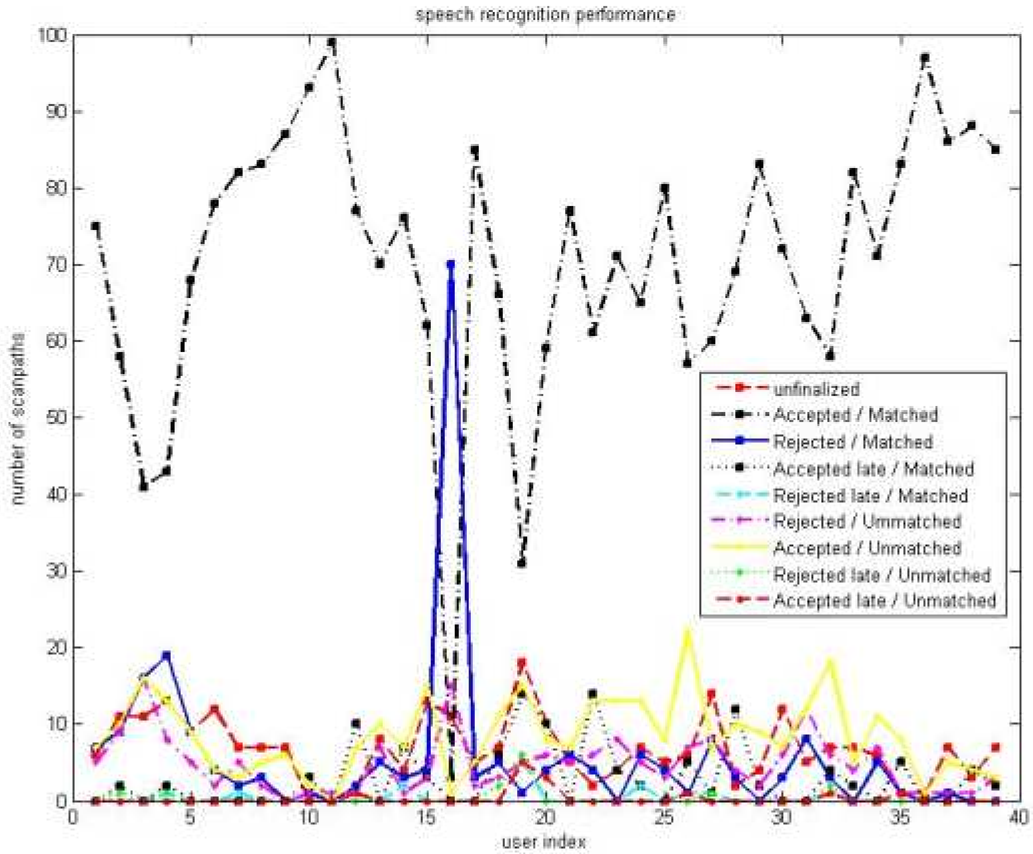
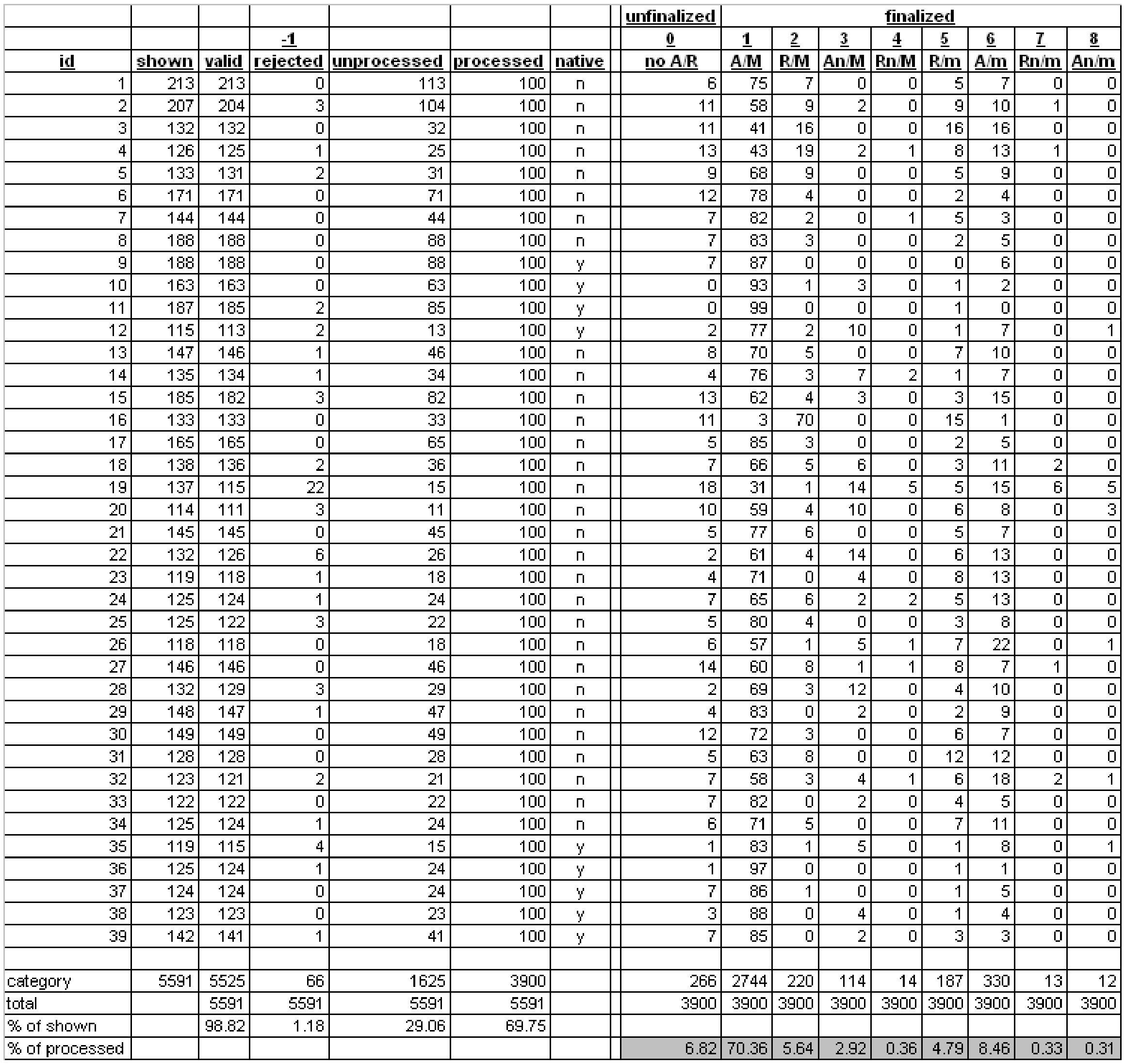


Figure 40 Scanpath classification using Recognition Categories in Constraint-Free Task

Figure 40 illustrates the proportions of samples with various categories during the Experiment 1. Only the first 100 samples of all subjects' usable samples are used for the analysis. The recognizer speech start/end events indicate approximate timings of the audio signal start/end from the recognizer. When the utterance is finalized with accept/reject event, the recognizer *adjusts* the start/end timestamps which reflect more accurate timestamps for the start/end of speech (described later in this chapter). The adaptive/predictive models can use the start/end events' timestamps when the finalized start/end timestamps are not available instead of rejecting the scanpaths. Notice also that there is considerable false rejection and false acceptance of subject utterances due to either user pronunciation errors or speech recognition errors.

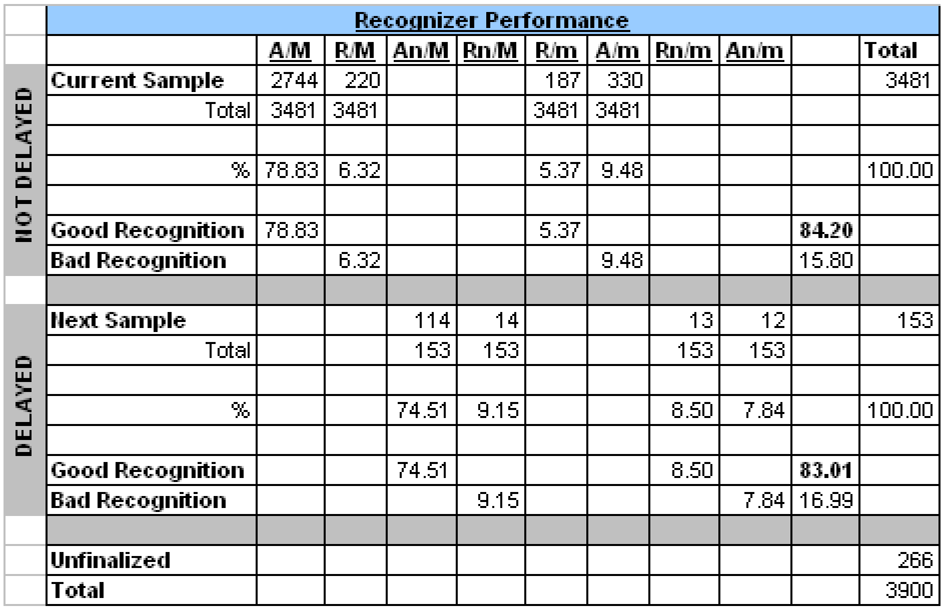


							unfinalized	finalized							
			-1				0	1	2	3	4	5	6	7	8
id	shown	valid	rejected	unprocessed	processed	native	no A/R	A/M	R/M	An/M	Rn/M	R/m	A/m	Rn/m	An/m
1	213	213	0	113	100	n	6	75	7	0	0	5	7	0	0
2	207	204	3	104	100	n	11	58	9	2	0	9	10	1	0
3	132	132	0	32	100	n	11	41	16	0	0	16	16	0	0
4	126	125	1	25	100	n	13	43	19	2	1	8	13	1	0
5	133	131	2	31	100	n	9	68	9	0	0	5	9	0	0
6	171	171	0	71	100	n	12	78	4	0	0	2	4	0	0
7	144	144	0	44	100	n	7	82	2	0	1	5	3	0	0
8	188	188	0	88	100	n	7	83	3	0	0	2	5	0	0
9	188	188	0	88	100	y	7	87	0	0	0	0	6	0	0
10	163	163	0	63	100	y	0	93	1	3	0	1	2	0	0
11	187	185	2	85	100	y	0	99	0	0	0	1	0	0	0
12	115	113	2	13	100	y	2	77	2	10	0	1	7	0	1
13	147	146	1	46	100	n	8	70	5	0	0	7	10	0	0
14	135	134	1	34	100	n	4	76	3	7	2	1	7	0	0
15	185	182	3	82	100	n	13	62	4	3	0	3	15	0	0
16	133	133	0	33	100	n	11	3	70	0	0	15	1	0	0
17	165	165	0	65	100	n	5	85	3	0	0	2	5	0	0
18	138	136	2	36	100	n	7	66	5	6	0	3	11	2	0
19	137	115	22	15	100	n	18	31	1	14	5	5	15	6	5
20	114	111	3	11	100	n	10	59	4	10	0	6	8	0	3
21	145	145	0	45	100	n	5	77	6	0	0	5	7	0	0
22	132	126	6	26	100	n	2	61	4	14	0	6	13	0	0
23	119	118	1	18	100	n	4	71	0	4	0	8	13	0	0
24	125	124	1	24	100	n	7	65	6	2	2	5	13	0	0
25	125	122	3	22	100	n	5	80	4	0	0	3	8	0	0
26	118	118	0	18	100	n	6	57	1	5	1	7	22	0	1
27	146	146	0	46	100	n	14	60	8	1	1	8	7	1	0
28	132	129	3	29	100	n	2	69	3	12	0	4	10	0	0
29	148	147	1	47	100	n	4	83	0	2	0	2	9	0	0
30	149	149	0	49	100	n	12	72	3	0	0	6	7	0	0
31	128	128	0	28	100	n	5	63	8	0	0	12	12	0	0
32	123	121	2	21	100	n	7	58	3	4	1	6	18	2	1
33	122	122	0	22	100	n	7	82	0	2	0	4	5	0	0
34	125	124	1	24	100	n	6	71	5	0	0	7	11	0	0
35	119	115	4	15	100	y	1	83	1	5	0	1	8	0	1
36	125	124	1	24	100	y	1	97	0	0	0	1	1	0	0
37	124	124	0	24	100	y	7	86	1	0	0	1	5	0	0
38	123	123	0	23	100	y	3	88	0	4	0	1	4	0	0
39	142	141	1	41	100	y	7	85	0	2	0	3	3	0	0
category	5591	5525	66	1625	3900		266	2744	220	114	14	187	330	13	12
total		5591	5591	5591	5591		3900	3900	3900	3900	3900	3900	3900	3900	3900
% of shown		98.82	1.18	29.06	69.75										
% of processed							6.82	70.36	5.64	2.92	0.36	4.79	8.46	0.33	0.31

Table 21 Scanpath Recognition Analysis in One Word Task

From Table 21, it can be seen that there is a considerable percentage (*i.e.*, 6.82%) of total data that did not receive accept/reject events or unfinalized. Note that only the **rejected** samples as defined in Table 19 cannot be processed because these samples do not have *SpeechStart / SpeechStop* events in addition to missing recognizer finalization events. If a speech start event is received then that sample can be processed though it may not be the accurate utterance start timestamp.

Apart from this kind of recognition categorization, Table 22 evaluates the recognizer's performance during good and bad recognitions. This is required to understand if the samples need to be treated any different if the recognition is delayed. It can be seen that the recognizer is 84.20% effective when processing the utterances which it can recognize quickly. Also, its performance is almost identical at 83.01% when the recognizer delayed recognizing the utterance.

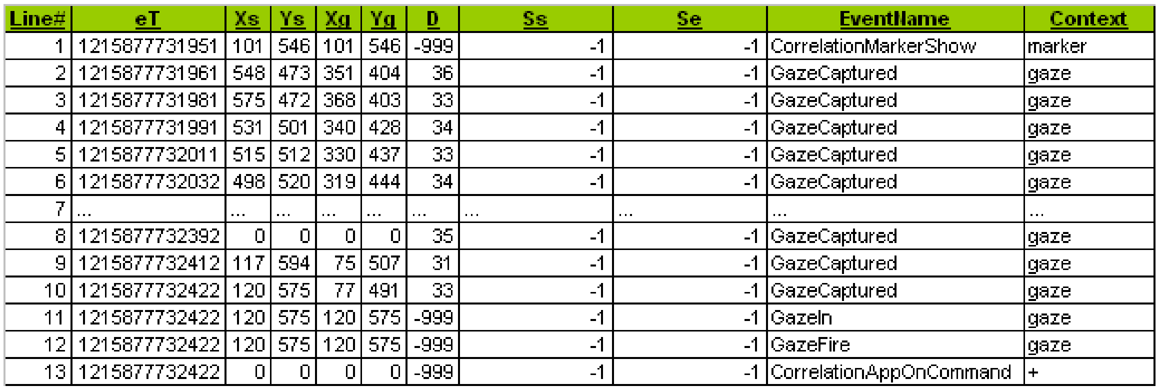


Recognizer Performance

		A/M	R/M	An/M	Rn/M	R/m	A/m	Rn/m	An/m		Total
NOT DELAYED	**Current Sample**	2744	220			187	330				3481
	Total	3481	3481			3481	3481				
	%	78.83	6.32			5.37	9.48				100.00
	Good Recognition	78.83				5.37				**84.20**	
	Bad Recognition		6.32				9.48			15.80	
DELAYED	**Next Sample**			114	14			13	12		153
	Total			153	153			153	153		
	%			74.51	9.15			8.50	7.84		100.00
	Good Recognition			74.51				8.50		**83.01**	
	Bad Recognition				9.15				7.84	16.99	
	Unfinalized										266
	Total										3900

Table 22 Performance of the Speech Recognizer in One Word Task

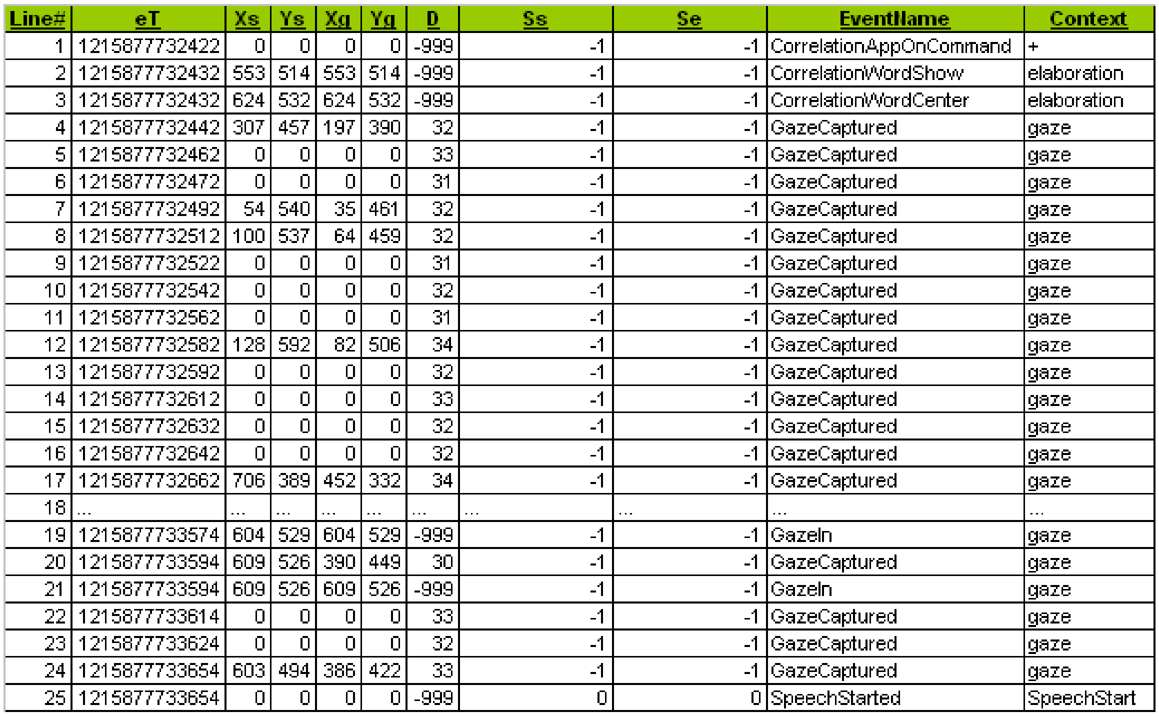
The *Marker Display / Dismissal* segment for Experiment 1 consists of events from the time the marker is shown to the point when the subject looks at the marker. This can be seen in Table 23 where the *EventName* in the first line indicates the marker display event and the *EventName* in the last line denotes that the subject has dismissed the marker by looking at it. The marker (i.e., + or cross-hair in Experiment 1) is shown as a reference before the word is looked at. When the subject looks at the marker it is registered as an application event *CorrelationAppOnCommand.* Line 7 (apart from header row) indicates that some of the gaze samples are skipped in displaying the data segment. Also, line 8 indicates that the gaze data is not recordable during that time possibly due to user movements.



Line#	eT	Xs	Ys	Xg	Yg	D	Ss	Se	EventName	Context
1	1215877731951	101	546	101	546	-999	-1	-1	CorrelationMarkerShow	marker
2	1215877731961	548	473	351	404	36	-1	-1	GazeCaptured	gaze
3	1215877731981	575	472	368	403	33	-1	-1	GazeCaptured	gaze
4	1215877731991	531	501	340	428	34	-1	-1	GazeCaptured	gaze
5	1215877732011	515	512	330	437	33	-1	-1	GazeCaptured	gaze
6	1215877732032	498	520	319	444	34	-1	-1	GazeCaptured	gaze
7	...	...	...	...	...	...	...	...	...	...
8	1215877732392	0	0	0	0	35	-1	-1	GazeCaptured	gaze
9	1215877732412	117	594	75	507	31	-1	-1	GazeCaptured	gaze
10	1215877732422	120	575	77	491	33	-1	-1	GazeCaptured	gaze
11	1215877732422	120	575	120	575	-999	-1	-1	GazeIn	gaze
12	1215877732422	120	575	120	575	-999	-1	-1	GazeFire	gaze
13	1215877732422	0	0	0	0	-999	-1	-1	CorrelationAppOnCommand	+

Table 23 Raw Speech Machine Data – Marker Display / Dismissal in Experiment 1

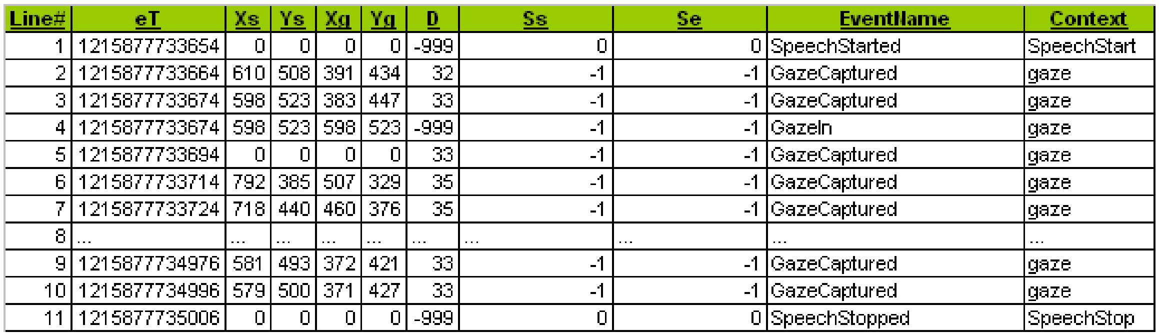
The *Word Display / Speech Start* data segment (Table 24) captures the data from the point when the subject looks at the marker to the point when the subject starts speaking. Gaze is captured continuously all the time in all the data segments. Each data segment is analyzed further to compute the accurate timestamps of all events in the data segments.



Line#	eT	Xs	Ys	Xg	Yg	D	Ss	Se	EventName	Context
1	1215877732422	0	0	0	0	-999	-1	-1	CorrelationAppOnCommand	+
2	1215877732432	553	514	553	514	-999	-1	-1	CorrelationWordShow	elaboration
3	1215877732432	624	532	624	532	-999	-1	-1	CorrelationWordCenter	elaboration
4	1215877732442	307	457	197	390	32	-1	-1	GazeCaptured	gaze
5	1215877732462	0	0	0	0	33	-1	-1	GazeCaptured	gaze
6	1215877732472	0	0	0	0	31	-1	-1	GazeCaptured	gaze
7	1215877732492	54	540	35	461	32	-1	-1	GazeCaptured	gaze
8	1215877732512	100	537	64	459	32	-1	-1	GazeCaptured	gaze
9	1215877732522	0	0	0	0	31	-1	-1	GazeCaptured	gaze
10	1215877732542	0	0	0	0	32	-1	-1	GazeCaptured	gaze
11	1215877732562	0	0	0	0	31	-1	-1	GazeCaptured	gaze
12	1215877732582	128	592	82	506	34	-1	-1	GazeCaptured	gaze
13	1215877732592	0	0	0	0	32	-1	-1	GazeCaptured	gaze
14	1215877732612	0	0	0	0	33	-1	-1	GazeCaptured	gaze
15	1215877732632	0	0	0	0	32	-1	-1	GazeCaptured	gaze
16	1215877732642	0	0	0	0	32	-1	-1	GazeCaptured	gaze
17	1215877732662	706	389	452	332	34	-1	-1	GazeCaptured	gaze
18	...	...	...	...	...	...	...	...	...	...
19	1215877733574	604	529	604	529	-999	-1	-1	GazeIn	gaze
20	1215877733594	609	526	390	449	30	-1	-1	GazeCaptured	gaze
21	1215877733594	609	526	609	526	-999	-1	-1	GazeIn	gaze
22	1215877733614	0	0	0	0	33	-1	-1	GazeCaptured	gaze
23	1215877733624	0	0	0	0	32	-1	-1	GazeCaptured	gaze
24	1215877733654	603	494	386	422	33	-1	-1	GazeCaptured	gaze
25	1215877733654	0	0	0	0	-999	0	0	SpeechStarted	SpeechStart

Table 24 Raw Speech Machine Data – Word Display / Speech Start in Experiment 1

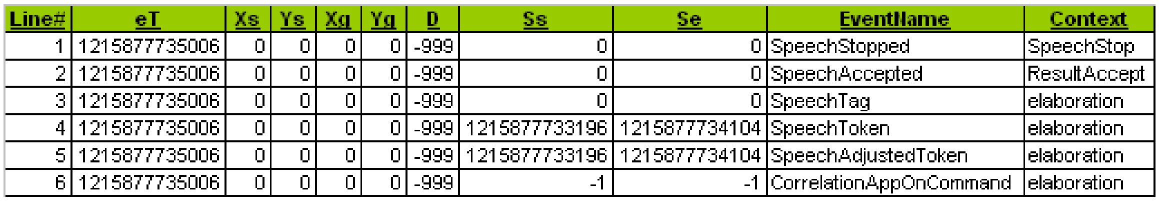
The *Speech Start / Stop* data segment (Table 25) captures the data from the point when the subject starts speaking to the point when the subject stops speaking. The first and last lines indicate these speech start/stop events. These events correspond to the speech recognizer’s application programming interface (API) events.



Line#	eT	Xs	Ys	Xg	Yg	D	Ss	Se	EventName	Context
1	1215877733654	0	0	0	0	-999	0	0	SpeechStarted	SpeechStart
2	1215877733664	610	508	391	434	32	-1	-1	GazeCaptured	gaze
3	1215877733674	598	523	383	447	33	-1	-1	GazeCaptured	gaze
4	1215877733674	598	523	598	523	-999	-1	-1	GazeIn	gaze
5	1215877733694	0	0	0	0	33	-1	-1	GazeCaptured	gaze
6	1215877733714	792	385	507	329	35	-1	-1	GazeCaptured	gaze
7	1215877733724	718	440	460	376	35	-1	-1	GazeCaptured	gaze
8	...	...	...	...	...	...	...	...	...	...
9	1215877734976	581	493	372	421	33	-1	-1	GazeCaptured	gaze
10	1215877734996	579	500	371	427	33	-1	-1	GazeCaptured	gaze
11	1215877735006	0	0	0	0	-999	0	0	SpeechStopped	SpeechStop

Table 25 Raw Speech Machine Data – Speech Start / Stop in Experiment 1

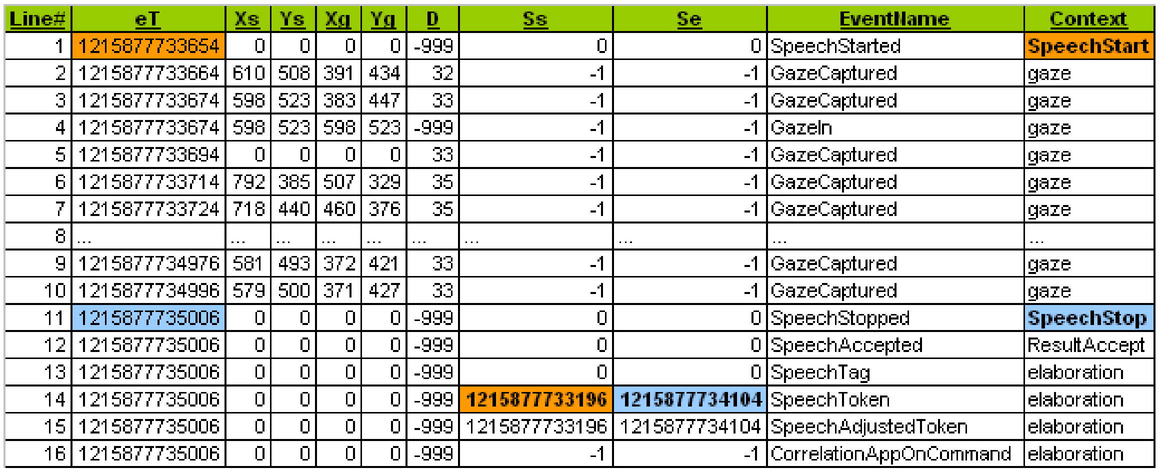
The *Recognizer Finalization* data segment (Table 26) captures the data from the point when the subject stops speaking to the point when the subject starts the next trial. In this segment, the recognizer's finalized events are received which gives the final and accurate timestamps of the word utterance. Note the Ss and Se columns now contain the finalized word timestamps from the recognizer. Also, the <Xs, Ys, Xg, Yg> columns are all 0 because the gaze data is not applicable during these events' collection. The value of D is set to -999 when it is not applicable to the event.



Line#	eT	Xs	Ys	Xg	Yg	D	Ss	Se	EventName	Context
1	1215877735006	0	0	0	0	-999	0	0	SpeechStopped	SpeechStop
2	1215877735006	0	0	0	0	-999	0	0	SpeechAccepted	ResultAccept
3	1215877735006	0	0	0	0	-999	0	0	SpeechTag	elaboration
4	1215877735006	0	0	0	0	-999	1215877733196	1215877734104	SpeechToken	elaboration
5	1215877735006	0	0	0	0	-999	1215877733196	1215877734104	SpeechAdjustedToken	elaboration
6	1215877735006	0	0	0	0	-999	-1	-1	CorrelationAppOnCommand	elaboration

Table 26 Raw Speech Machine Data – Recognizer Finalization in Experiment 1

In the IBM Via Voice speech recognizer, it is observed that the timestamps for the API events for *SpeechStart* and *SpeechStop* do not match with the finalized timestamps of word utterance. The finalized timestamps are the accurate timestamps from the recognizer. Sometimes the recognition engine may not be able to recognize the speech due to ambient noise and in those cases the API events *SpeechStart* and *SpeechStop* can help analyze the interaction. In Table 27, the *SpeechStarted* event (line 1) timestamp is greater than finalized speech start-timestamp (column Ss in line 14) from the recognizer. Similarly the *SpeechStopped* event (line 11) timestamp is greater than the finalized speech stop-timestamp (column Se in line 14) from the recognizer. This is because the subject's speech utterance starts before the recognizer can issue an API speech-start-event and similarly the subject's speech utterance stops before the recognizer can issue an API speech-stop-event.



Line#	eT	Xs	Ys	Xg	Yg	D	Ss	Se	EventName	Context
1	1215877733654	0	0	0	0	-999	0	0	SpeechStarted	SpeechStart
2	1215877733664	610	508	391	434	32	-1	-1	GazeCaptured	gaze
3	1215877733674	598	523	383	447	33	-1	-1	GazeCaptured	gaze
4	1215877733674	598	523	598	523	-999	-1	-1	GazeIn	gaze
5	1215877733694	0	0	0	0	33	-1	-1	GazeCaptured	gaze
6	1215877733714	792	385	507	329	35	-1	-1	GazeCaptured	gaze
7	1215877733724	718	440	460	376	35	-1	-1	GazeCaptured	gaze
8	...	...	...	...	...	...	...	...	...	...
9	1215877734976	581	493	372	421	33	-1	-1	GazeCaptured	gaze
10	1215877734996	579	500	371	427	33	-1	-1	GazeCaptured	gaze
11	1215877735006	0	0	0	0	-999	0	0	SpeechStopped	SpeechStop
12	1215877735006	0	0	0	0	-999	0	0	SpeechAccepted	ResultAccept
13	1215877735006	0	0	0	0	-999	0	0	SpeechTag	elaboration
14	1215877735006	0	0	0	0	-999	1215877733196	1215877734104	SpeechToken	elaboration
15	1215877735006	0	0	0	0	-999	1215877733196	1215877734104	SpeechAdjustedToken	elaboration
16	1215877735006	0	0	0	0	-999	-1	-1	CorrelationAppOnCommand	elaboration

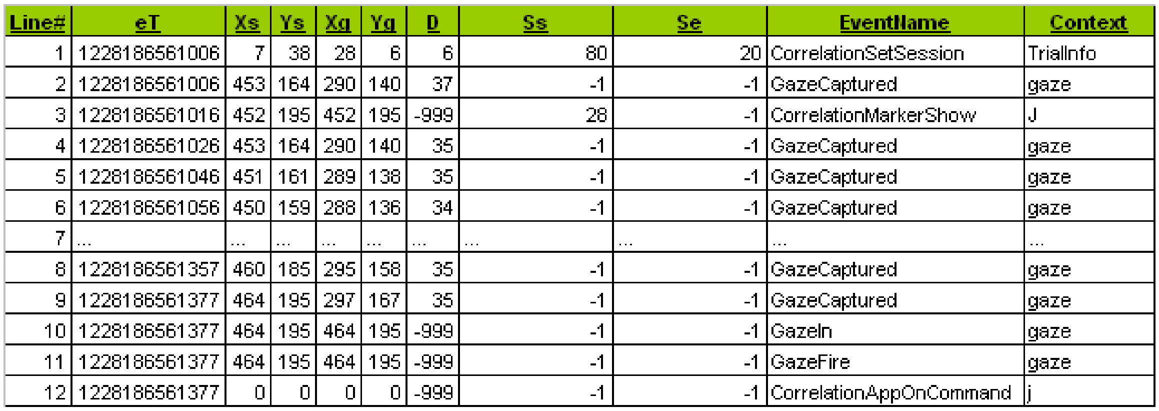
Table 27 Raw Speech Machine Data – Recognizer API events / Finalization in Experiment 1

E.4. Experiment 2 Data Capture

In Experiment 2, the subjects speak only 5 words in each session. One might expect that a word repeatedly pronounced should make the recognition of the word consistent. However, Table 28 shows recognition performance for different users (i.e., 20 from the total number of subjects in Experiment 2) and for different words by a single user. Column 1 in Table 28 corresponds to the session number in Experiment 2. Column 2 is the target word location in the 6x6 array. Column 3 (starting with u1) to 28 indicate the number of times a word is successfully recognized out of the total number of times the word is uttered, for each subject. Each row corresponds to a single word recognition performance by different users. It can be seen that a word can not be recognized all the time for a single user. Also, a single word cannot be recognized when spoken by multiple users. This poses a challenge in speech interfaces as to which word should be chosen as the command in the interface. It has never been a problem in non-multimodal (i.e. keyboard and mouse) systems because the user always clicks the command required. With the advent of multimodal systems, the different pronunciations of the command require additional disambiguation like gaze.

			All Users in Menu Interaction																													
session	target		u1	u2	u3	u4	u5	u6	u7	u8	u9	u10	u11	u12	u13	u14	u15	u16	u17	u18	u19	u20	u21	u22	u23	u24	u25	u26	Recognized	Total	P(word)		word
1	10		4	0	1	0	0	0	2	0	0	1	0	6	1	0	2	4	9	0	1	3	1	5	9	7	9	5	70	260	0.2692		jail
1	23		8	9	9	4	9	0	10	4	8	8	5	5	7	4	8	9	0	7	9	1	3	8	8	8	9	8	168	260	0.6462		keg
1	21		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	4	0	0	0	0	0	8	7	6	8	40	260	0.1538		earth
1	29		8	3	1	8	4	7	7	3	7	7	8	2	9	2	9	9	1	7	6	3	3	8	9	7	8	9	155	260	0.5962		venom
1	26		10	9	6	10	1	8	1	0	0	9	6	7	7	3	9	9	8	8	8	0	2	9	9	9	9	9	166	260	0.6385		abode
2	8		10	5	2	9	0	2	5	3	6	4	4	2	8	5	3	10	6	9	0	5	8	5	9	8	9	9	146	260	0.5615		abyss
2	15		10	8	9	9	8	10	6	3	5	7	9	9	9	9	9	9	9	9	6	9	7	3	9	6	9	7	203	260	0.7808		lake
2	17		8	5	8	10	8	8	5	6	0	8	7	8	9	9	8	9	2	6	9	2	6	7	4	1	4	7	164	260	0.6308		inn
2	20		10	10	9	10	5	10	1	8	8	8	9	8	9	9	6	9	6	7	9	8	8	9	7	8	8	9	208	260	0.8000		mast
2	22		9	8	4	9	5	7	10	7	9	8	7	5	9	7	8	9	9	8	9	4	9	7	7	7	9	9	199	260	0.7654		tank
3	9		5	4	0	4	7	0	7	0	0	2	0	8	0	0	4	8	0	1	0	8	1	0	5	8	8	8	88	260	0.3385		adage
3	11		10	10	6	8	8	1	7	5	6	2	7	4	1	9	9	9	2	9	7	9	7	8	9	8	8	9	178	260	0.6846		thief
3	16		10	10	10	9	10	9	10	8	10	10	4	9	9	7	9	9	9	9	9	8	8	7	8	9	9	6	225	260	0.8654		joke
3	28		10	9	10	8	1	10	7	9	9	6	8	10	9	9	8	10	2	7	9	9	5	7	8	9	8	9	206	260	0.7923		elbow
3	14		8	2	10	9	9	8	5	9	1	6	9	9	9	8	9	9	3	5	9	3	7	7	9	9	1	8	181	260	0.6962		nun
4	20		4	3	10	1	10	2	9	8	8	10	7	9	8	9	7	4	4	7	9	9	8	6	7	7	5	8	179	260	0.6885		law
4	16		6	0	0	4	1	0	10	1	0	0	0	8	6	0	0	9	0	6	1	2	0	9	8	8	9	8	96	260	0.3692		dell
4	28		1	0	0	0	1	0	1	0	0	2	0	0	1	0	0	8	0	1	5	1	0	0	5	2	2	4	34	260	0.1308		thorn
4	9		6	8	10	9	7	9	4	5	10	7	4	8	9	6	4	9	7	8	8	3	8	5	9	8	9	9	189	260	0.7269		rock
4	27		10	10	7	10	10	10	10	10	10	6	7	9	8	5	8	9	8	7	9	8	7	9	9	7	9	8	220	260	0.8462		agony
5	14		1	3	4	7	0	3	4	0	0	0	8	1	9	6	0	9	0	4	1	0	0	4	9	6	7	9	95	260	0.3654		oven
5	27		10	10	10	10	10	10	8	10	10	7	9	9	9	9	7	9	10	9	9	9	8	9	9	9	9	9	237	260	0.9115		idea
5	23		2	3	2	0	4	2	3	2	3	6	1	9	5	6	3	9	7	3	8	1	4	2	9	9	4	9	116	260	0.4462		lawn
5	21		10	2	5	6	6	3	1	5	8	4	6	7	2	1	7	9	1	3	7	2	5	6	9	9	8	9	141	260	0.5423		judge
5	17		7	0	0	0	0	1	2	8	0	0	1	6	3	0	0	9	9	1	0	1	1	8	9	9	9	9	93	260	0.3577		nail
6	22		3	7	1	3	9	8	1	10	5	6	7	9	4	0	8	9	5	7	9	6	2	7	9	8	9	9	161	260	0.6192		rosin
6	15		10	10	10	5	9	9	9	8	8	9	8	9	9	8	9	9	8	9	9	8	9	8	9	9	9	9	226	260	0.8692		unit
6	8		9	6	7	6	4	9	10	10	10	9	7	9	9	7	9	9	1	9	9	8	9	9	5	9	9	7	205	260	0.7885		lemon
6	11		9	10	10	8	9	10	10	5	10	10	8	9	9	8	9	9	9	9	9	1	8	9	9	8	9	9	223	260	0.8577		idiom
6	26		7	1	7	2	0	1	0	1	0	1	0	7	1	6	4	9	7	1	5	2	6	8	9	3	7	7	102	260	0.3923		owner
7	9		10	8	10	0	8	9	10	3	10	9	8	9	8	8	8	8	6	8	9	8	8	7	9	3	6	8	198	260	0.7615		gist
7	14		10	9	10	9	3	8	5	0	5	5	9	8	9	9	5	9	1	9	6	7	5	8	5	9	8	7	178	260	0.6846		elbow
7	16		4	8	9	4	4	0	6	1	1	2	2	1	0	0	0	9	5	5	5	0	0	3	8	8	7	9	101	260	0.3885		tomb
7	17		10	6	10	10	8	9	4	5	3	5	9	9	9	8	9	9	5	8	9	2	7	7	9	7	5	9	191	260	0.7346		nun
7	29		10	0	10	9	6	1	6	5	0	9	2	8	5	5	0	9	7	1	3	6	7	7	9	7	7	8	147	260	0.5654		jury
8	26		7	10	4	3	1	8	9	6	10	2	9	0	9	6	6	7	7	9	1	3	4	7	9	8	6	6	157	260	0.6038		venom
8	22		9	4	10	1	8	9	7	2	4	6	2	9	9	3	3	9	1	9	5	7	8	9	9	8	9	8	168	260	0.6462		kine
8	15		9	3	7	0	0	1	6	4	0	1	1	2	0	0	0	9	8	6	0	0	0	0	2	9	9	9	86	260	0.3308		monk
8	10		8	3	1	3	5	7	8	4	3	1	6	2	4	7	1	9	9	6	7	4	5	9	9	8	9	9	147	260	0.5654		tool
8	11		6	10	7	9	4	7	1	5	10	4	4	7	9	4	2	9	8	9	7	2	9	6	9	9	9	9	175	260	0.6731		rock
Total			298	226	246	226	202	216	227	183	197	207	208	256	250	202	210	345	203	246	241	172	203	252	320	298	303	325					
P(user)			0.75	0.57	0.62	0.57	0.51	0.54	0.57	0.46	0.49	0.52	0.52	0.64	0.63	0.51	0.53	0.86	0.51	0.62	0.60	0.43	0.51	0.63	0.80	0.75	0.76	0.81					

Table 28 Recognition Performance in Menu System for a subset of users in Experiment 2

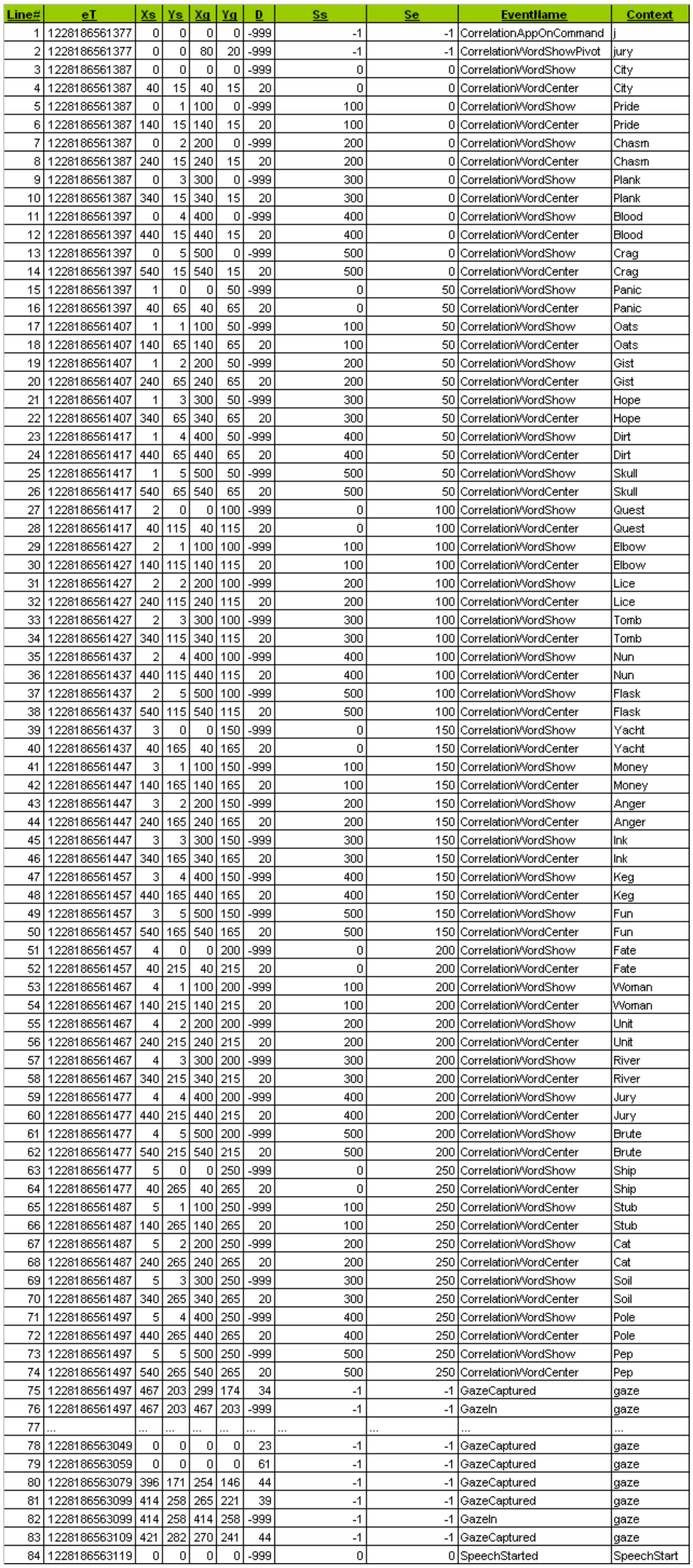


Line#	eT	Xs	Ys	Xg	Yg	D	Ss	Se	EventName	Context
1	1228186561006	7	38	28	6	6	80	20	CorrelationSetSession	TrialInfo
2	1228186561006	453	164	290	140	37	-1	-1	GazeCaptured	gaze
3	1228186561016	452	195	452	195	-999	28	-1	CorrelationMarkerShow	J
4	1228186561026	453	164	290	140	35	-1	-1	GazeCaptured	gaze
5	1228186561046	451	161	289	138	35	-1	-1	GazeCaptured	gaze
6	1228186561056	450	159	288	136	34	-1	-1	GazeCaptured	gaze
7	...	...	...	...	...	...	...	...	...	...
8	1228186561357	460	185	295	158	35	-1	-1	GazeCaptured	gaze
9	1228186561377	464	195	297	167	35	-1	-1	GazeCaptured	gaze
10	1228186561377	464	195	464	195	-999	-1	-1	GazeIn	gaze
11	1228186561377	464	195	464	195	-999	-1	-1	GazeFire	gaze
12	1228186561377	0	0	0	0	-999	-1	-1	CorrelationAppOnCommand	j

Table 29 Raw Speech Machine Data – Marker Display / Dismissal in Experiment 2

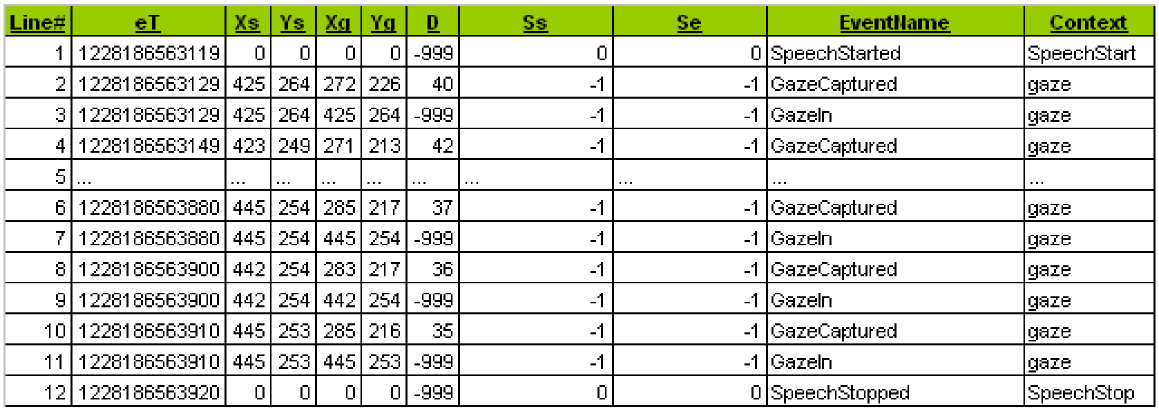
Table 29 through Table 32 indicates all the four data segments of a single trial in Experiment 2. Although experiments 1 and 2 are identical in that the subject speaks only a single word in a trial, there are several differences in the task complexity with respect to gaze. The subject looks at a cross-hair '+' in Experiment 1 (Table 23) whereas the subject looks at a letter in Experiment 2 (Table 29).

There is more gaze activity due to interference (i.e., surrounding objects) in Experiment 2 than in Experiment 1(Table 30 and Table 24). The subject need not have to remember anything after looking at the cross-hair in Experiment 1 whereas the subject needs to remember the letter until the subject speaks the word in Experiment 2. Experiment 1 displays only one word whereas Experiment 2 displays 36 words.



Line#	eT	Xs	Ys	Xg	Yg	D	Ss	Se	EventName	Context
1	1228186561377	0	0	0	0	-999	-1	-1	CorrelationAppOnCommand	j
2	1228186561377	0	0	80	20	-999	-1	-1	CorrelationWordShowPivot	jury
3	1228186561387	0	0	0	0	-999	0	0	CorrelationWordShow	City
4	1228186561387	40	15	40	15	20	0	0	CorrelationWordCenter	City
5	1228186561387	0	1	100	0	-999	100	0	CorrelationWordShow	Pride
6	1228186561387	140	15	140	15	20	100	0	CorrelationWordCenter	Pride
7	1228186561387	0	2	200	0	-999	200	0	CorrelationWordShow	Chasm
8	1228186561387	240	15	240	15	20	200	0	CorrelationWordCenter	Chasm
9	1228186561387	0	3	300	0	-999	300	0	CorrelationWordShow	Plank
10	1228186561387	340	15	340	15	20	300	0	CorrelationWordCenter	Plank
11	1228186561397	0	4	400	0	-999	400	0	CorrelationWordShow	Blood
12	1228186561397	440	15	440	15	20	400	0	CorrelationWordCenter	Blood
13	1228186561397	0	5	500	0	-999	500	0	CorrelationWordShow	Crag
14	1228186561397	540	15	540	15	20	500	0	CorrelationWordCenter	Crag
15	1228186561397	1	0	0	50	-999	0	50	CorrelationWordShow	Panic
16	1228186561397	40	65	40	65	20	0	50	CorrelationWordCenter	Panic
17	1228186561407	1	1	100	50	-999	100	50	CorrelationWordShow	Oats
18	1228186561407	140	65	140	65	20	100	50	CorrelationWordCenter	Oats
19	1228186561407	1	2	200	50	-999	200	50	CorrelationWordShow	Gist
20	1228186561407	240	65	240	65	20	200	50	CorrelationWordCenter	Gist
21	1228186561407	1	3	300	50	-999	300	50	CorrelationWordShow	Hope
22	1228186561407	340	65	340	65	20	300	50	CorrelationWordCenter	Hope
23	1228186561417	1	4	400	50	-999	400	50	CorrelationWordShow	Dirt
24	1228186561417	440	65	440	65	20	400	50	CorrelationWordCenter	Dirt
25	1228186561417	1	5	500	50	-999	500	50	CorrelationWordShow	Skull
26	1228186561417	540	65	540	65	20	500	50	CorrelationWordCenter	Skull
27	1228186561417	2	0	0	100	-999	0	100	CorrelationWordShow	Quest
28	1228186561417	40	115	40	115	20	0	100	CorrelationWordCenter	Quest
29	1228186561427	2	1	100	100	-999	100	100	CorrelationWordShow	Elbow
30	1228186561427	140	115	140	115	20	100	100	CorrelationWordCenter	Elbow
31	1228186561427	2	2	200	100	-999	200	100	CorrelationWordShow	Lice
32	1228186561427	240	115	240	115	20	200	100	CorrelationWordCenter	Lice
33	1228186561427	2	3	300	100	-999	300	100	CorrelationWordShow	Tomb
34	1228186561427	340	115	340	115	20	300	100	CorrelationWordCenter	Tomb
35	1228186561437	2	4	400	100	-999	400	100	CorrelationWordShow	Nun
36	1228186561437	440	115	440	115	20	400	100	CorrelationWordCenter	Nun
37	1228186561437	2	5	500	100	-999	500	100	CorrelationWordShow	Flask
38	1228186561437	540	115	540	115	20	500	100	CorrelationWordCenter	Flask
39	1228186561437	3	0	0	150	-999	0	150	CorrelationWordShow	Yacht
40	1228186561437	40	165	40	165	20	0	150	CorrelationWordCenter	Yacht
41	1228186561447	3	1	100	150	-999	100	150	CorrelationWordShow	Money
42	1228186561447	140	165	140	165	20	100	150	CorrelationWordCenter	Money
43	1228186561447	3	2	200	150	-999	200	150	CorrelationWordShow	Anger
44	1228186561447	240	165	240	165	20	200	150	CorrelationWordCenter	Anger
45	1228186561447	3	3	300	150	-999	300	150	CorrelationWordShow	Ink
46	1228186561447	340	165	340	165	20	300	150	CorrelationWordCenter	Ink
47	1228186561457	3	4	400	150	-999	400	150	CorrelationWordShow	Keg
48	1228186561457	440	165	440	165	20	400	150	CorrelationWordCenter	Keg
49	1228186561457	3	5	500	150	-999	500	150	CorrelationWordShow	Fun
50	1228186561457	540	165	540	165	20	500	150	CorrelationWordCenter	Fun
51	1228186561457	4	0	0	200	-999	0	200	CorrelationWordShow	Fate
52	1228186561457	40	215	40	215	20	0	200	CorrelationWordCenter	Fate
53	1228186561467	4	1	100	200	-999	100	200	CorrelationWordShow	Woman
54	1228186561467	140	215	140	215	20	100	200	CorrelationWordCenter	Woman
55	1228186561467	4	2	200	200	-999	200	200	CorrelationWordShow	Unit
56	1228186561467	240	215	240	215	20	200	200	CorrelationWordCenter	Unit
57	1228186561467	4	3	300	200	-999	300	200	CorrelationWordShow	River
58	1228186561467	340	215	340	215	20	300	200	CorrelationWordCenter	River
59	1228186561477	4	4	400	200	-999	400	200	CorrelationWordShow	Jury
60	1228186561477	440	215	440	215	20	400	200	CorrelationWordCenter	Jury
61	1228186561477	4	5	500	200	-999	500	200	CorrelationWordShow	Brute
62	1228186561477	540	215	540	215	20	500	200	CorrelationWordCenter	Brute
63	1228186561477	5	0	0	250	-999	0	250	CorrelationWordShow	Ship
64	1228186561477	40	265	40	265	20	0	250	CorrelationWordCenter	Ship
65	1228186561487	5	1	100	250	-999	100	250	CorrelationWordShow	Stub
66	1228186561487	140	265	140	265	20	100	250	CorrelationWordCenter	Stub
67	1228186561487	5	2	200	250	-999	200	250	CorrelationWordShow	Cat
68	1228186561487	240	265	240	265	20	200	250	CorrelationWordCenter	Cat
69	1228186561487	5	3	300	250	-999	300	250	CorrelationWordShow	Soil
70	1228186561487	340	265	340	265	20	300	250	CorrelationWordCenter	Soil
71	1228186561497	5	4	400	250	-999	400	250	CorrelationWordShow	Pole
72	1228186561497	440	265	440	265	20	400	250	CorrelationWordCenter	Pole
73	1228186561497	5	5	500	250	-999	500	250	CorrelationWordShow	Pep
74	1228186561497	540	265	540	265	20	500	250	CorrelationWordCenter	Pep
75	1228186561497	467	203	299	174	34	-1	-1	GazeCaptured	gaze
76	1228186561497	467	203	467	203	-999	-1	-1	GazeIn	gaze
77	...	...	...	...	...	...	...	...	...	...
78	1228186563049	0	0	0	0	23	-1	-1	GazeCaptured	gaze
79	1228186563059	0	0	0	0	61	-1	-1	GazeCaptured	gaze
80	1228186563079	396	171	254	146	44	-1	-1	GazeCaptured	gaze
81	1228186563099	414	258	265	221	39	-1	-1	GazeCaptured	gaze
82	1228186563099	414	258	414	258	-999	-1	-1	GazeIn	gaze
83	1228186563109	421	282	270	241	44	-1	-1	GazeCaptured	gaze
84	1228186563119	0	0	0	0	-999	0	0	SpeechStarted	SpeechStart

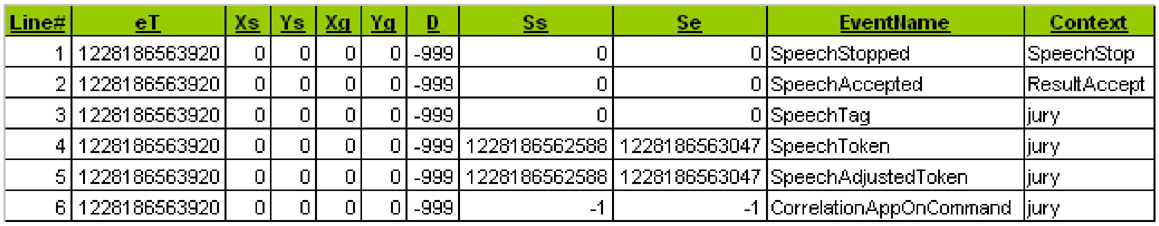
Table 30 Raw Speech Machine Data – Word Display / Speech Start in Experiment 2



Line#	eT	Xs	Ys	Xg	Yg	D	Ss	Se	EventName	Context
1	1228186563119	0	0	0	0	-999	0	0	SpeechStarted	SpeechStart
2	1228186563129	425	264	272	226	40	-1	-1	GazeCaptured	gaze
3	1228186563129	425	264	425	264	-999	-1	-1	GazeIn	gaze
4	1228186563149	423	249	271	213	42	-1	-1	GazeCaptured	gaze
5	...	...	...	...	...	...	...	...	...	...
6	1228186563880	445	254	285	217	37	-1	-1	GazeCaptured	gaze
7	1228186563880	445	254	445	254	-999	-1	-1	GazeIn	gaze
8	1228186563900	442	254	283	217	36	-1	-1	GazeCaptured	gaze
9	1228186563900	442	254	442	254	-999	-1	-1	GazeIn	gaze
10	1228186563910	445	253	285	216	35	-1	-1	GazeCaptured	gaze
11	1228186563910	445	253	445	253	-999	-1	-1	GazeIn	gaze
12	1228186563920	0	0	0	0	-999	0	0	SpeechStopped	SpeechStop

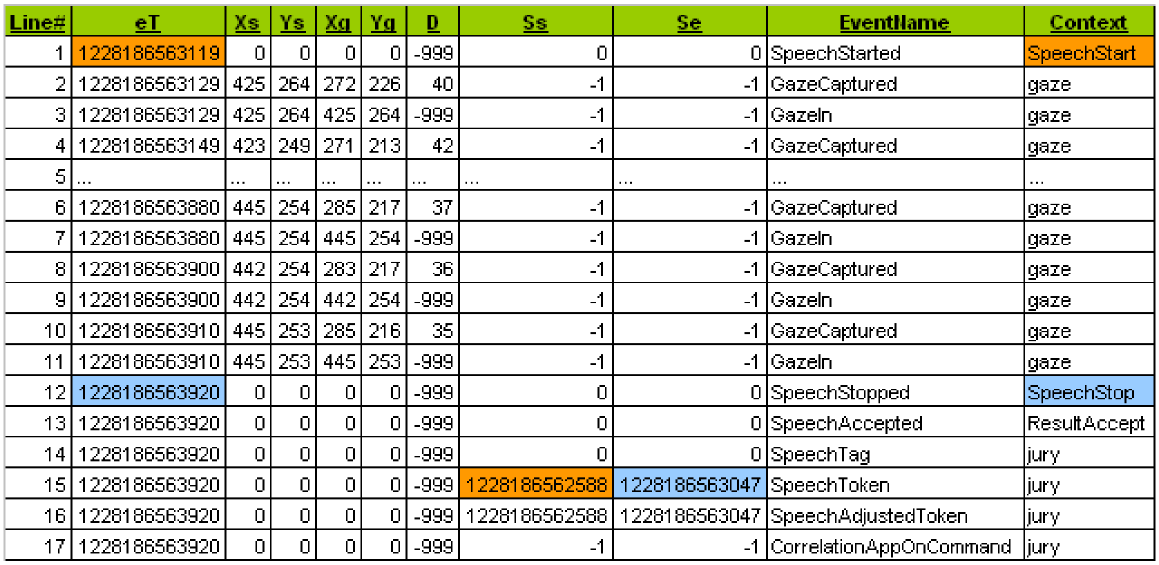
Table 31 Raw Speech Machine Data –Speech Start / Stop in Experiment 2

Table 25 and Table 31 are identical for *SpeechStart / SpeechStop* events. Similarly Table 26 and Table 32 are identical for recognizer finalization events of the trial.



Line#	eT	Xs	Ys	Xg	Yg	D	Ss	Se	EventName	Context
1	1228186563920	0	0	0	0	-999	0	0	SpeechStopped	SpeechStop
2	1228186563920	0	0	0	0	-999	0	0	SpeechAccepted	ResultAccept
3	1228186563920	0	0	0	0	-999	0	0	SpeechTag	jury
4	1228186563920	0	0	0	0	-999	1228186562588	1228186563047	SpeechToken	jury
5	1228186563920	0	0	0	0	-999	1228186562588	1228186563047	SpeechAdjustedToken	jury
6	1228186563920	0	0	0	0	-999	-1	-1	CorrelationAppOnCommand	jury

Table 32 Raw Speech Machine Data – Recognizer Finalization in Experiment 2



Line#	eT	Xs	Ys	Xg	Yg	D	Ss	Se	EventName	Context
1	1228186563119	0	0	0	0	-999	0	0	SpeechStarted	SpeechStart
2	1228186563129	425	264	272	226	40	-1	-1	GazeCaptured	gaze
3	1228186563129	425	264	425	264	-999	-1	-1	GazeIn	gaze
4	1228186563149	423	249	271	213	42	-1	-1	GazeCaptured	gaze
5	...	...	...	...	...	...	...	...	...	...
6	1228186563880	445	254	285	217	37	-1	-1	GazeCaptured	gaze
7	1228186563880	445	254	445	254	-999	-1	-1	GazeIn	gaze
8	1228186563900	442	254	283	217	36	-1	-1	GazeCaptured	gaze
9	1228186563900	442	254	442	254	-999	-1	-1	GazeIn	gaze
10	1228186563910	445	253	285	216	35	-1	-1	GazeCaptured	gaze
11	1228186563910	445	253	445	253	-999	-1	-1	GazeIn	gaze
12	1228186563920	0	0	0	0	-999	0	0	SpeechStopped	SpeechStop
13	1228186563920	0	0	0	0	-999	0	0	SpeechAccepted	ResultAccept
14	1228186563920	0	0	0	0	-999	0	0	SpeechTag	jury
15	1228186563920	0	0	0	0	-999	1228186562588	1228186563047	SpeechToken	jury
16	1228186563920	0	0	0	0	-999	1228186562588	1228186563047	SpeechAdjustedToken	jury
17	1228186563920	0	0	0	0	-999	-1	-1	CorrelationAppOnCommand	jury

Table 33 Raw Speech Machine Data – Recognizer API events / Finalization in Experiment 2

Table 33 shows how the recognizer’s finalized timestamps differ from the API events from the recognizer, similar to Table 27.

E.5. Gaze/Speech Data Preprocessing

Each subject performs a set of trials of an experiment. Each trial constitutes of a set of speech events and gaze samples collectively called a *scanpath.* The data collected for each scanpath needs to be verified and preprocessed before the data can be used as input to the predictive/adaptive models. Figure 41 illustrates the pre-processing where the two event datasets, the raw gaze data generated by the gaze machine *<subject>.tda* and the raw speech machine data *<subject>.txt,* are processed to generate a set of internal data structures. The *GazePreProcessor* (Figure 41) creates a gaze data file *g.g* as sent by the gaze machine. The *EventIndexer* and *SpeechMachineDataPreProcessor* create another gaze data file *s.g,* which is the gaze data as received by the speech machine. The two gaze data files *g.g* and *s.g* are aligned to verify that the data sent by the gaze machine is the same as the gaze data received by the speech machine. The *EventAnalyzer* creates additional internal data structures required by the prediction/adaptation models.

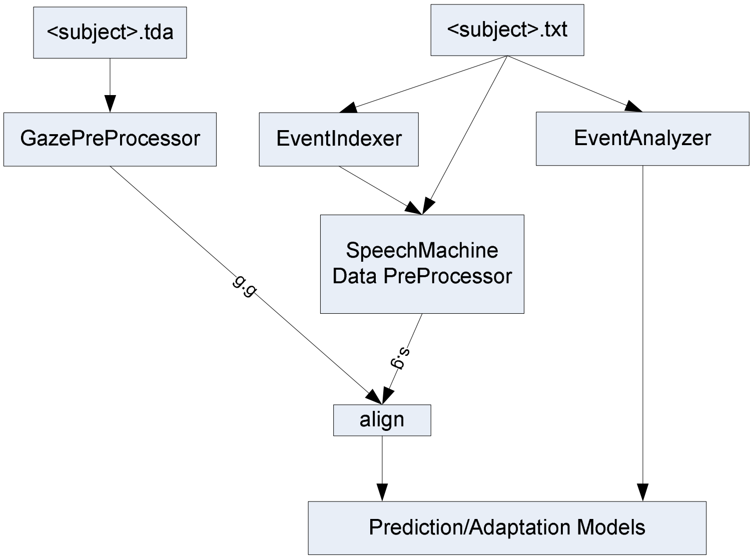
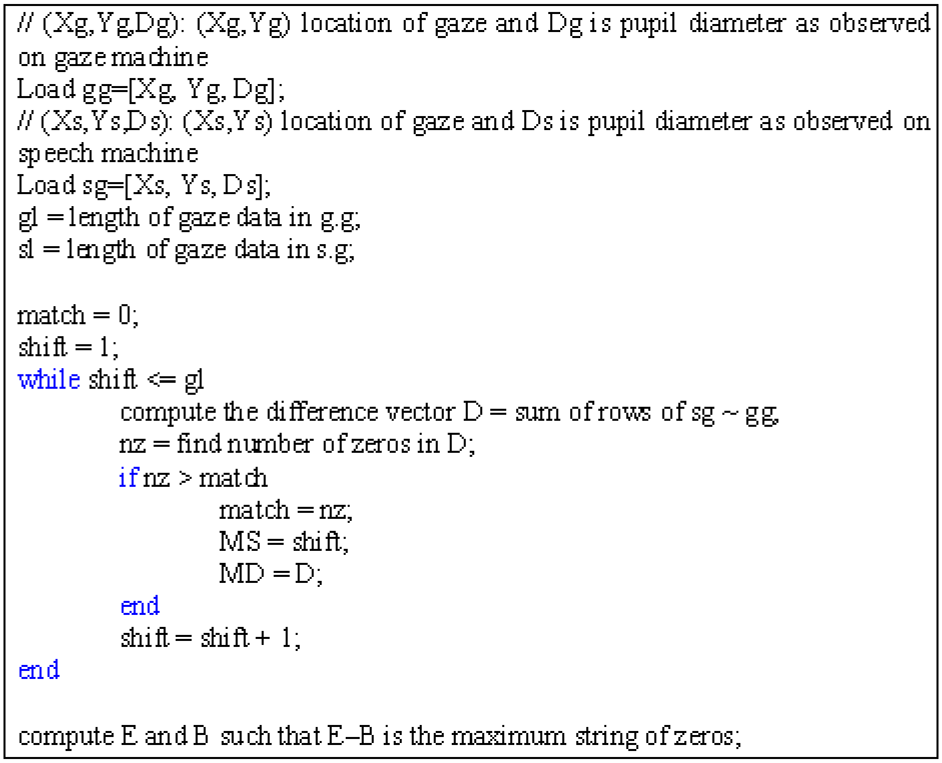


Figure 41 Preprocessing of Event Data

The two gaze data files collected on gaze and speech machines ***g.g*** and ***s.g*** (Figure 41) are compared for <x, y, d> tuples to ensure there is no synchronization problems in data transmission between the two machines. Figure 42 indicates the alignment algorithm to verify the data transmission between the two machines. B and E mark the beginning and ending of the string of zeros which are used to calculate the maximum length of the sequence of (0, 0)'s. The maximum length sequence ensures that the data is aligned between the machines from some point of time. Data alignment verification ensures data integrity and data transmission order between speech and gaze machines. It is only a validation mechanism and does not have any impact on experimental results.



```
// (Xg,Yg,Dg): (Xg,Yg) location of gaze and Dg is pupil diameter as observed
on gaze machine
Load gg=[Xg, Yg, Dg];
// (Xs,Ys,Ds): (Xs,Ys) location of gaze and Ds is pupil diameter as observed on
speech machine
Load sg=[Xs, Ys, Ds];
gl = length of gaze data in g.g;
sl = length of gaze data in s.g;

match = 0;
shift = 1;
while shift <= gl
        compute the difference vector D = sum of rows of sg ~ gg,
        nz = find number of zeros in D;
        if nz > match
                match = nz;
                MS = shift;
                MD = D;
        end
        shift = shift + 1;
end

compute E and B  such that E–B is the maximum string of zeros;
```

Figure 42 Data Alignment Algorithm

A gaze sample *g(x, y, d)* obtained on gaze machine at time *t1* is received by the speech machine at time *t2>t1*. Ideally speaking *t2-t1* should be zero or close to zero. In reality, there are several factors contributing to a non-zero time delay. There is always a finite non-zero delay in transmitting the gaze sample data from gaze machine to speech machine over a serial port. Also, the clocks on the two machines could be different attributing to a fixed time delay term regardless of careful clock setting on both machines. Two time delays *global* and *local* are computed to ensure the time delay of gaze data collected from the two machines is not reflected in the results in any manner. The *global time delay* is the difference of the gaze sample timestamp on the speech machine and the corresponding gaze sample’s timestamp on gaze machine. The *local time delay* of a gaze sample is calculated with reference to the scanpath’s starting time using the 60Hz sampling rate of the eye tracker. The timestamp of gaze samples in each scanpath are calculated as multiples of 16 (with a fixed offset from scanpath’s starting timestamp) and then compared with the actual event timestamp. Both the local (Figure 43) and global (Figure 44) time delays are periodic in nature. The amplitude of the global or local time delay is of the order of 10ms indicating that the error in gaze sample’s timestamp due to experimental setup is of the order of 10ms. Note that this correction is not applied in the data processing and is left for future studies. The fixed offset of global time delay compared to local delay is due to the clock settings on both machines and can safely be ignored. One can also observe an occasional large spike (Figure 43 and Figure 44) in the time delay which can be attributed to serial port communication delay due to buffering. Even if both speech recognition and eye tracking were to be running on a single machine system, they can compete for system resources and potentially introduce timestamp errors. It may not be even possible to compute these timestamp errors accurately because the operating system or the device driver’s log information would be needed to analyze them. And logging such low level information would potentially invalidate the whole results because of additional delay in writing such low level information. So, an experimental setup on two machines for this kind of high computational task is reasonable provided the data capture process is validated.

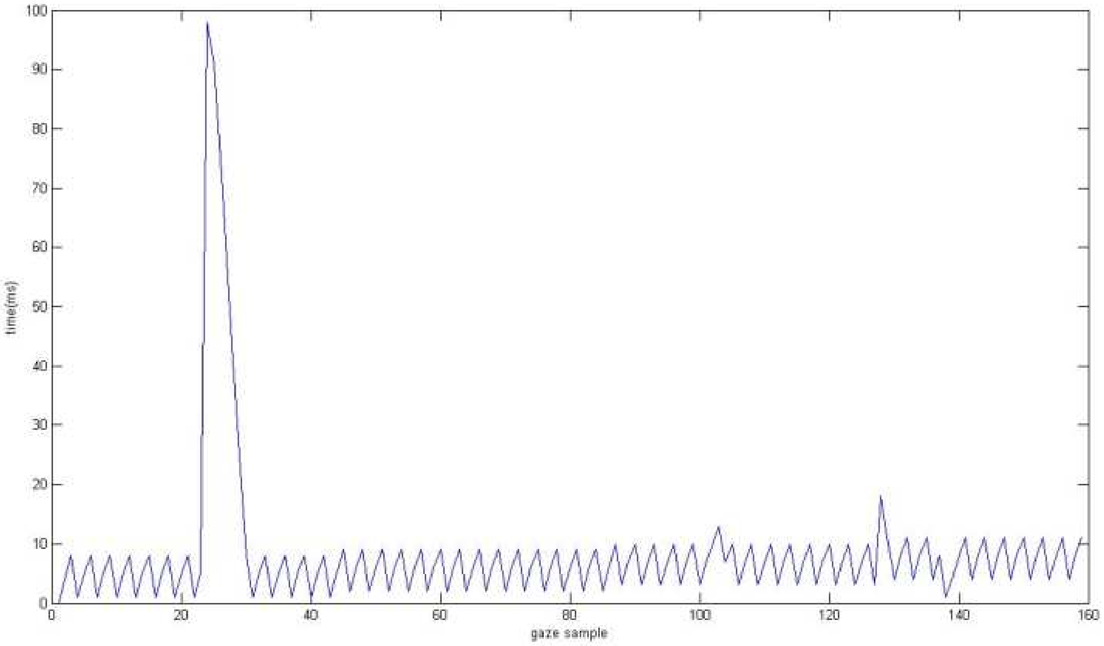


Figure 43 Gaze Sample Local Timestamp Differences

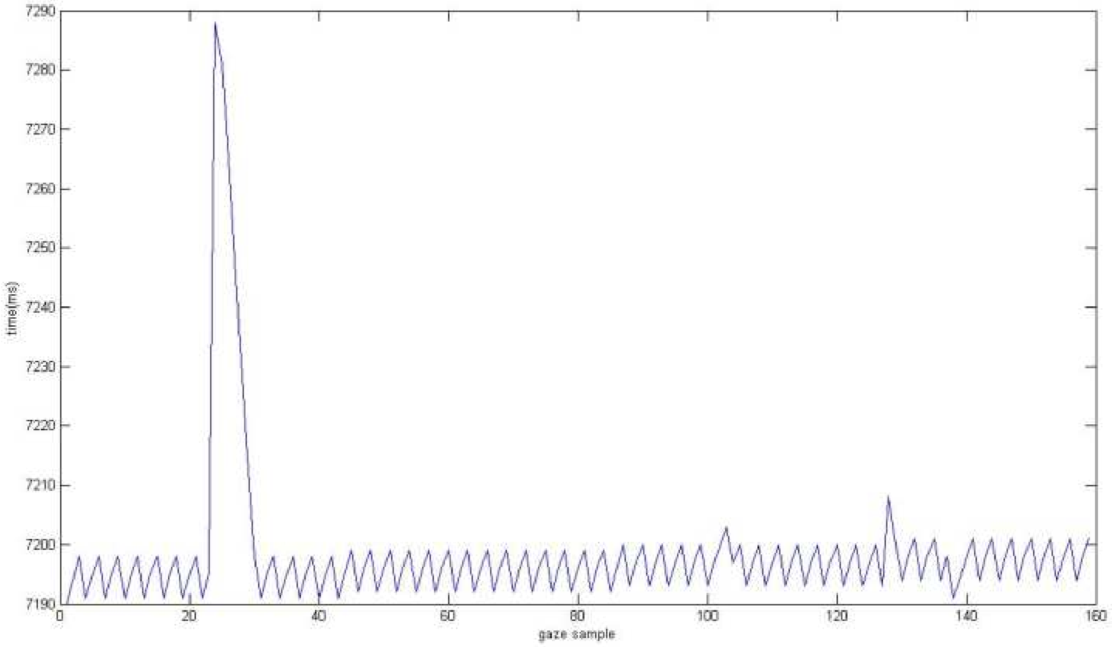
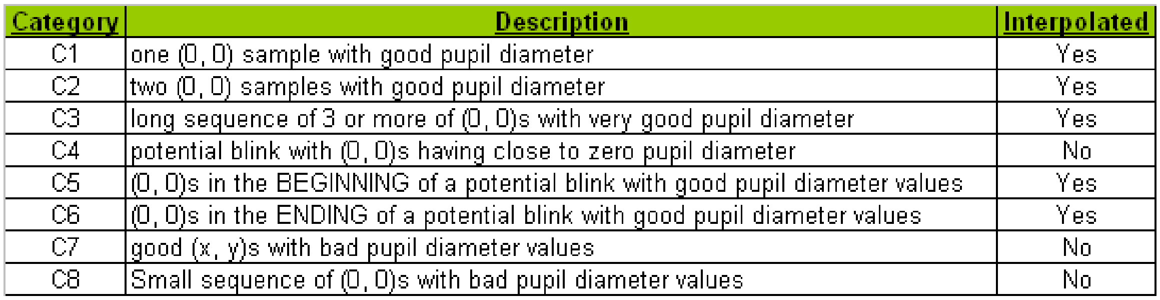


Figure 44 Gaze Sample Global Timestamp Differences

E.6. Scanpath Analysis

A *scanpath* can be informally defined as a set of gaze samples a subject traverses through on the screen in an interaction which contains both *fixations* and *saccades* (*i.e.,* sudden displacements of the eye from one location in space to another or between two fixations). Each trial of both the experiments is treated as a scanpath. The scanpath contains gaze tuples ***(x, y, d)*** where ***(x, y)*** is the gaze location as computed by the eye tracker that is interpolated to the target screen and ***d*** is the pupil diameter. Factors like user movements and blinks cause the ***(x, y, d)*** tuples to contain (0, 0) values for location coordinates. To avoid the effect of these zeros on fixations, the scanpath's gaze data is interpolated to fill these zero-sequences. Table 34 defines various categories in the gaze data and illustrates whether these categories are interpolated or not. For each scanpath (*i.e.,* a trial in either Experiment 1 or 2), mean d*m* and standard deviation d*s* of pupil diameter for all gaze samples is computed and for each sequence of (0, 0) in the gaze data, an average value of pupil diameter D*nz* is calculated. A sequence of (0, 0) values is considered for interpolation if *Dnz* > *dm* – *ds* otherwise it is not considered for interpolation (*i.e.,* a potential blink).

Data has been analyzed with and without interpolating blinks but no significant impact has been found on speech/gaze integration. The results are shown without interpolating the blinks in order to align closely with the underlying physical process. However, further research is necessary to understand the full impact of interpolation on integration model.



Category	Description	Interpolated
C1	one (0, 0) sample with good pupil diameter	Yes
C2	two (0, 0) samples with good pupil diameter	Yes
C3	long sequence of 3 or more of (0, 0)s with very good pupil diameter	Yes
C4	potential blink with (0, 0)s having close to zero pupil diameter	No
C5	(0, 0)s in the BEGINNING of a potential blink with good pupil diameter values	Yes
C6	(0, 0)s in the ENDING of a potential blink with good pupil diameter values	Yes
C7	good (x, y)s with bad pupil diameter values	No
C8	Small sequence of (0, 0)s with bad pupil diameter values	No

Table 34 Gaze Categories

Figure 45 shows the raw coordinates ***(Xs, Ys)*** of all gaze samples of a single scan path of a subject (a, c) and the same data after the interpolation (b, d). In this case the interpolation doesn't have any affect on the raw gaze data other than a smoothing effect. Notice that the blink *i.e.*, long sequence of (0, 0) values is not interpolated.

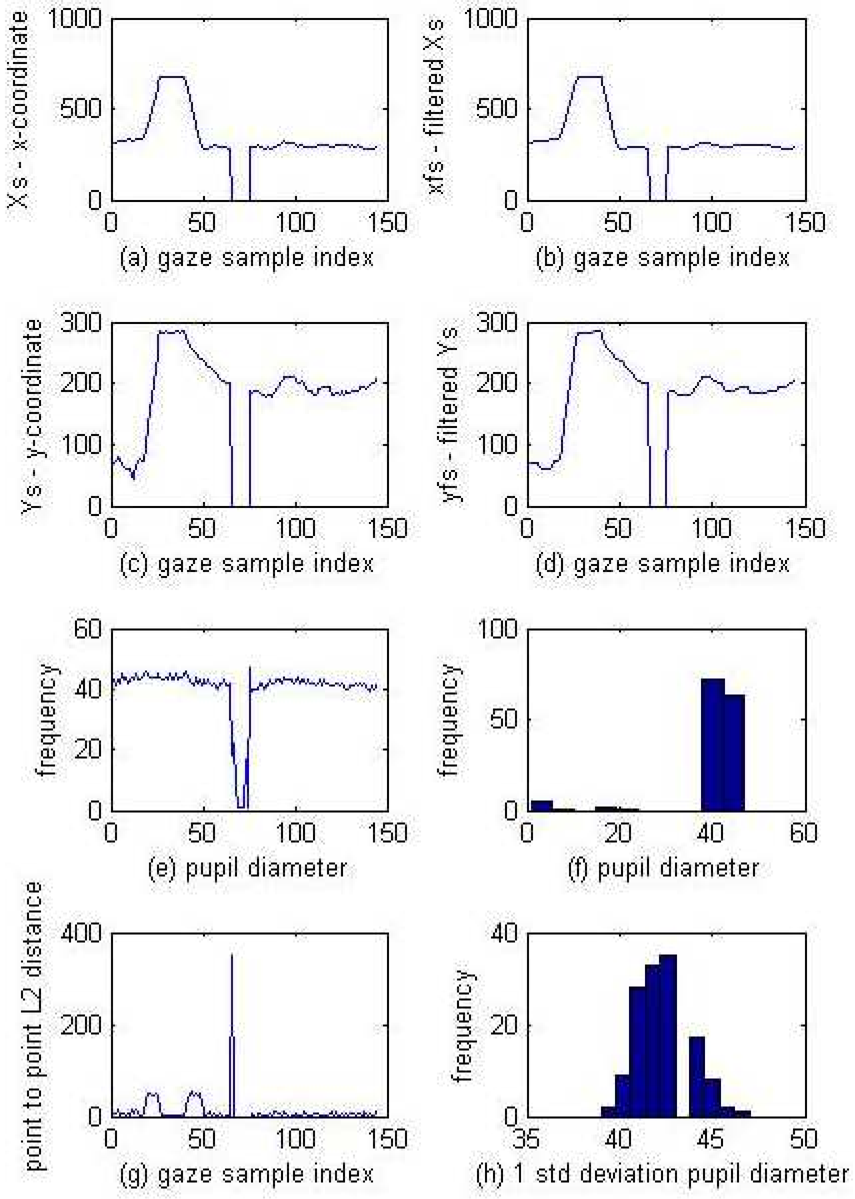


Figure 45 Good Scanpath Gaze Interpolation

Figure 45(e, f, and h) shows the pupil diameter while the Figure 45g shows the point to point distance. As can be observed from Figure 45e, all the pupil diameter values are very well distributed within a small range of values. It indicates that the eye is wide open (except in blink) during the entire scanpath consistently and the eye tracker has been able to track the eye efficiently.

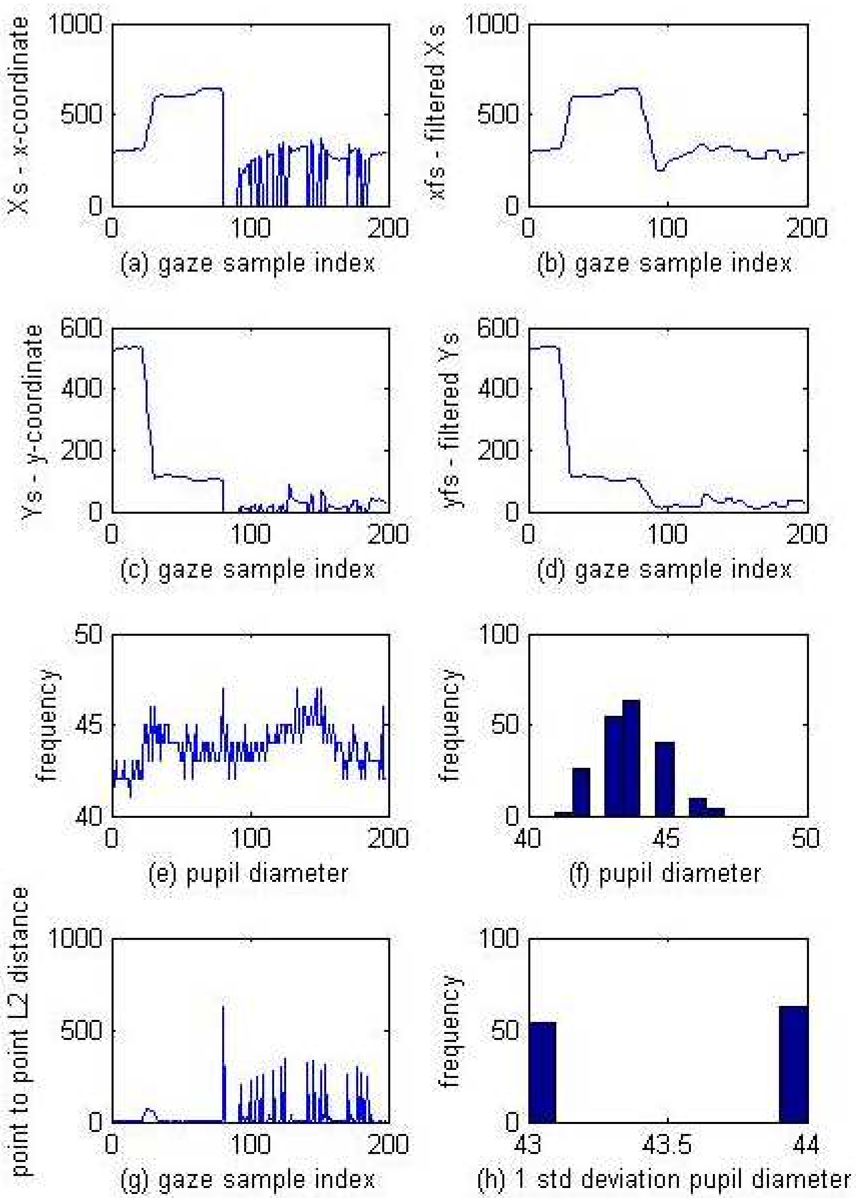


Figure 46 Bad Scanpath Gaze Interpolation

Figure 46 (a, c) shows that many (0, 0) coordinates received for eye measurements during this scanpath. The filtered result in Figure 46 (b, d) shows the effectiveness of filtering the scanpath to eliminate the (0, 0) values. This large number of (0, 0) values can be attributed to either the subject movements or equipment errors where the eye tracker is not able to measure the eye position accurately. The (0, 0) values in the raw data have been interpolated to avoid an impact on fixation computations. Figure 46 shows the pupil diameter statistics in (e, f, and h) while the point to point distance (Figure 46g) indicates the point to point L2 distance.

Figure 47 and Figure 48 show similar curves for another bad scanpath when the gaze data is interpolated during blinks. Notice that the filtered curves in Figure 47(b, d) do not include any (0, 0) values in the (x, y) coordinates. The last dip in the Figure 47(a, c) which corresponds to the blink (as can be verified from Figure 47a) has also been interpolated.

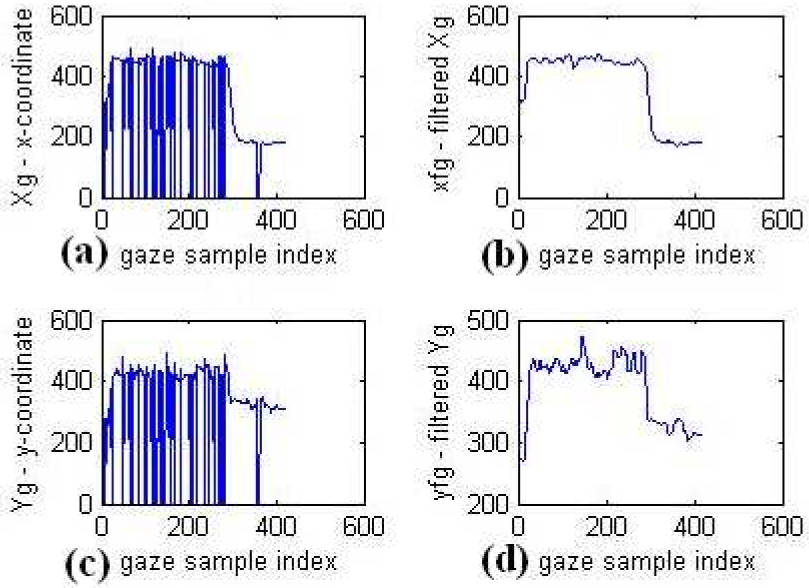


Figure 47 Bad Scanpath Gaze Interpolation (blink interpolated)

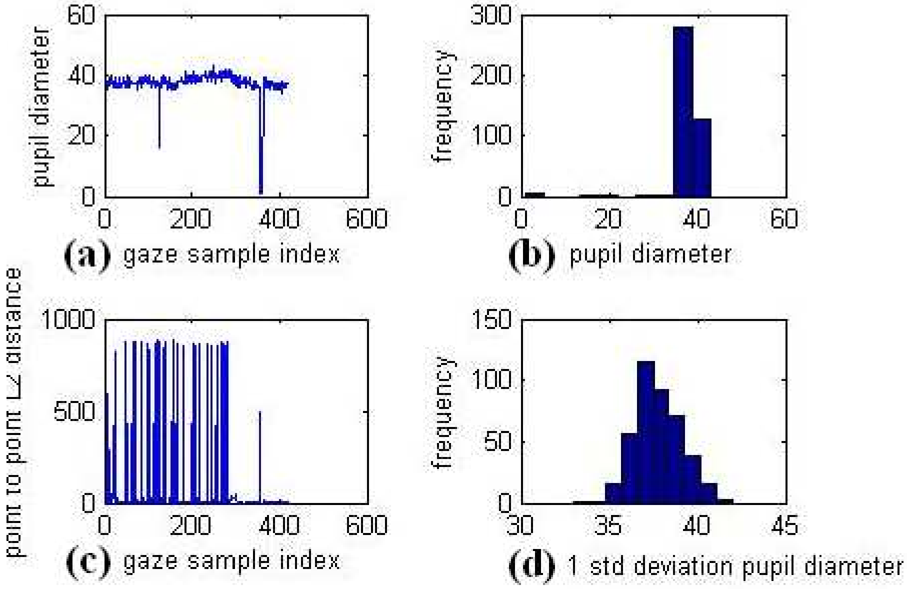


Figure 48 Bad Scanpath Parameters (blink interpolated)

Figure 47 and Figure 48 are illustrated to indicate how interpolation affects blinks if blinks were to be interpolated. But the data analysis is performed without interpolating blinks to model the underlying physical process as accurately as possible.

A good scanpath (*i.e.*, small number of <0, 0> gaze samples) or bad scanpath (*i.e.*, very large number of <0, 0> gaze samples) after interpolation becomes usable for fixation computations. Each scanpath consists of various events as illustrated in Figure 49. The events file, *<subject>.txt,* generated by the experiments is analyzed by *EventAnalyzer* to create scanpath samples. Each scanpath contains 4 data segments (*Marker Display / Dismissal* segment, *Word(s) Display / Speech Start* segment, *Speech Start / Stop* segment, and *Recognizer Finalization* segment) and each subject’s scanpath samples are analyzed to produce various internal data structures.

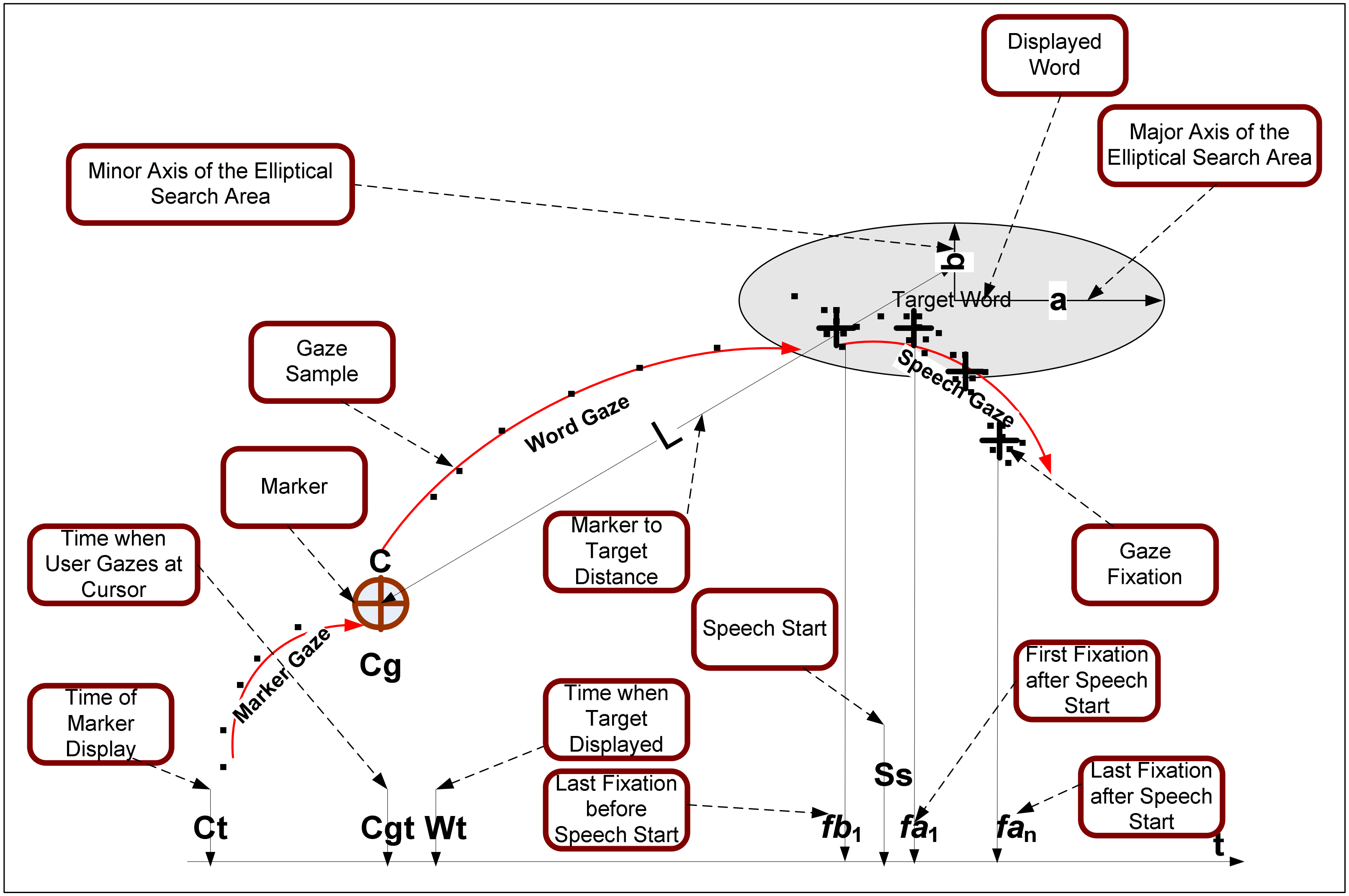

Displayed Word
Major Axis of the Elliptical Search Area
Minor Axis of the Elliptical Search Area
b
Target Word
a
Gaze Sample
Word Gaze
Speech Gaze
Marker
Marker to Target Distance
Gaze Fixation
Time when User Gazes at Cursor
C
Cg
Speech Start
First Fixation after Speech Start
Marker Gaze
Time of Marker Display
Time when Target Displayed
Last Fixation before Speech Start
Ss
Last Fixation after Speech Start
Ct
Cgt
Wt
fb1
fa1
fan
t

Figure 49 Scanpath Detailed Analysis

E.7. Search Discriminants

Target objects can be different in size and location in an application interface. Instead of the target object properties, a search region on the screen is established to determine if the gaze is in that search region. There are 6 different search regions or discriminants defined in the data analysis namely: small rectangle, big rectangle, small circle, big circle, small ellipse, and big ellipse. Small rectangle is the tightest rectangle fitting the target object while the big rectangle is bigger by 100% in height and 50% in width. Small circle is chosen to have 50 pixels radius while big circle is double that of small circle. The small ellipse bounds the target object from outside. The big ellipse leaves some gap around the target object. Big exscribed (*i.e.*, bounding the target object from outside) ellipse has been the most widely used discriminant throughout the data analysis. For each of the scanpaths, ***(x, y)*** of the gaze feature and predicted location of gaze are tested to check if it falls inside the search region defined by these discriminants. If the point is inside the search region it is considered to be a hit otherwise a miss. For each subject, a probability of prediction is calculated using the number of scanpaths inside the search region out of the total number of scanpaths. Figure 50, Figure 51, and Figure 52 illustrate how the big ellipse circumscribes the small, medium, and large words respectively. For the small word it has a little more room for fixations to fall within the search region than for the large word. However, the big ellipse is big enough to capture the fixation falling in the search region for all target objects because it accounts for the object dimensions. It provides a tight search region without overlapping on the nearby objects, thus helping build effective user interfaces without wasting much of the screen space. Table 35 and Table 36 show the detailed descriptions of the discriminants and the parameters involved.

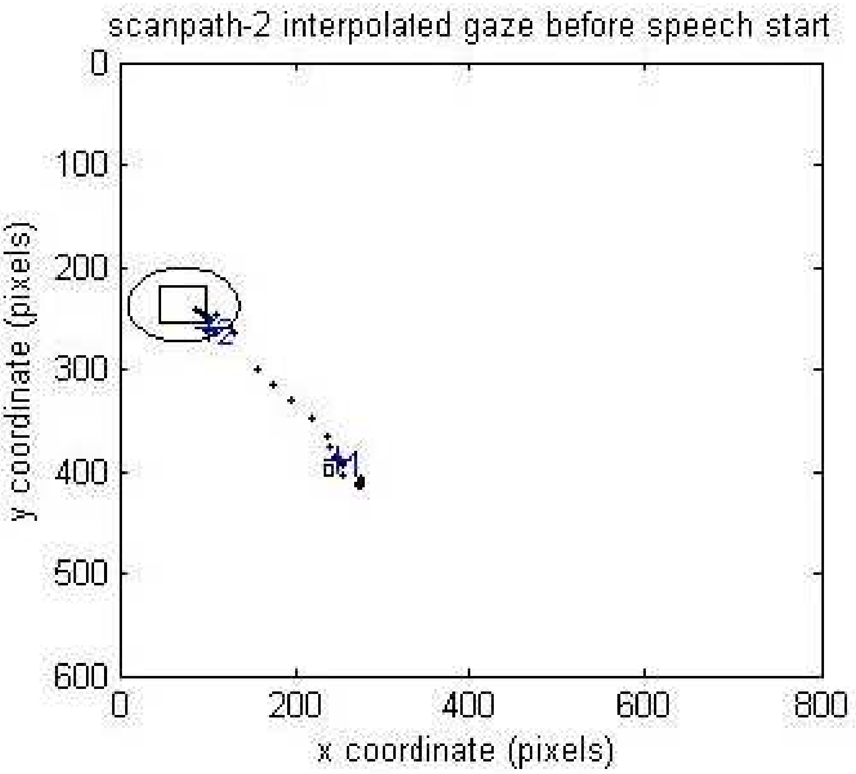


Figure 50 Scanpath Detailed Analysis of Small Word

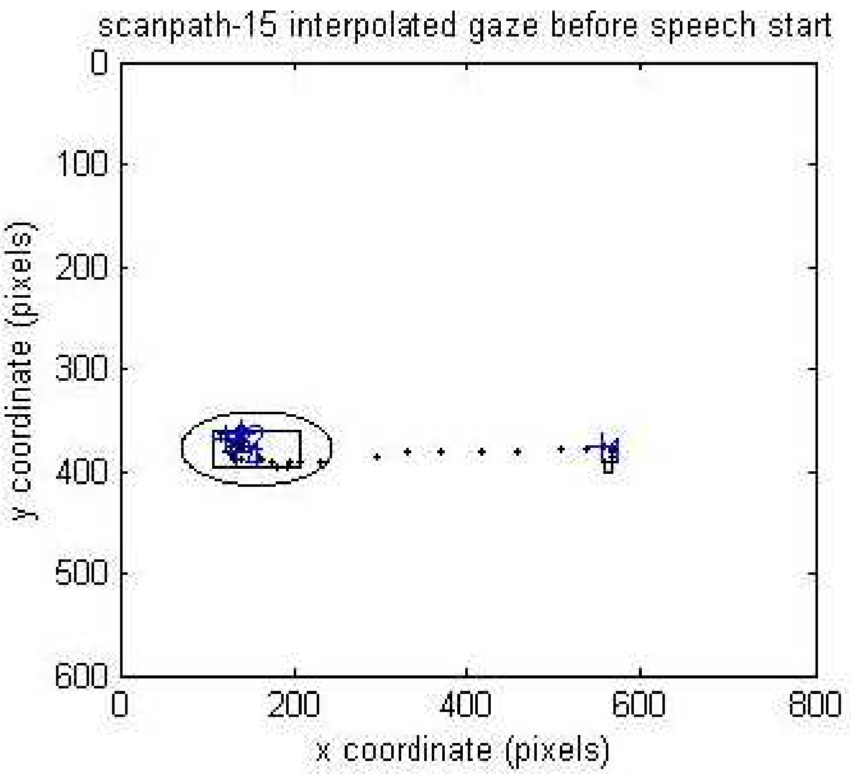


Figure 51 Scanpath Detailed Analysis of Medium Word

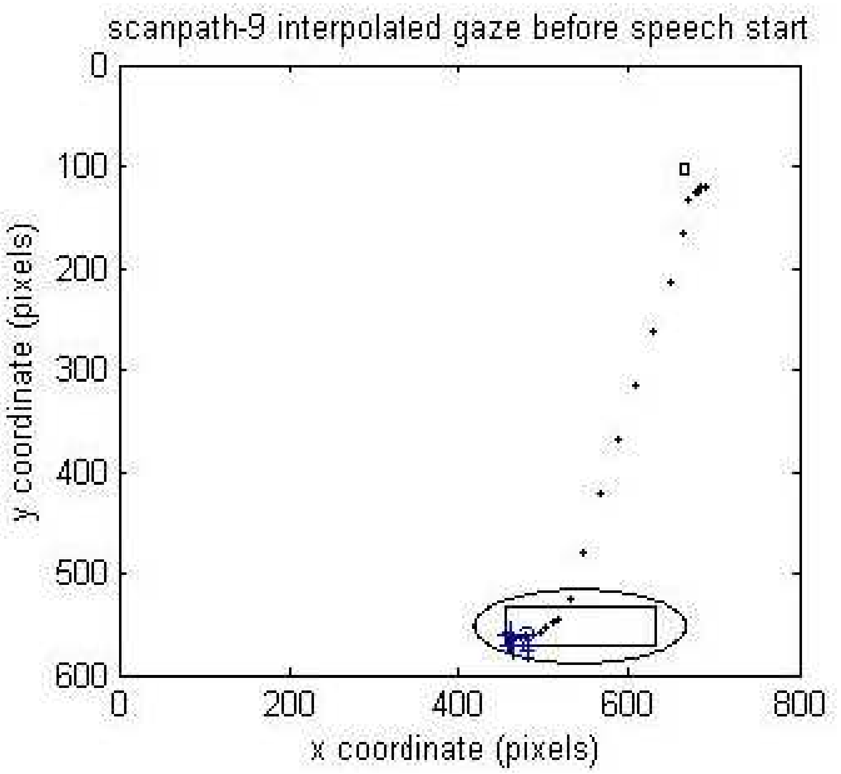
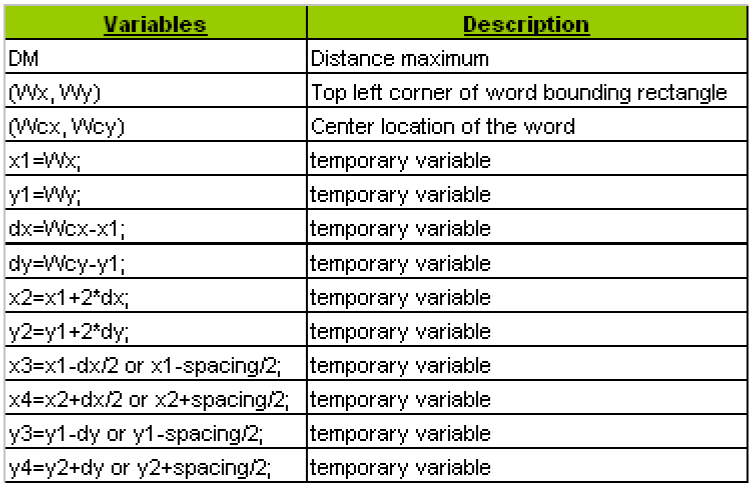
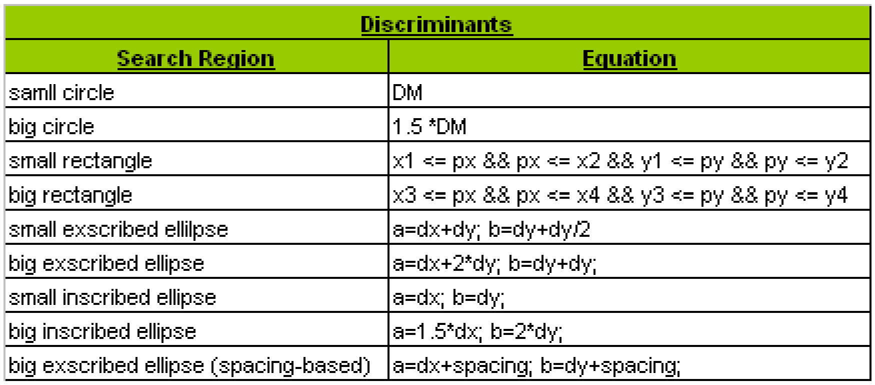


Figure 52 Scanpath Detailed Analysis of Large Word



Variables	Description
DM	Distance maximum
(Wx, Wy)	Top left corner of word bounding rectangle
(Wcx, Wcy)	Center location of the word
x1=Wx;	temporary variable
y1=Wy;	temporary variable
dx=Wcx-x1;	temporary variable
dy=Wcy-y1;	temporary variable
x2=x1+2*dx;	temporary variable
y2=y1+2*dy;	temporary variable
x3=x1-dx/2 or x1-spacing/2;	temporary variable
x4=x2+dx/2 or x2+spacing/2;	temporary variable
y3=y1-dy or y1-spacing/2;	temporary variable
y4=y2+dy or y2+spacing/2;	temporary variable

Table 35 Search Region Discriminants’ Variables



Discriminants	
Search Region	**Equation**
samll circle	DM
big circle	1.5 *DM
small rectangle	x1 <= px && px <= x2 && y1 <= py && py <= y2
big rectangle	x3 <= px && px <= x4 && y3 <= py && py <= y4
small exscribed ellilpse	a=dx+dy; b=dy+dy/2
big exscribed ellipse	a=dx+2*dy; b=dy+dy;
small inscribed ellipse	a=dx; b=dy;
big inscribed ellipse	a=1.5*dx; b=2*dy;
big exscribed ellipse (spacing-based)	a=dx+spacing; b=dy+spacing;

Table 36 Search Region Discriminants

E.8. Interaction Scanpaths

Fixations, instead of raw gaze samples, are a better estimate of the user’s gaze when analyzing scanpaths. Moreover, using raw gaze samples is error prone due to equipment problems and calibration inaccuracies. So each scanpath produces a set of fixations, which are indexed based on speech onset time, which are plotted as shown in Figure 53 through Figure 58. The fixations obtained depend on the fixation algorithm and these figures show dispersion based fixations along with the raw gaze data in each plot. Each scanpath is analyzed based on the raw gaze data (Figure 53, Figure 54, and Figure 55) or interpolated gaze data (Figure 56, Figure 57, and Figure 58) for calculating fixations. The gaze data of a scanpath is split into three sections: gaze data before the word/menu is displayed (Figure 53 and Figure 56), gaze data after the word is displayed and before the speech starts (Figure 54 and Figure 57), and gaze data after the speech started (Figure 55 and Figure 58). Similarly, Figure 59 through Figure 64 illustrates a sample scanpath in the menu interaction task in Experiment 2. Note that Figure 63 has several gaze samples interpolated when compared to Figure 60.

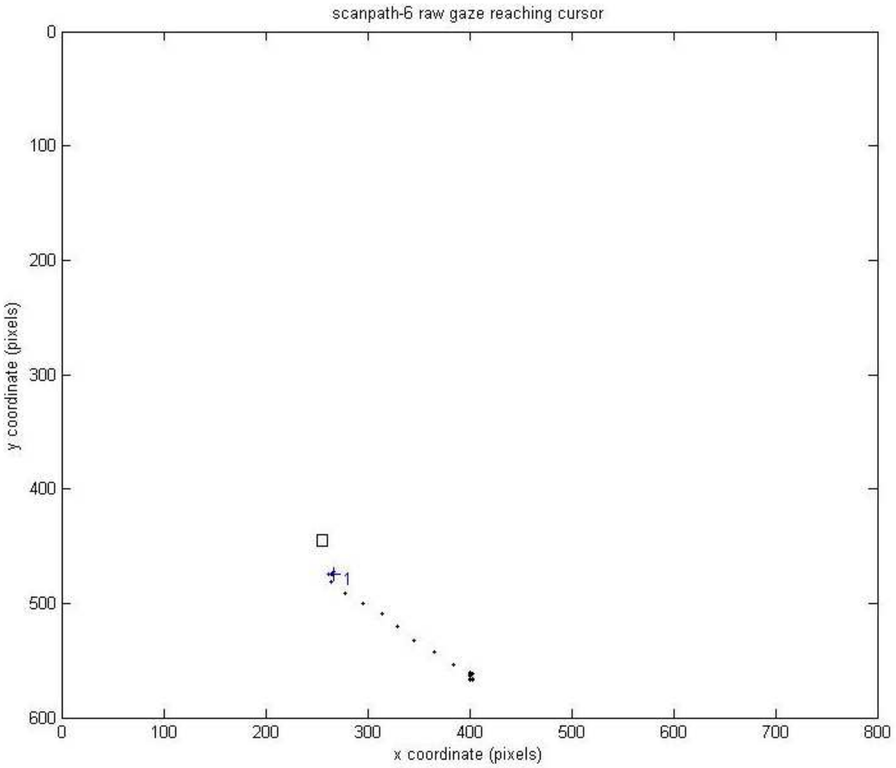


Figure 53 Scanpath of Word Reading Experiment – raw gaze reaching cursor

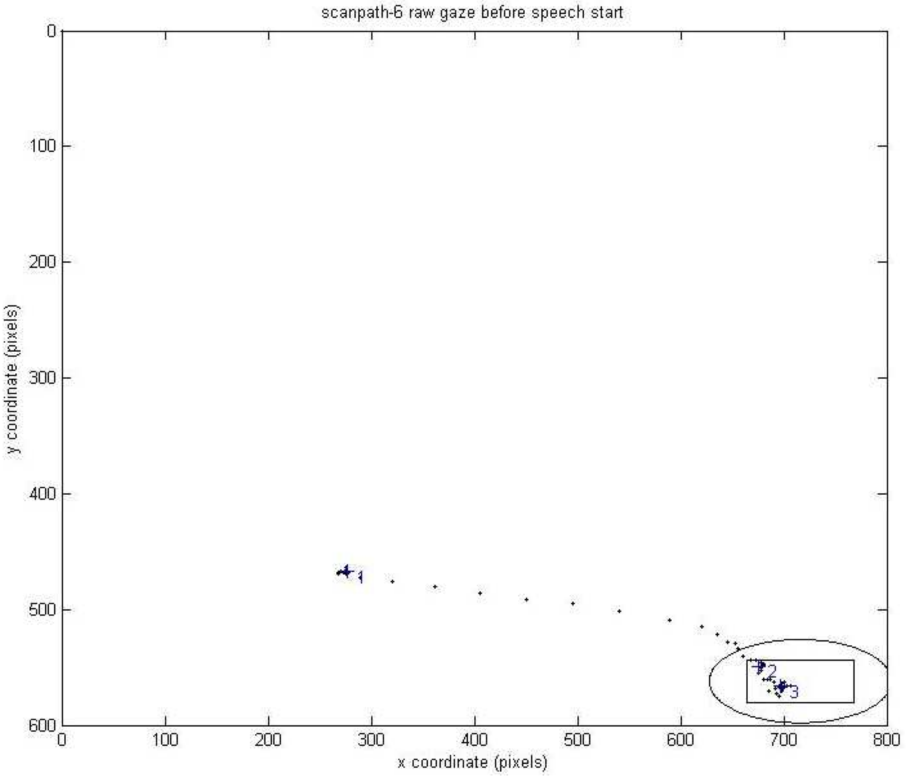


Figure 54 Scanpath of Word Reading Experiment – raw gaze before speech start

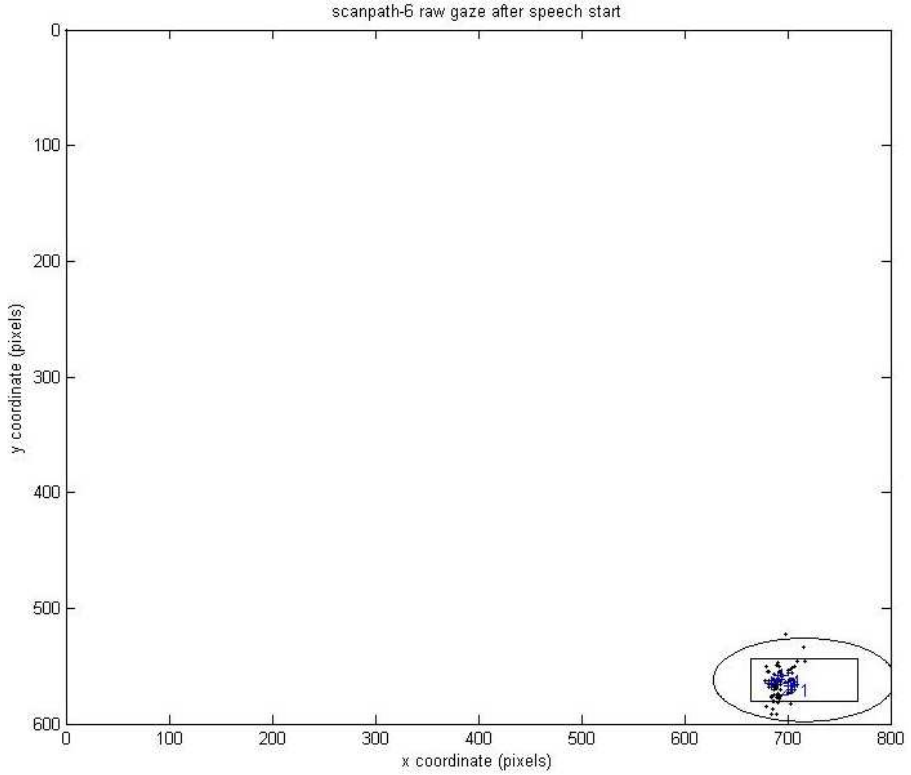


Figure 55 Scanpath of Word Reading Experiment – raw gaze after speech start

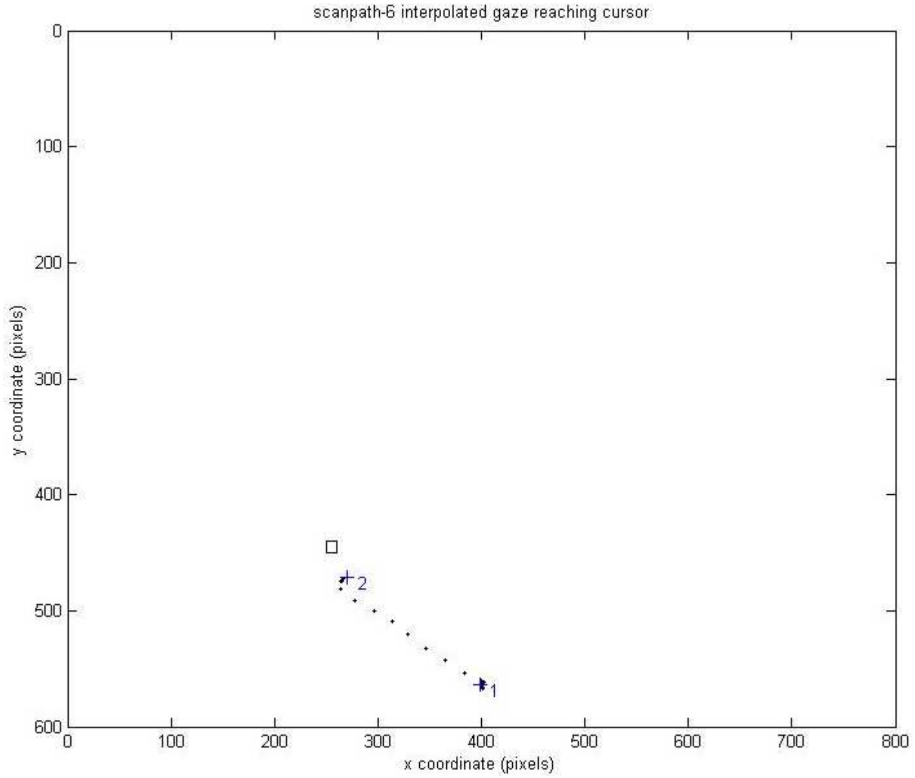


Figure 56 Scanpath of Word Reading Experiment – interpolated gaze reaching cursor

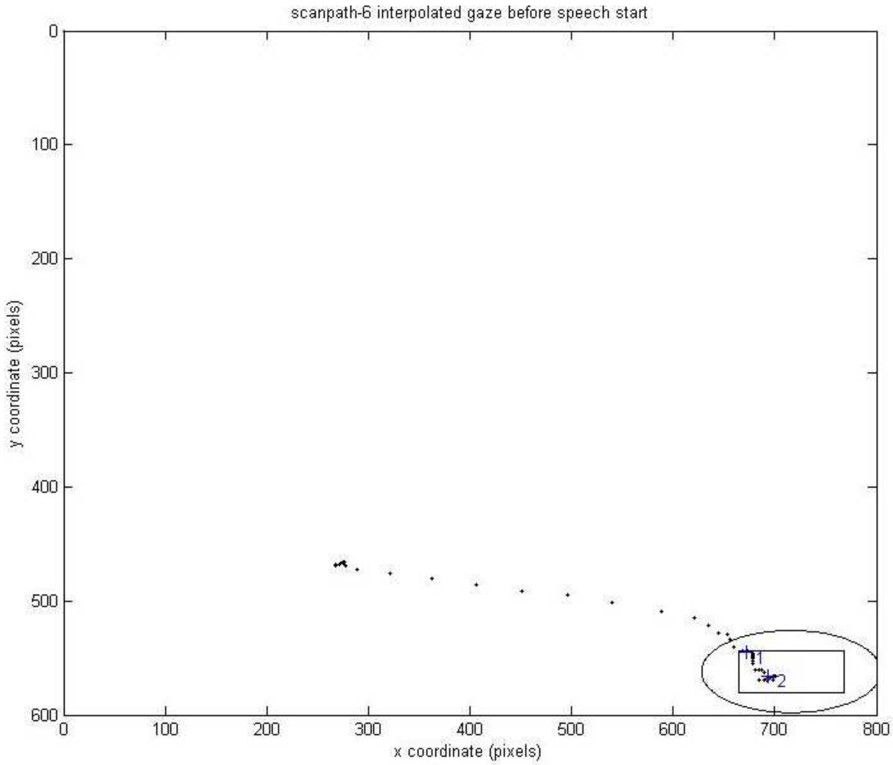


Figure 57 Scanpath of Word Reading Experiment – interpolated gaze before speech start

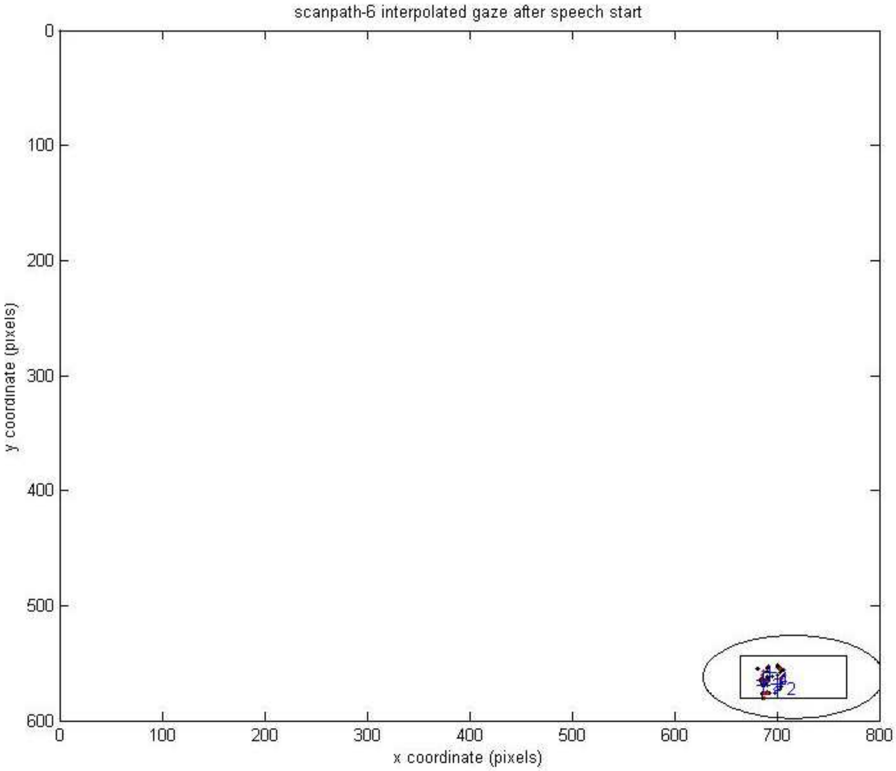


Figure 58 Scanpath of Word Reading Experiment – interpolated gaze after speech start

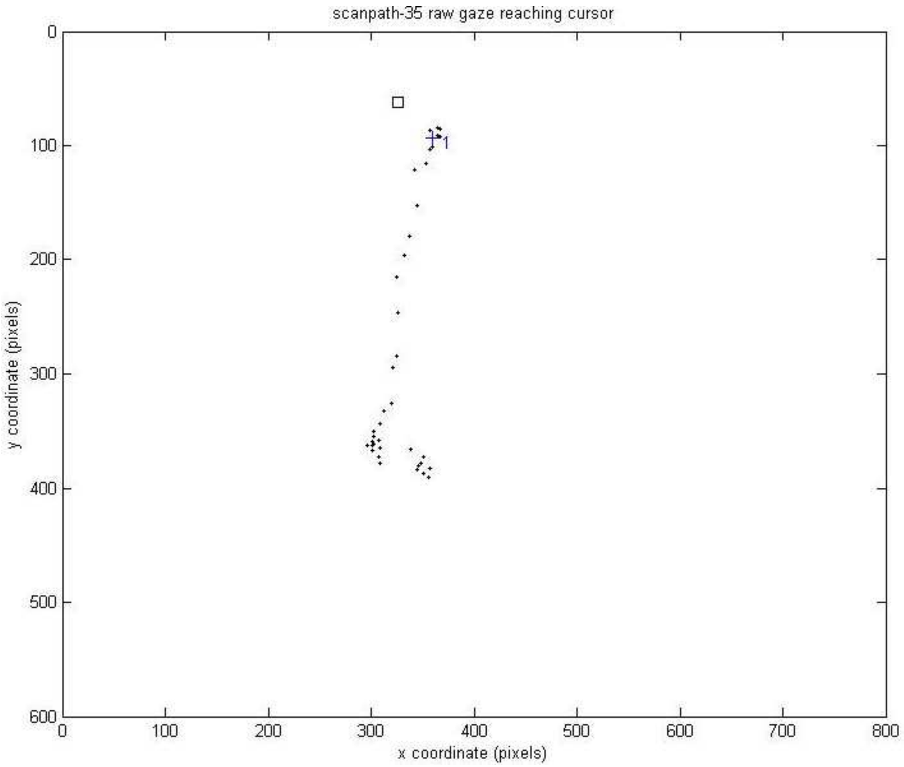


Figure 59 Scanpath of Menu System Interaction – raw gaze reaching cursor

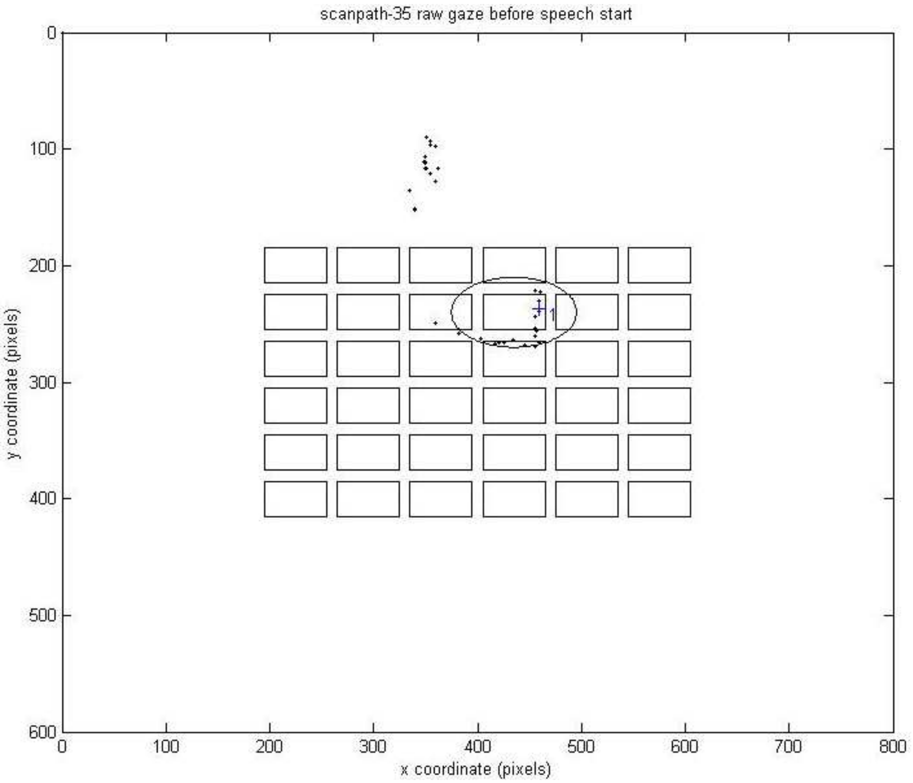


Figure 60 Scanpath of Menu System Interaction – raw gaze before speech start

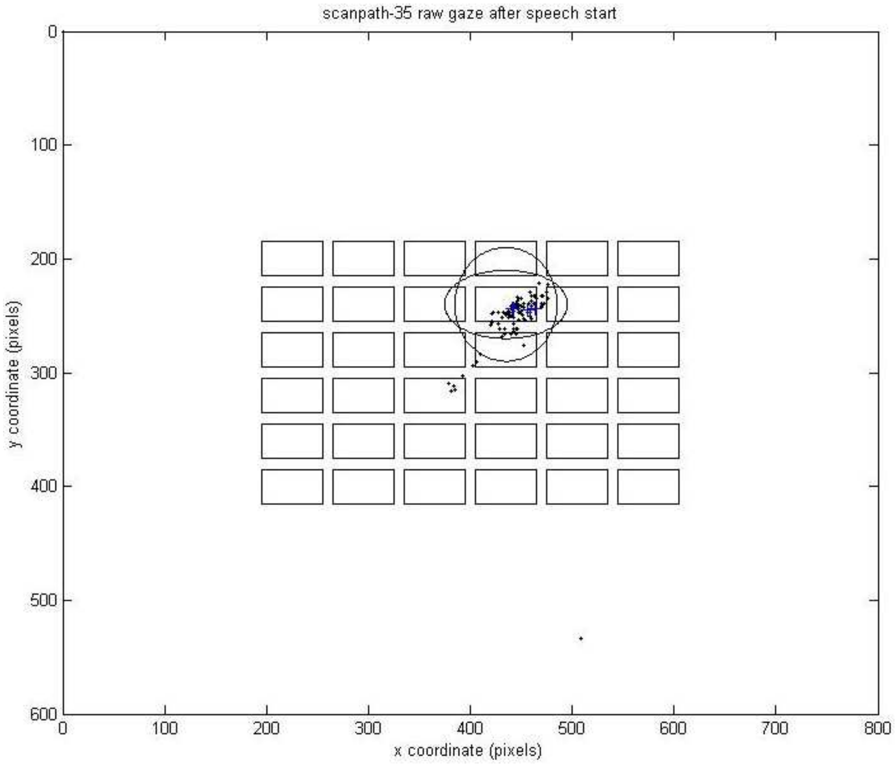


Figure 61 Scanpath of Menu System Interaction – raw gaze after speech start

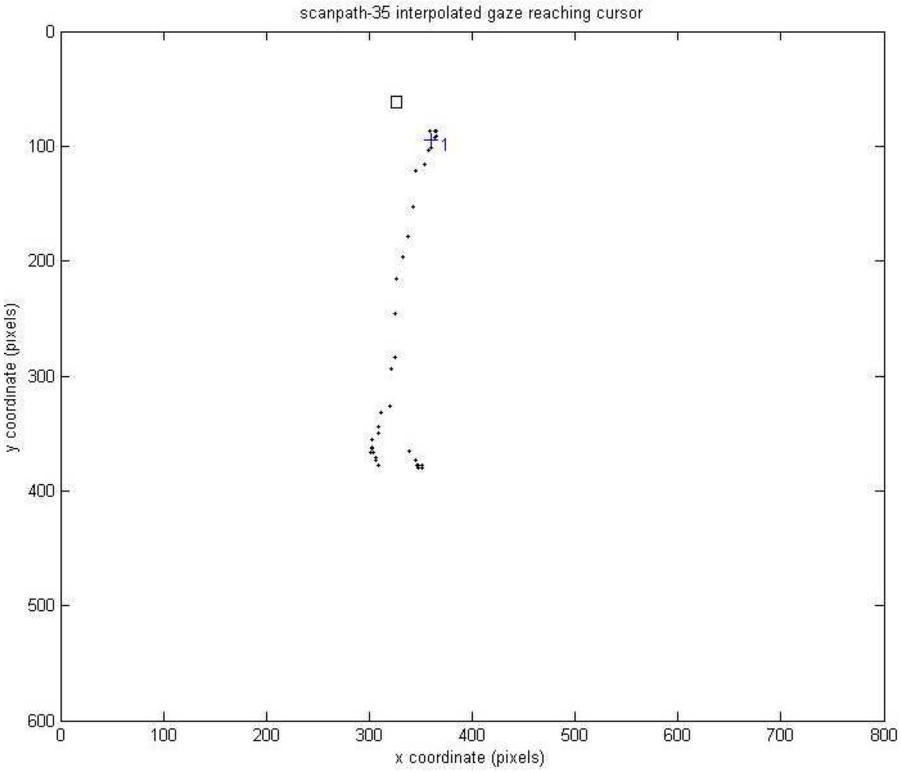


Figure 62 Scanpath of Menu System Interaction – interpolated gaze reaching cursor

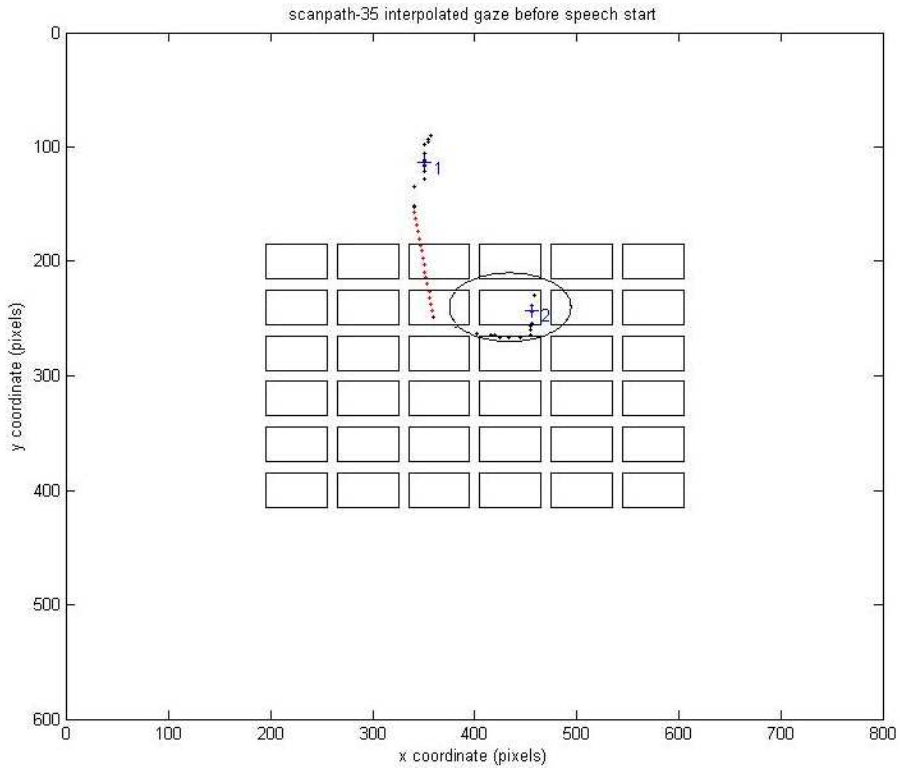


Figure 63 Scanpath of Menu System Interaction – interpolated gaze before speech start

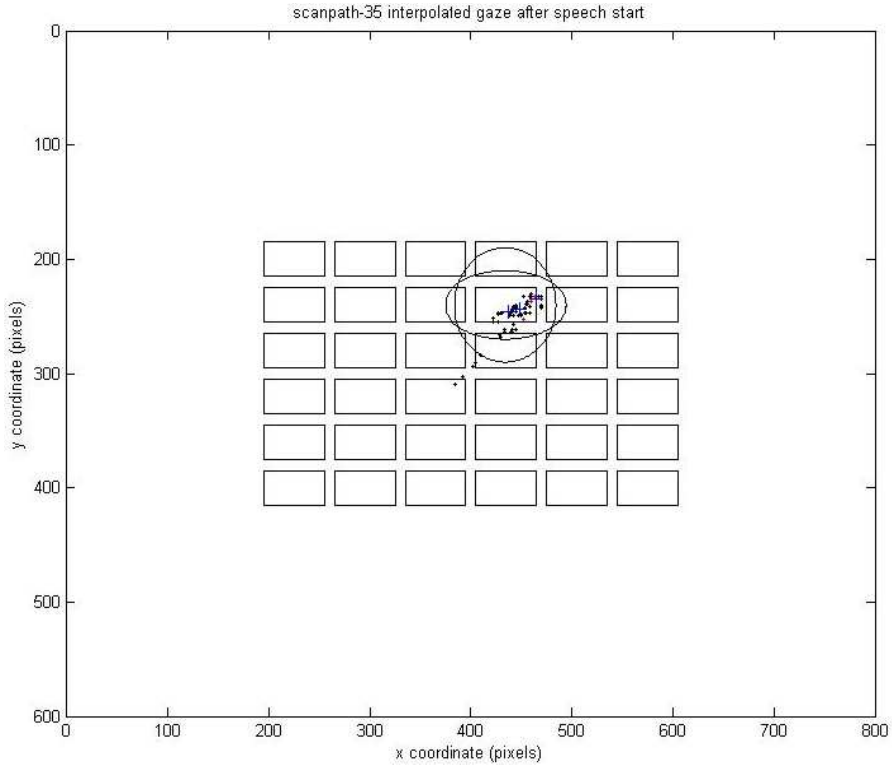
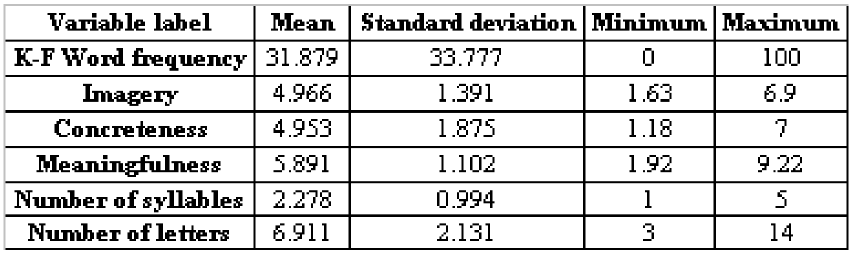


Figure 64 Scanpath of Menu System Interaction – interpolated gaze after speech start

E.9. Experiment Limitations/Design Choices

In this section, some of the main limitations and design choices pertinent to the speech/gaze integration model will be described. These are applicable equally to both experiments. First, the words selected for both the experiments are analyzed for their characteristics. Second, the effect of speech training on recognition accuracy is discussed. Third, gaze calibration impact on eye tracking is discussed. Finally, the importance of obtaining the accurate timestamps from the speech recognizer is emphasized.

Choosing words for the subjects to utter, is an important criterion in the speech/gaze interaction behavior. If the word is familiar to the subject, then it may be possible for the subject to use peripheral vision to read the word quickly. When the word is not known, the user has to look at the word before reading it. The time to read the word depends on the complexity of the word being displayed. It is hypothesized that the more complex and unfamiliar a word is, the longer the gaze span is around the word. But only the speech onset time alone is of interest, which reflects gaze location around the word. It is not clear how the onset time is affected depending on the word though. So a mix of multi-syllable words was chosen in the experiment for even distribution of the interaction onset times.



Variable label	Mean	Standard deviation	Minimum	Maximum
K-F Word frequency	31.879	33.777	0	100
Imagery	4.966	1.391	1.63	6.9
Concreteness	4.953	1.875	1.18	7
Meaningfulness	5.891	1.102	1.92	9.22
Number of syllables	2.278	0.994	1	5
Number of letters	6.911	2.131	3	14

Table 37 Words Statistics

Table 37 shows the word statistics for words chosen for Experiment 1 [Appendix B]. These words are generated using a word list generator [132]. *K-F Frequency* refers to the word frequency as defined in "*Computational Analysis of Present-Day American English*" published in 1967 by Henry Kucera and Nelson Francis. They analyzed and compiled several statistics on about a million words collected from various sources. *Correctness* refers to whether the word has any semantic meaning associated with it. *Imagery* measures the extent of associating a word to any image in cognitive memory. *Meaningfulness* measures the ease with which a word can be learned.

The length of words (chosen in this dissertation) in pixels and the number of characters are illustrated in Figure 65. Note that the word list for Experiment 2 is a subset of Experiment 1. Figure 65(a, c) shows the length of words in pixels and number of characters for all users individually. Figure 65(b, d) shows the average length of words in pixels and number of characters of all users.

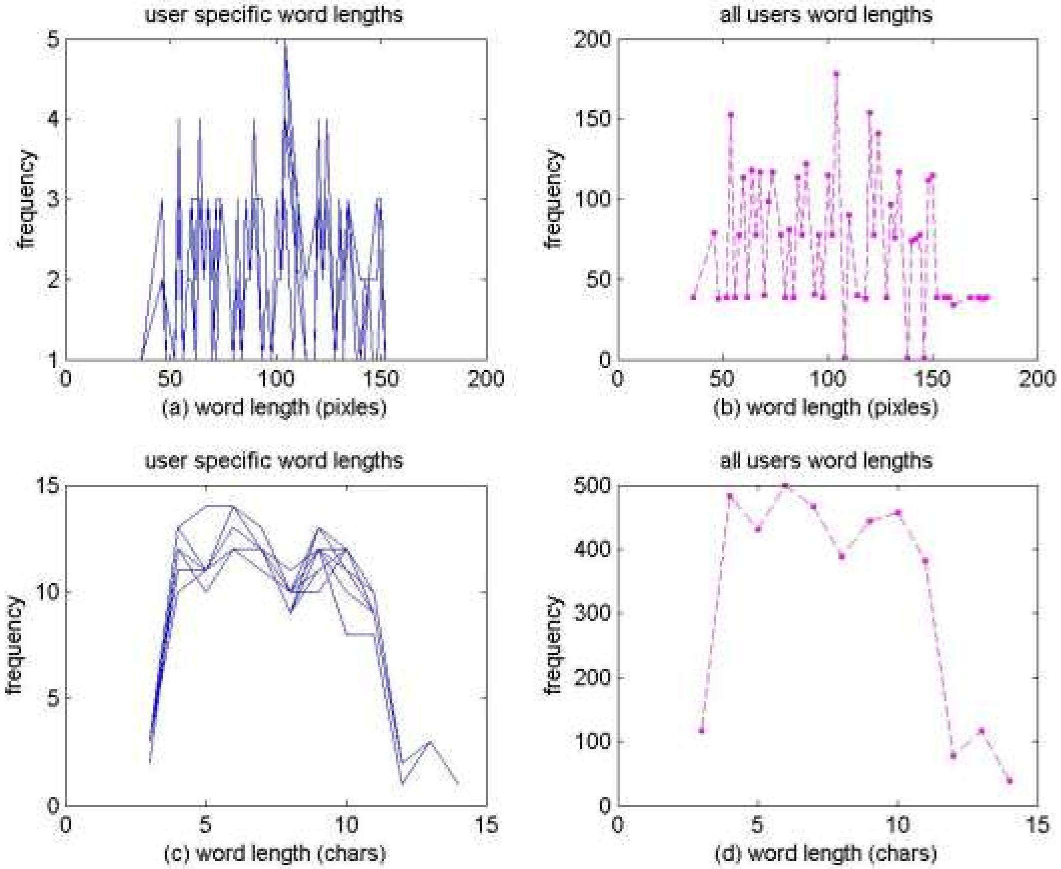


Figure 65 Word Length Statistics in Constraint-Free Interaction

E.9.1. Effect of Speech Training/Recognition

In current speech and gaze experiments, the speech recognizer issues various kinds of events indicating the state changes inside the recognizer. These events help applications determine how to process the information coming from the recognizer. For example, events like *SpeechStarted*, *SpeechStopped*, and *SpeechAccepted* are very critical in data processing. The following illustrate different scenarios that exist in the current speech/gaze experiments.

- Displayed word and the word recognized by the recognizer are the same. This is a perfectly valid interaction. Gaze and speech can reinforce each other for more effectiveness of the interface. Note that there is no guarantee that gaze is really looking at the word being spoken, but it can help reinforce a valid interaction in a noisy environment.
- Recognizer recognizes the acoustic signal as a valid word in the grammar but it is not the current word shown on the screen i.e., the displayed word is “run” and recognizer heard it as “trumpet” from the signal processing. This is an “Accepted Successful Recognition” from the recognizer point of view but it is not the word on the screen. Gaze in this case can help verify what’s on the screen and whether it is same as what’s heard. Thus, “*gaze aids speech*”.
- Recognizer totally doesn’t understand the word or it rejects the acoustic signal for not being able to process. Then the only information available is gaze pointing at the command button. Thus, gaze aids speech.
- Recognizer couldn’t give any conclusive answer other than saying when speech started and ended. These kinds of scanpaths can also be used because the speech start and end event times are still available from API events (but these are not accurate because the recognizer didn’t finalize them with an Accept/Reject). But in this case these times are not going to be accurate because they are not the timestamps as

> perceived by the recognizer. However, they can be used for an approximation of the start/end times of speech to make the interface more robust. Thus, gaze aids speech.

The recognizer issues events to the application when speech starts and ends. When an event like this is received, it is only an approximation from the engine for that event. When the engine finalizes the acceptance/reject of the utterance as a valid/invalid word, it ‘corrects’, internally, the timestamps and provides more accurate timestamps. Sometimes the engine can be simply quiet and may not issue the speech start/end events. It doesn’t even give the accept/reject finalization events. Basically the recognizer doesn’t give any information for that interaction/scanpath sample. Then there is no way of knowing anything about that utterance/interaction. It can happen because of the recognizer’s implementation. Even with a better recognizer these kinds of things *may* potentially happen. These scanpaths cannot be used in the data analysis at all because it is insufficient information to construct a scanpath as the corresponding speech start time is not available to analyze the fixations. Currently these are being filtered out.

In the current speech/gaze experiments, each subject speaks around 100-250 words in Experiment 1 [Appendix B] and 400 words in Experiment 2 [Appendix C]. Not all the words may be successfully recognized by the speech recognizer. Although recognizers claim to be very efficient in recognizing the user’s speech, the successful recognition depends on several factors like pronunciation, accents, ambient conditions *etc*. Some recognizers claim zero training time while others claim minimal training time. However, IBM Via Voice, the recognizer that is used in the speech/gaze interaction experiments, has given reasonable recognition results among all the recognizers that have been tried (Via Voice, SpeechWorks, MS Speech API). Dragon’s naturally speaking, HTK/ATK and Sphinx are among other recognizers which need to be explored in the future to understand the speech recognizer’s influence on the speech/gaze interaction. Even though Via Voice claimed to be recognizing with the least training time, experience has shown that for best results, it requires a minimum of 30 mins to about 3 hours of training time.

Sometimes it’s not even the time spent in training that determines the recognition accuracy. It is mostly to do with the quality of training and the distribution of words in the training text. Each subject is trained with the same text from the recognizer before running the experiments. The following factors need to be considered when a recognition error occurs.

- Should the experiment make a second recognition attempt?
- How many recognition retry attempts should be used? What will be the effect of these retries in speech/gaze interaction?
- Will the fixation formation and speech/gaze interaction be affected by the recognition errors?

If the word is recognized successfully, the word disappears from the screen immediately starting the next interaction. However, when a word is not recognized, it is not a failed interaction. Speech onset time captured in a failed (i.e., not-recognized successfully by the speech recognizer) interaction is enough to understand the speech/gaze interaction behavior. The problem in further attempts to recognize the word is that it may not give accurate interaction behavior. Moreover, it is not necessary that the gaze is around the word after the first attempt because the user is familiar with the word after the first attempt. Only during the first attempt is the cognitive processing triggered which captures the speech/gaze interaction accurately. During subsequent attempts, gaze may be jumping off to a different location giving a saccade before coming back to the word. Capturing the speech onset time during the first attempt accurately reflects speech/gaze interaction regardless of the speech recognition success/failure. Currently a word appearing on the screen disappears if the recognition is successful or if the word it timed out. The timeout depends on the task being performed. For simple tasks like in Experiment 1 timeout is 3 seconds and for complex tasks like in Experiment 2 the timeout is 15 seconds. The system’s dependency on speech onset time instead of speech recognition accuracy reduces the need for higher accuracy of speech recognition and also allows for the system to be usable under less constrained environments.

E.9.2. Effect of Gaze Calibration

The speech/gaze interaction behavior depends more heavily on the gaze data being accurate than on the speech because of the high sensitivity of gaze over speech. Also, gaze is much faster than speech and is difficult to track. In order to obtain the gaze data accurately, gaze needs to be calibrated carefully and accurately. Even after careful calibration, gaze can drift over the time due to data acquisition problems with the equipment. Also, the users may be moving their head position when interacting with the machine which can offset the gaze. In order to mitigate gaze errors and to calculate error in gaze data, each subject looks at a few scattered points on the screen after completing the gaze calibration. The user looks at these predefined points on the screen and the system calculates the gaze accuracy. Although it can not be established that the user is really looking at the predefined point, it can at least compute the gaze calibration and tracking errors.

Gaze can drift over time due to several factors like equipment correction, user movements, *etc*. Its accuracy may decrease as the experiment runs for longer intervals. In order to mitigate the effect of gaze degradation, each session is run independently calibrating gaze every time (Note that there is an ongoing development effort among eye tracker suppliers to simplify and improve the calibration process). Each session is allowed to run a maximum of 10-15 minutes. The session length depends on the recognition accuracy and the subject’s ability to complete the task. The break between sessions also provides enough rest to the subject reducing any fatigue.

E.9.3. Effect of Word Timestamps on Fixation Identification

Typically, speech recognizers use pauses between user utterances to extract recognition results. The current IBM Via Voice recognizer issues events to applications in the beginning of the utterance and at the end of utterance (Figure 66a). In addition to these events, it issues other recognition events (e.g., *SpeechAccepted*) indicating the recognition results. The lead-time for gaze with respect to the *SpeechStarted* event is relatively consistent for a given user, and it also varies from user to user [19]. For example, for commands like “put it there” which contain multiple words, Figure 66a shows the gaze fixations and recognizer events. The word timestamps from the recognizer and fixations corresponding to the words are retrieved *after* the end of the complete utterance. By searching the closest fixation before the word timestamp, fixations 3, 6, and 8 could be retrieved for words “put”, “it”, and “there” respectively. However, fixations 3 and 8 are not completely formed at the beginning of the *word utterance* as shown in the Figure 66b. A simple search back in time at the end of the complete utterance can yield incorrect fixations, as compared to more accurate fixations 2, 5, and 7 shown in Figure 66b.

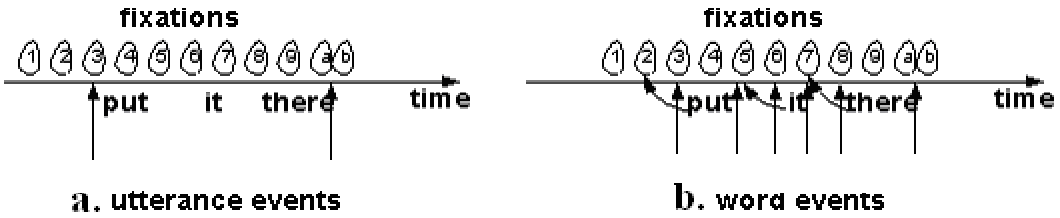


Figure 66 Recognizer Word Timestamp Analysis

Moreover, fixations closely spaced in time don’t necessarily mean that they are located closely in space. Any error in fixation identification can yield an error in ascertaining the user’s intent. The modality integration process may not think that gaze and speech are aligned and will prompt the user for a clarification. In commands like “put it there” which contain deictic references, the word timestamps for “it” and “there” are highly important to extract the referential information pointed to by gaze. Thus, word timestamps retrieved from the speech recognizer have very high influence on the fixations associated with the words. Both Experiments 1 and 2 use only single words for targets and multi-word commands need to be explored in the future to resolve ambiguities between total utterance timestamps and individual word timestamps.

REFERENCES

[1] Manpreet Kaur, Marilyn Tremaine, Ning Huang, Joseph Wilder, Zoran Gacovski, Frans Flippo, Chandra Sekhar Mantravadi, "*Where is "it"? Event Synchronization in Gaze-Speech Input Systems,*" International conference on Multimodal interfaces, 2003, Vancouver, Canada, Pages: 151 – 158.

[2] L. R. Rabiner, "*A tutorial on hidden Markov models and selected applications in speech recognition,*" Proceedings of the IEEE, vol. 37, no. 2, pp.257-86, February 1989.

[3] L.R. Rabiner and B-H Juang: *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[4] L.R. Rabiner and S.E. Levinson, "*Isolated and Connected Word Recognition – Theory and Selected Applications,*" IEEE Transactions on Communications, vol. COM-29, no. 5, pp. 621-69, May 1981.

[5] Frans Flippo, Allen Krebs, Ivan Marsic, "*A Framework for Rapid Development of Multimodal Interfaces*", ICMI 2003, November 5-7 2003 Vancouver, Canada.

[6] http://faculty.washington.edu/chudler/nsdivide.html.

[7] http://www.sirinet.net/~jgjohnso/nervous.html.

[8] Marc O. Ernst, "*The 'Puzzle' of Sensory Perception: Putting Together Multisensory Information*", ICMI 2005, October 4-6, 2005, Trento, Italy.

[9] Jullien Bouchet, Laurence Nigay, Thierry Ganille, "*ICARE Software Components for Rapidly Developing Multimodal Interfaces*", ICMI 2004, pp. 251-258, October 13-15, State College, Pennsylvania, USA.

[10] Glass et al., "*A Framework for Developing Conversational User Interfaces*", Proceedings of CADUI'2004 (2004) 354-365.

[11] Krahnstoever et al., "*A Real-Time Framework for Natural Multimodal Interaction with Large Screen Displays*", Proceedings of ICMI'02 (2002).

[12] Oviatt, S. et al., "*Designing the User Interface for Multimodal Speech and Gesture Applications: State-of-the-Art Systems and Future Research Directions*", HCI, 15, 4 (2000), 263-322.

[13] Luca Nardelli, Marco Orlandi, Daniele Falavigna, "*A Multi-Modal Architecture for Cellular Phones*", ICMI 2004, October 13-15, 2004, State College, Pennsylvania, USA.

[14] Bee-Wah Lee, Alvin W. Yeo, "*Integrating Sketch and Speech Inputs using Spatial Information*", ICMI 2005, October 4-6, 2005, Trento, Italy.

[15] R. A. Bolt, "*Put-that-there": Voice and Gesture at the Graphic Interface*", Computer Graphics (SIGGRAPHS'80 Proceedings), 14(3):262-270, July 1980.

[16] Oleg Spakov, Darius Miniotas, "*Gaze-Based Selection of Standard-Size Menu Items*", ICMI 2005, October 4-6, 2005, Trento, Italy.

[17] Peter Gorniak, Deb Roy, "*Probabilistic Grounding of Situated Speech using Plan Recognition and Reference Resolution*", ICMI 2005, October 4-6, 2005, Trento, Italy.

[18] A. J. Cheyer and D. L. Martin, "*The Open Agent Architecture*", Autonomous Agents and Multi-Agent Systems, 4(1-2):143-148, 2001.

[19] S. Oviatt, "*Ten Myths of Multimodal Interaction*", Communications of the ACM, 42(11):74-81, 1999.

[20] Dominic W. Massaro, "*A Framework for Evaluating Multimodal Integration by Humans and A Role for Embodied Conversational Agents*", ICMI 2005, October 4-6, 2005, Trento, Italy.

[21] Niels Ole Bernsen and Laila Dybkjaer, "*Evaluation of Spoken Multimodal Conversation*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[22] Paulo Barthelmess and Clarence A. Ellis, "*The ThreadMill Architecture for Stream-oriented Human Communication Analysis Applications*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[23] G. Herzog, H. Kirchmann, S. Merten, A. Ndiaye, and P. Poller, "*Multiplatform Testbed: An Integration Platform for Multimodal Dialog Systems,"* H. Cunningham and J. Patrick, editors, In Proceedings of HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS), pp. 75-82, Edmonton, Canada, 2003.

[24] P. Oreizy, M. M. Gorlick, R.N. Taylor, D. Heimbigner, G. Johnson, N. Medvidovic, A. Quilici, D.S. Rosenblum, and A. L. Wolf, "*An Architecture-Based Approach to Self-Adaptive Software*", IEEE Intelligent Systems and Their Applications, 14(3):54-62, May/June 1999.

[25] Sharon Oviatt, Rachel Coulston, Rebecca Lunsford, "*When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[26] Zhinhong Zeng, Jilin Tu, Ming Liu, Tong Zhang, Nicholas Rizzolo, Zhenqiu Zhang, Thomas S. Huang, Dan Roth and Stephen Levinson, "*Bimodal HCI-related Affect Recognition*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[27] Andre D Milota, "*Modality Fusion For Graphic Design Applications*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[28] Vo. T, and Wood, C., "*Building an Application Framework for Speech and Pen Input Integration in Multimodal Learning Interfaces*", In Proceedings of the International Conference on acoustics speech and signal processing (IEEE-ICAASP 1996), Vol. 6, 3545-3548, IEEE Press.

[29] Kaiser, C., and Cohen, P.R., "*Implementation Testing of a Hybrid Symbolic/Statistical Multimodal Architecture*", In Proceedings of the International Conference on Spoken Language Processing (Denver, September 2002) pp. 173- 176.

[30] Johnston, M., and Bangalore, S. "*Finite-State Multimodal Parsing and Understanding*", In Proceedings of COLING-2000 citeseer.nj.nec.com/413176.html.

[31] Touraine, D., Bourdot, P., Bellik, Y., and Bolot, L.A., "*Framework to Manage Multimodal Fusion of Events for Advanced Interactions within Virtual Environments*", In Proceedings of the workshop on Virtual environments 2002, pp. 159-168.

[32] Hartwig Holzapfel, Kai Nickel, Rainer Stiefelhagen, "*Implementation and Evaluation of a Constraint-Based Multimodal Fusion System for Speech and 3D pointing Gestures*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[33] L. Wu, S.L. Oviatt, and P. R. Cohen, "*Multimodal Integration – A Statistical View*", IEEE Transactions on Multimedia, 1(4):334-341, 1999.

[34] J. Eisenstein and C. M. Christoudias, "*A Salience-based Approach to gesture-Speech Alignment*", In Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting, 2004.

[35] M. Johnston, "*Unification-based Multimodal Parsing*", In COLING-ACL, pages 624-630, 1998.

[36] I. Wachsmuth, "*Communicative Rhythm in Gesture and Speech*", In Gesture-Based Communication in Human-Computer Interaction: International Gesture Workshop, GW'99, Gif-sur-Yvette, France, March 1999.

[37] Daniel Bauer and James D. Hollan, "*IRYS: A Visualization Tool for Temporal Analysis of Multimodal Interaction*", ICMI'03, November 5-7, 2003, Vancouver, British Columbia, Canada.

[38] Norbert Pfleger, "*Context Based Multimodal Fusion*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[39] J. R. Anderson and C. Lebiere, "*The Atomic Components of Thought*", Erlbaum, Mahwah, NJ, 1998, http://act-r.psy.cmu.edu/book/.

[40] R. Travis Rose, Francis Quek, and Yang Shi, "*MacVisSTA: A System for Multimodal Analysis*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[41] S. Dusan and J. L. Flanagan, "*Human Language Acquisition by Computers*", in Proceedings of the International Conference on Robotics, Distance Learning and Intelligent Communication Systems, WSES/IEEE, Malta, 2001, pp. 387-392.

[42] Peter Pal Boda, "*A Maximum Entropy Based Approach for Multimodal Integration*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[43] Cristy Ho, "*Using Spatial Warning Signals to Capture a Driver's Visual Attention*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[44] Yeow Kee Tan, Nasser Sherkat, Tony Allen, "*Error Recovery in a Blended Style Eye Gaze and Speech Interface*", ICMI'03, November 5-7, 2003, Vancouver, British Columbia, Canada.

[45] Norbert Reithinger, Jan Alexandersson, Tilman Becker, Anselm Blocher, Ralf Engel, Markus Lockelt, Jochen Muller, Norbert Pfleger, Peter Poller, Michael Streit, Valentin Tschernomas, "*SmartKom – Adaptive and Flexible Multimodal Access to Multiple Applications*", ICMI'03, November 5-7, 2003, Vancouver, British Columbia, Canada.

[46] Eric Horvitz, Johnson Apacible, "*Learning and Reasoning about Interruption*", ICMI'03, November 5-7, 2003, Vancouver, British Columbia, Canada.

[47] Nuria Oliver, Eric Horvitz, "*Selective Perception Policies for Guiding Sensing and Computation in Multimodal Systems: A Comparative Analysis*", ICMI'03, November 5-7, 2003, Vancouver, British Columbia, Canada.

[48] Sharon Oviatt, Rachel Coulston, Stefanie Tomko, Benfang Xiao, Rebecca Lunsford, Matt Wesson, Lesley Carmichael, "*Toward a Theory of Organized Multimodal Integration Patterns during Human-Computer Interaction*", ICMI'03, November 5-7, 2003, Vancouver,

British Columbia, Canada.

[49] Robert Snelick, Mike Indovina, James Yen, Alan Mink, "*Multimodal Biometrics: Issues in Design and Testing*", ICMI'03, November 5-7, 2003, Vancouver, British Columbia, Canada.

[50] Christian Elting, Stefan Rapp, "*Architecture and Implementation of Multimodal Plug and Play*", ICMI'03, November 5-7, 2003, Vancouver, British Columbia, Canada.

[51] Darrell S. Rudmann, George W. McConkie, Xianjun Sam Zheng, "*Eyetracking in Cognitive State Detection for HCI*", ICMI'03, November 5-7, 2003, Vancouver, British Columbia, Canada.

[52] Johnston, M., and Bangalore, S. "*Finite-state Methods for Multimodal Parsing and Integration*", In Finite-state Methods Workshop ESSLLI Summer School on Logic Language and Information, Helsinki, Finland, August, 2001.

[53] S. Wrede, M. Hanheide, C. Bauckhage, and G. Sagerer, "*An Active Memory as a Model for Information Fusion*", Int. Conf. on Information Fusion, number 1, pp. 198-205, 2004.

[54] J.B. Allen, "*How do Humans Process and Recognize Speech?*" IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, pp. 567-77, October 1994.

[55] Benfang Xiao, Rebecca Lunsford, Rachel Coulston, Matt Wesson, and Sharon Oviatt, "*Modeling Multimodal Integration Patterns and Performance in Seniors: Toward Adaptive Processing of Individual Differences"*, ICMI 2003, November 5-7 2003 Vancouver, Canada.

[56] Benfang Xiao, Cynthia Girand, and Sharon Oviatt, "*Multimodal Integration Patterns in Childern"*, Proceedings of ICSLP2002, Casual Productions, Ltd. pp. 629-632.

[57] Sharon Oviatt, Rebecca Lunsford, Rachel Coulston, "*Individual Differences in Multimodal Integration Patterns: What Are They and Why Do They Exist?*", CHI 2005, April 2-7, 2005, Portland, Oregon, USA.

[58] Sharon Oviatt, "*Toward Adaptive Information Fusion in Multimodal Systems"*, NICTA-HCSNet Multimodal user Interaction Workshop (MMUI2005), Sydney, Australia; Conferences in Research and Practice in Information Technology, Vol. 57.

[59] Anastasio, T. J., & Patton, P. E. (2004). Analysis and modeling of multisensory enhancement in the deep superior colliculus. In G. Calvert, C. Spence & B. E. Stein (Eds.), *Handbook of Multisensory Processes* (pp. 265-283). Cambridge, MA: MIT Press.

[60] Massaro, D.W. (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, MA: MIT Press.

[61] Computational methods of Signal Recovery by R. Mammone

[62] Darius Miniotas1, Oleg Špakov2, Ivan Tugoy2, I. Scott MacKenzie3, “Speech-Augmented Eye Gaze Interaction with Small Closely Spaced Targets”, ETRA 2006, San Diego, California, 27–29 March 2006.

[63] Stefan Graf1, Wolfgang Spiessl1, Albrecht Schmidt2, Anneke Winter1 and Gerhard Rigoll, “In-car Interaction using Search-Based User Interfaces”, CHI 2008 April 5-10, 2008, Florence, Italy.

[64] Péter Pál Boda , “A Contextual Multimodal Integrator”, ICMI’06, November 2–4, 2006, Banff, Alberta, Canada

[65] Nuria Oliver and Eric Horvitz, “S-SEER Selective Perception in a Multimodal Office Activity Recognition System”, MLMI 2004, LNCS 3361, pp. 122–135, 2005

[66] Matt Feusner, Brian Lukoff, "Testing for statistically significant differences between groups of scan patterns", ETRA 2008, Savannah, Georgia, March 26–28, 2008.

[67] Vladim´ır Bergl, Martin Cˇ mejrek, Martin Fanta, Martin Labsky´, Ladislav Seredi, Jan Sˇ edivy´, Lubosˇ Ures, "CarDialer—MultiModal InVehicle Cellphone Control Application", ICMI’06, November 2–4, 2006, Banff, Alberta, Canada.

[68] Ludo Maat and Maja Pantic, "Gaze-X: Adaptive Affective Multimodal Interface for Single-User Office Scenarios", ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.

[69] Ing-Marie Jonsson, Helen Harris, Clifford Nass, "How Accurate must an In-Car Information System be? Consequences of Accurate and Inaccurate Information in Cars", CHI 2008, April 5–10, 2008, Florence, Italy.

[70] Maja Pantic, Alex Pentland, Anton Nijholt and Thomas Huan4, "Human Computing and Machine Understanding of Human Behavior: A Survey", ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.

[71] C. Mario Christoudias, Kate Saenko, Louis-Philippe Morency and Trevor Darrell, "Co-Adaptation of Audio-Visual Speech and Gesture Classifiers", ICMI’06, November 2–4, 2006, Banff, Alberta, Canada.

[72] Howell Istance, "Communication Through Eye-Gaze: Where We Have Been, Where We Are Now and Where We Can Go From Here", ETRA 2006, San Diego, California, 27–29

March 2006.

[73] Tim J. Smith, Martyn Whitwell, John Lee, "Eye Movements and Pupil Dilation During Event Perception", ETRA 2006, San Diego, California, 27–29 March 2006.

[74] Leah Findlater and Joanna McGrenere, "Impact of Screen Size on Performance, Awareness, and User Satisfaction With Adaptive Graphical User Interfaces", CHI 2008, April 5–10, 2008, Florence, Italy.

[75] Geoffrey Tien, M. Stella Atkins, "Improving Hands-free Menu Selection Using Eyegaze Glances and Fixations", ETRA 2008, Savannah, Georgia, March 26–28, 2008.

[76] Emiliano Castellina, Fulvio Cornoy, Paolo Pellegrino, "Integrated Speech and Gaze Control for Realistic Desktop Environments", ETRA 2008, Savannah, Georgia, March 26–28, 2008.

[77] Yu-Chi Tai, James E. Sheedy, John Hayes, "Effect of Letter Spacing On Eye Movements and Reading Performance", ETRA 2006, San Diego, California, 27–29 March 2006.

[78] Tao Lin, Atsumi Imamiya, "Evaluating Usability Based on Multimodal Information: An Empirical Study", ICMI'06, November 2–4, 2006,, Banff, Alberta, Canada.

[79] Yvonne Kammerer, Katharina Scheiter, Wolfgang Beinhauer, "Looking my Way through the Menu: The Impact of Menu Design and Multimodal Input on Gaze-based Menu Selection", ETRA 2008, Savannah, Georgia, March 26–28, 2008.

[80] Pilar Manchón Portillo, Guillermo Pérez García, Gabriel Amores Carredano, "Multimodal Fusion: A New Hybrid Strategy for Dialogue Systems", ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.

[81] Jeff Klingner, Rakshit Kumar, and Pat Hanrahan, "Measuring the Task-Evoked Pupillary Response with a Remote Eye Tracker", ETRA 2008, Savannah, Georgia, March 26–28, 2008.

[82] Krzysztof Z. Gajos, Katherine Everitt, Desney S. Tan, Mary Czerwinski, Daniel S. Weld, "Predictability and Accuracy in Adaptive User Interfaces", CHI 2008, April 5–10, 2008, Florence, Italy.

[83] Xiao Huang and Sharon Oviatt, "Toward Adaptive Information Fusion in Multimodal Systems", MLMI 2005, LNCS 3869, pp. 15–27, 2006.

[84] Yuan-Chi Tseng, Andrew Howes, "The Adaptation of Visual Search Strategy to Expected Information Gain", CHI 2008, April 5–10, 2008, Florence, Italy.

[85] Yoshiko Habuchi, Muneo Kitajima, Haruhiko Takeuchi, "Comparison of Eye Movements in Searching for Easy-to-Find and Hard-to-Find Information in a Hierarchically Organized Information Structure", ETRA 2008, Savannah, Georgia, March 26–28, 2008.

[86] Akiko Yamazaki, Keiichi Yamazaki, Yoshinori Kuno, Matthew Burdelski, Michie Kawashima, Hideaki Kuzuoka, "Precision Timing in Human-Robot Interaction: Coordination of Head Movement and Utterance", CHI 2008, April 5–10, 2008, Florence, Italy.

[87] Daniel Gepner, Jérôme Simonin, Noëlle Carbonell, "Gaze as a Supplementary Modality for Interacting with Ambient Intelligence Environments", Campus Scientifique, BP 239, F54506, Vandoeuvre-lès-Nancy Cedex, France

[88] Carlos Duarte, Lu´ıs Carric, "A Conceptual Framework for Developing Adaptive Multimodal Applications", IUI'06, January 29–February 1, 2006, Sydney, Australia.

[89] Qiaohui Zhang, Atsumi Imamiya, Kentaro Go, Xiaoyang Mao, "A Gaze and Speech Multimodal Interface", Proceedings of the 24th International Conference on Distributed Computing Systems Workshops (ICDCSW'04).

[90] Yong Sun1, Fang Chen1, Yu (David) Shi, Vera Chung, "A Novel Method for Multi-sensory Data Fusion in Multimodal Human Computer Interaction", OZCHI 2006, November 20-24, 2006, Sydney, Australia.

[91] Andrew F. Monk and Leon Watts, "A poor quality video link affects speech but not gaze", CHI' Companion 95, Denver, Colorado, USA.

[92] Peter Gorniak, Deb Roy, "Augmenting User Interfaces with Adaptive Speech Commands", ICMI'03, November 5–7, 2003, Vancouver, British Columbia, Canada.

[93] Joyce Y. Chai, Pengyu Hong, Michelle X. Zhou, "Combining Semantic and Temporal Constraints for Multimodal Integra-tion in Conversation Systems".

[94] Jiazhi Ou, Lui Min Oh, Jie Yang, Susan R. Fussell, "Effects of Task Properties, Partner Actions, and Message Content on Eye Gaze Patterns in a Collaborative Task", CHI 2005, April 2 – 7, 2005, Portland, Oregon, USA.

[95] Yeow Kee Tan, Nasser Sherkat, Tony Allen, "Error Recovery in a Blended Style Eye Gaze and Speech Interface", ICMI'03, November 5–7, 2003, Vancouver, British Columbia, Canada.

[96] Alice Oh, Harold Fox, Max Van Kleek, Aaron Adler, Krzysztof Gajos, Louis-Philippe Morency, and Trevor Darrell, "Evaluating Look-to-Talk: A Gaze-Aware Interface in a Collaborative Environment", CHI 2002, April 20-25, 2002, Minneapolis, Minnesota, USA.

[97] Linda E. Sibert, Robert J.K. Jacob, "Evaluation of Eye Gaze Interaction", CHI '2000 The Hague, Amsterdam.

[98] Emilio Schapira, Rajeev Sharma, "Experimental Evaluation of Vision and Speech based Multimodal Interfaces", PUI 2001 Orlando, FL USA.

[99] Jian-Gang Wang, Eric Sung, Ronda Venkateswarlu, "Eye Gaze Estimation from a Single Image of One Eye", Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV’03).

[100] Zahar Prasov and Joyce Y. Chai and Hogyeong Jeong, "Eye Gaze for Attention Prediction in Multimodal Human-Machine Conversation", Department of Computer Science Michigan State University East Lansing, MI 48823, USA.

[101] Paul P. Maglio, Teenie Matlock, Christopher S. Campbell, Shumin Zhai, and Barton A. Smith, "Gaze and Speech in Attentive User Interfaces, ICMI 2000, LNCS 1948, pp. 1-7, 2000.

[102] Tetsuro Chino, Kazuhiro Fukui, and Kaoru Suzuki, "GazeToTalk: A Nonverbal Interface with Meta-Communication Facility", Eye Tracking Research & Applications Symposium 2000 Palm Beach Gardens, FL, USA.

[103] Leah M. Reeves, Jennifer Lai, James A. Larson, Sharon Oviatt, T.S. Balaji, Stéphanie Buisine, Penny Collings, Phil Cohen, Ben Kraal, Jean-Claude Martin, Michael McTear, TV Raman, Kay M. Stanney, Hui Su, and QianYing Wang, "Guidelines for Multimodal User Interface Design", Communications of the ACM, January 2004/Vol. 47, No. 1.

[104] Jian Wang, "Integration of Eye-gaze, Voice and Manual Response in Multimodal User Interface", Department of Psychology, National Key Laboratory of Human Factors, Hangzhou University, Hangzhou, Zhejiang 310028,China, IEEE 1995.

[105] Joyce Y. Chai Zahar Prasov Joseph Blaim Rong Jin, "Linguistic Theories in Efficient Multimodal Reference Resolution: An Empirical Investigation", IUI’05, January 9–12, 2005, San Diego, California, USA.

[106] Zahar Prasov, Joyce Chai, "Predicting User Attention using Eye Gaze in Conversational Interfaces", Department of Computer Science, Michigan State University, East Lansing, MI 48823

[107] Alessandra Pireddu, "Multimodal Interaction: an integrated speech and gaze approach", Politecnico di Torino.

[108] Ivan Marsic, Attila Medl, and James Flanagan, "Natural Communication with Information Systems", IEEE 2000.

[109] Joyce Y. Chai, Pengyu Hong, Michelle X. Zhou, Zahar Prasov, "Optimization in Multimodal Interpretation", Computer Science and Engineering, Michigan State University, East Lansing, MI 48824.

[110] Qiaohui Zhang, Kentaro Go, Atsumi Imamiya, Xiaoyang Mao, "Overriding Errors in a Speech and Gaze Multimodal Architecture", IUI'04, Jan. 13-16, 2004, Madeira, Funchal, Portugal.

[111] Moran Cerf, Jonathan Harel, Wolfgang Einh¨auser, Christof Koch, "Predicting human gaze using low-level saliency combined with face detection", California Institute of Technology/Swiss Federal Institute of Technology.

[112] Susan R. Fussell, Robert E. Kraut, Jane Siegel, Susan E. Brennan, "Relationships Among Speech, Vision, and Action in Collaborative Physical Tasks", CHI 2002, April 20-25, 2002, Minneapolis, Minnesota, USA.

[113] Qiaohui Zhang, Atsumi Imamiya, Kentaro Go, Xiaoyang Mao, "Resolving Ambiguities of a Gaze and Speech Interface", ACM 2004.

[114] Qiaohui Zhang, Atsumi Imamiya, Kentaro Go, Xiaoyang Mao, "Robust Object-Identification from Inaccurate Recognition-Based Inputs", AVI '04, May 25-28, 2004, Gallipoli (LE), Italy.

[115] Darius Miniotas, Oleg Špakov, Ivan Tugoy, I. Scott MacKenzie, "Speech-Augmented Eye Gaze Interaction with Small Closely Spaced Targets", ETRA 2006, San Diego, California, 27–29 March 2006.

[116] Niels Ole Bernsen, NISLab, Denmark (editor), "Speech-Related Technologies, Where will the field go in 10 years?"

[117] Michael Johnston, Philip R. Cohen, David McGee, Sharon L. Oviatt, James A. Pittman, Ira Smith, "Unification-based Multimodal Integration", Center for Human Computer Communication Department of Computer Science and Engineering Oregon Graduate Institute, PO BOX 91000, Portland, OR 97291, USA.

[118] Sharon Oviatt, "User-Centered Modeling and Evaluation of Multimodal Interfaces", Proceedings of the IEEE, Vol. 91, No. 9, September 2003.

[119] Zahar Prasov and Joyce Y. Chai, "What's in a Gaze? The Role of Eye-Gaze in Reference Resolution in Multimodal Conversational Interfaces", IUI'08, January 13-16, 2008, Maspalomas, Canary Islands, Spain.

[120] Susan R. Fussell, Leslie D. Setlock, Elizabeth M. Parker, "Where do Helpers Look? Gaze Targets During Collaborative Physical Tasks", CHI 2003, April 5-10, 2003, Ft. Lauderdale, Florida, USA.

[121] Candace L. Sidner, Cory D. Kidd, Christopher Lee and Neal Lesh, "Where to Look: A Study of Human-Robot Engagement", IUI'04, January 13-16, 2004, Madeira, Funchal, Portugal.

[122] Dario D. Salvucci, "An Interactive Model-Based Environment for Eye-Movement Protocol Analysis and Visualization", Eye Tracking Research & Applications Symposium 2000 Palm Beach Gardens, FL, USA.

[123] Dario D. Salvucci, Joseph H. Goldberg, "Identifying Fixations and Saccades in Eye-Tracking Protocols", Eye Tracking Research & Applications Symposium 2000 Palm Beach Gardens, FL, USA.

[124] Minoru Nakayama, Koji Takahashi, "The Act of Task Difficulty and Eye-movement Frequency for the Oculo-motor indices", ETRA 2002 New Odeans Louisiana USA.

[125] Paivi Majaranta and Kari-Jouko Raiha, "Twenty Years of Eye Typing: Systems and Design Issues", ETRA 2002 New Odeans Louisiana USA.

[126] Brian Goldiez, Glenn Martin, Jason Daly, Donald Washburn, and Todd Lazarus, "*Software Infrastructure for Multi-Modal Virtual Environments*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[127] Erum Tanvir, Jonathan Cullen, Pourang Irani, Andy Cockburn, "AAMU: Adaptive Activation Area Menus for Improving Selection in Cascading Pull-Down Menus", CHI 2008 Proceedings Mixed-Initiative Interaction April 5-10, 2008 · Florence, Italy.

[128] Xiao Huang, Sharon Oviatt, and Rebecca Lunsford, "Combining User Modeling and Machine Learning to Predict Users' Multimodal Integration Patterns", MLMI 2006, LNCS 4299, pp. 50 – 62, 2006.

[129] Sharon Oviatt, Colin Swindells, Alex Arthur, "Implicit User-Adaptive System Engagement in Speech and Pen Interfaces", CHI 2008 Proceedings Mixed-Initiative Interaction April 5-10, 2008 Florence, Italy.

[130] M. Sodhi, B. Reimer, J. L. Cohen E. Vastenburg, It. Kaars, "On-Road Driver Eye Movement Tracking Using Head-Mounted Devices", ETRA 2002 New Odeans Louisiana USA.

[131] Qiaohui Zhang, Atsumi Imamiya, Kentaro Go, Xiaoyang Mao, "Resolving Ambiguities of a Gaze and Speech Interface", ACM 2004.

[132] http://www.math.yorku.ca/SCS/Online/paivio

[133] Arlo Faria, "Accent Classification for Speech Recognition", MLMI 2005, LNCS 3869, pp. 285–293, 2006.

[134] Manolis Perakakis, Alexandros Potamianos, "The Effect of Input Mode on Inactivity and Interaction Times of Multimodal Systems", ICMI'07, November 12-15, 2007, Nagoya, Aichi, Japan.

[135] Matthias W¨olfel, Kai Nickel, and John McDonough, "Microphone Array Driven Speech Recognition: Influence of Localization on the Word Error Rate", MLMI 2005, LNCS 3869, pp. 320–331, 2006.

[136] Frederick Shic, Katarzyna Chawarska, "The Incomplete Fixation Measure", ETRA 2008, Savannah, Georgia, March 26–28, 2008.

[137] Boris M, Velichkovsky, Sascha M. Domhoefer, Sebastian Pannasch and Pieter J.A.Unema, "Visual Fixations and Level of Attentional Processing", Eye Tracking Research & Applications Symposium 2000 Palm Beach Gardens, FL, USA.

[138] Marc Erich Latoschik, "*A User Interface Framework for Multimodal VR Interactions*", ICMI 2005, October 4-6, 2005, Trento, Italy.

[139] Nigay, L., Coutaz, J. "*A Generic Platform for Addressing the Multimodal Challenge*", Proceedings of CHI'95 (1995), 98-105.

[140] Norbert Reithinger, Simon Bergweiler, Ralf Engel, Gerd Herzog, Norbert Pfleger, Massimo Romanelli, Daniel Sonntag, "*A Look Under the Hood – Design and Development of the First SmartWeb System Demonstrator*", ICMI 2005, October 4-6, 2005, Trento, Italy.

[141] Mitre Corporation, "*Galaxy Communicator Documentation*", 2002, Available on the web at http://communicator.sourceforge.net/sites/MITRE/distributions/GalaxyCommunicatordocs/manual/index.html.

[142] Remi Bastide, David Navarre, Philippe Palanque, Amelie Schyn & Pierre Dragicevic, "*A Model-Based Approach for Real-Time Embedded Multimodal Systems in Military Aircrafts*", ICMI'04, October 13-15, State College, Pennsylvania, USA.

[143] Westeyn, T. et. al., "*Georgia Tech Gesture Toolkit: Supporting Experiments in Gesture Recognition*", Proceedings of ICMI'03 (2003), 85-92.

Curriculum Vita

Chandra Sekhar Mantravadi

1989-1993

Chaithanya Bharathi Institute of Technology, Osmania University, Hyderabad, India
Bachelor of Engineering in Electronics and Communication Engineering

1993-1994

University of Kentucky, Lexington, KY, USA

1994-1996

Midwestern State University, Texas, TX, USA
Master of Sciences in Computer Science

1997-2009

Rutgers, The State University of New Jersey, New Jersey, NJ, USA
Doctor of Philosophy in Electrical and Computer Engineering

1996-Current

Project Manager for a large scale global real-time data warehouse platform for Equities division in a financial firm

Publications:

- M. Kaur, M. Tremaine, N. Huang, J. Wilder, F. Flippo, Z. Gacovski, C. S. Mantravadi, "Where is "it"? Event Synchronization in Gaze-Speech Input Systems", Proceedings of Fifth International ACM Conf., ICMI'03 (2003 International Conference on Perceptive and Multimodal User Interfaces), Nov 2003
- C. S. Mantravadi, J. Wilder, D. Grove, X. Yuan, "A Java-based multimodal human-computer interface architecture," presented at the *Third International Conference on Information, Communications and Signal Processing*, Singapore, (Proceedings on CD), Oct. 2001