Conveying Routes: Multimodal Generation and Spatial Intelligence in Embodied Conversational Agents

by

Thomas A. Stocky

Submitted to the Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science in Computer Science and Engineering

and Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

May 24, 2002

Author_____

Department of Electrical Engineering and Computer Science
May 24, 2002

Certified by_____

Justine Cassell
Thesis Supervisor

Accepted by_____

Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Conveying Routes: Multimodal Generation and Spatial Intelligence in Embodied Conversational Agents

by

Thomas A. Stocky

Submitted to the

Department of Electrical Engineering and Computer Science

May 24, 2002

In Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science in Computer Science and Engineering

and Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

In creating an embodied conversational agent (ECA) capable of conveying routes, it is necessary to understand how to present spatial information in an effective and natural manner. When conveying routes to someone, a person uses multiple modalities – e.g., speech, gestures, and reference to a map – to present information, and it is important to know precisely how these modalities are coordinated. With an understanding of how humans present spatial intelligence to give directions, it is then possible to create an ECA with similar capabilities. Two empirical studies were carried out to observe natural human-to-human direction-giving interactions. From the results, a direction-giving model was created, and then implemented in the MACK (**M**edia Lab **A**utonomous **C**onversational **K**iosk) system.

Thesis Supervisor: Justine Cassell
Title: Associate Professor, MIT Media Laboratory

Acknowledgements

I would especially like to thank my advisor, Professor Justine Cassell, for her support and guidance over the past year and throughout the entire thesis process. I am especially grateful for her patience in helping me understand the essence of *motivated* research.

I am also extremely grateful to the other members of the GNL team. In addition to kindly welcoming me into the group, they have served as a constant source of help and support. I would like to especially thank:

Tim Bickmore, for serving as a mentor and intellectual resource regarding everything from programming design patterns to demonstrative pronouns.

Ian Gouldstone, for sharing his graphical talents with the MACK system, and for his invariable good nature and sense of humor while teaching me the intricacies of Pantomime and KQML.

Hannes Vilhjálmsson, for his valuable insights in response to my daily bombardment of discourse questions, and for his continued patience as I misused Pantomime Server in every way possible.

Kimiko Ryokai and Catherine Vaucelle, for all their guidance along the way, and especially for their much-needed moral support when it was needed most.

Yukiko Nakano, for her help in teaching me the fundamentals of data collection and analysis, and for all her patience and understanding.

Dona Tversky, for her consistent willingness to lend a helping hand, and for her commitment to subduing my sarcasm to a tolerable level.

I would like to dedicate this thesis to my parents, whose unwavering love and support continues to provide me the strength to pursue my aspirations. And I would like to thank my sister, whose patience and kindness have proved endless, as well as my entire family for all their love and encouragement.

Contents

Contents 4

I. Introduction 5
- i. Motivation 5
- ii. Research Overview 6
- iii. Thesis Layout 7

II. Related Work 8
- i. Embodied Agents 8
- ii. Kiosks and Information Displays 9
- iii. Spatial Reference 11
- iv. Route Directions 12

III. My Research 15
- i First Study 15
- ii Second Study 18

IV. Direction-Giving Model 21
- i. Relative Descriptions (RD) 21
- ii. Speech and Gesture (SG) 21
- iii. Map-Based (MB) 22

V. Implementation 25
- i. Introduction to MACK 25
- ii. Map Representation 27
- iii. Path Calculation 29
- iv. Direction Generation 30

VI. Future Work 33

VII. Conclusions 37

VIII. References 38

I. Introduction

I. i. Motivation

Over the last few years, research in computational linguistics, multimodal interfaces, computer graphics, and intelligent interactive systems has led to the development of more and more sophisticated autonomous or semi-autonomous virtual humans. Increasingly, there has been a shift from modeling physical human motion to instead creating the underlying models of behavior and intelligence.

The last few years also represents a continuing trend to develop, and to situate in public locations, information access systems that can deliver resources and services to the general public. These kiosk systems place interesting constraints on interface design. Kiosk systems must stand out so that they will be noticed by casual passers-by, and their purpose must be self-evident. Users of kiosk systems do not have time for lengthy training, and so interaction must be intuitive and self-explanatory. The user base for a typical public space tends to represent a diverse set of backgrounds, and so systems must be geared toward the least common denominator and demonstrate the ability for effective error-recovery.

Despite clear advances in technology, however, the old-fashioned information booth in railway stations, department stores, and museums had one significant advantage over today's information kiosk: staff members could rely on the physical space shared with a visitor in order to give directions, describe processes, and spatialize relationships among places and things. Uniformed railway personnel could point to the proper train platform; department store hostesses could show specific product locations on unfolded

store maps; and museum staff could illustrate nonverbally the size of the dinosaurs in the great hall.

This issue of communicating spatial intelligence is clearly an important consideration in designing a public information kiosk. In creating an embodied conversational agent (ECA) capable of conveying routes, it is necessary to understand how to present spatial information in an effective and natural manner. When giving directions to someone, a person uses multiple modalities – e.g., speech, gestures, and reference to a map – to present information, and it is important to know precisely how these modalities are coordinated. Past work has looked at the use of hand-drawn street maps in direction-giving tasks [Tversky and Lee, 1999], but unconstrained direction-giving within a building was previously unstudied.

I. ii. Research Overview

The goal of this thesis is to provide for an ECA a direction-giving framework that can coordinate speech, gesture, and map-based reference, based on an underlying model of human-to-human direction-giving behavior. In addition to this framework, three back-end structures are explored in this thesis: (1) map representation, (2) path calculation, and (3) generation of map-based reference and deictic gestures.

To create a direction-giving model, two empirical studies were carried out to observe natural human-to-human interactions. The first study was an observation of unconstrained direction-giving interactions in the Lobby of the MIT Media Lab, where a map was easily accessible on the wall between the elevators. The second study was in a more controlled environment to examine direction generation in more detail. The

resulting model was implemented in the MACK (**M**edia Lab **A**utonomous **C**onversational **K**iosk) system [Cassell and Stocky et al., 2002].

I. iii. Thesis Layout

This chapter presents the motivation behind my research, along with an overview of the research work. The following chapter introduces the research context in the form of related work in embodied agents, kiosks and information displays, spatial reference, and route directions. Chapter III describes the empirical studies on human-to-human direction-giving, the results of which produced the direction-giving model defined in Chapter IV. Chapter V details the implementation of a direction generation system into MACK, followed by suggestions for future work and conclusions.

II. Related Work

II. i. Embodied Agents

In past research, embodiment has proven its effectiveness in engaging users [Koda and Maes, 1996; Reeves and Nass, 1996], and has shown a qualitative advantage over non-embodied interfaces, enabling the exchange of multiple levels of information in real time [Cassell, Bickmore, Vilhjálmsson, and Yan, 2000]. One example of this is Rea, an ECA real estate agent capable of both multimodal input and output. Users that interacted with Rea found the interface intuitive and natural, as conversation is an intrinsically human skill that requires no introduction [Cassell et al., 1999]. With this in mind, an ECA is the natural choice for implementing an interactive system capable of giving directions.

Cambridge Research Laboratory has also explored this concept of embodiment while trying to create a better way for people to obtain information in public spaces. They began with a traditional public access kiosk, and enhanced it with an animated head and synthesized speech. The kiosk was also able to do face-tracking on passing users [Waters and Levergood, 1993]. They deployed these kiosks in public places, and one of the foremost lessons learned was that people were attracted to an animated face that watched them [Christian and Avery, 2000]. CRL also report, however, that while a face-only avatar did well at attracting and entertaining people, it was not successful at conveying or interacting with content. These animated heads lack the ability to indicate spatiality through hand and arm gestures, which is crucial in a system that gives directions.

II. ii. Kiosks and Information Displays

Research indicates that public information kiosks are useful and effective interfaces. They have been shown to increase user acceptance of the online world in that they serve a wide range of individuals. Knowledge transfer is also improved, as kiosk users have demonstrated that they gain new information and tend to use the system repeatedly after initial interactions. Further, kiosks increase business utility by increasing the likelihood of purchase and reducing the time needed for physical staff to provide advice [Steiger and Suter, 1994].

However, current kiosks have been limited in interaction techniques, requiring literacy on the part of users, and the use of one's hands to type or choose information. Replacing text and graphics with an ECA may result in systems that are more flexible, allowing for a wider diversity in users. ECAs allow for hands-free multimodal input and output (speech and gesture), which produces a more natural, more intuitive interaction [Cassell et al., 1999]. These communication protocols come without need for user training, as all users have these skills and use them daily. Natural language and gesture take full advantage of the shared environment, creating a spatial bridge between the user and the agent.

Significant research has been conducted to find ways of effectively presenting information, as well as ways to allow users to interact with that information. Much research, for example, has concentrated on using touch screens to allow more intuitive interaction with bodies of information. Additional research has examined the most natural kinds of linkages between those bodies of information, in order to allow users to

engage in “social navigation” – following the trails of others, or patterns generated by their own evolving interests.

The MINELLI system was created as a hypermedia public information kiosk with a touch screen interface. Rather than the standard, static, text-with-graphics content, MINELLI used short films, musical and graphical content, and interactive games to engage users and make them feel comfortable using the system [Steiger and Suter, 1994]. While MINELLI was certainly an improvement over standard kiosks, it required user training, which is not ideal for a public access system. Raisamo’s Touch’n’Speak kiosk demonstrated another approach, employing natural language input in conjunction with a touch screen. These modalities were selected with the goal of creating an intuitive interface that required no user training [Raisamo, 1999]. While a touch screen is perhaps more intuitive than a keyboard and mouse, both MINELLI and Touch’n’Speak remain limited to a primarily menu-driven process flow. Embedding an ECA into the interface addresses this limitation with the use of dialogue-based interaction, which has the added benefit of not requiring user literacy.

Others have looked at multimodal display of information, and multimodal input. Looking at the combination of text and graphics in information display, Kerbedjiev proposed a methodology for realizing communicative goals in graphics. He suggested that more appealing multimedia presentations take advantage of both natural language and graphics [Kerpedjiev et al., 1998]. Such findings have paved the way for ECAs, capable of natural language and its associated nonverbal behaviors.

Feiner and McKeown made a similar distinction between the function of pictures and words. Pictures describe physical objects more clearly, while language is more adept

in conveying information about abstract objects and relations. This research led to their COMET system, which generates text and 3D graphics on the fly [Feiner and McKeown, 1998]. Similarly, Maybury's TEXTPLAN generated multimedia explanations, tailoring these explanations based on the type of communicative act required [Maybury, 1998].

Similar to TEXTPLAN, Wahlster's WIP model stressed the idea that the various constituents of a multimodal presentation should be generated from a common representation of what is to be conveyed [Wahlster et al., 1993]. This is an important point because it stresses the importance of correctly coordinating the multimodal output. Using an ECA, this problem does not become any easier, but the resulting natural output better achieves this goal of coordination [Cassell et al., 2001].

II. iii. Spatial Reference

In designing interfaces capable of spatial reference, there have been a number of different approaches. For example, Billinghurst implemented an immersive virtual reality system in the form of an intelligent medical interface. It was designed to allow surgeons to interact with virtual tissue and organ models, achieved through the use of continuous voice recognition coupled with gesture input via pointing and grasping of simulated medical instruments [Billinghurst et al., 1996]. This approach achieved spatial reference through the use of a shared virtual reality.

Other implementations have come closer to creating shared *physical* reality through the use of devices that can accompany the user, such as PDAs. One such interface, PalmGuide, is a hand-held tour guidance system that refers to objects in the user's reality, recommending exhibits that may be of interest. The interface is primarily

text-based, accented by user-specified character icons that give it an added sense of familiarity. When in the vicinity of a computer kiosk, PalmGuide is able to exchange information so that the kiosk can present information in a way that is comfortable and natural to the user [Sumi and Mase, 2000].

The ability to refer to space multimodally is addressed by OGI's QuickSet system, which allows users to reference maps through the use of pen and voice inputs. For example, a user can create an open space on a map by drawing an area and saying, "open space." This multimodal input allows users to interact with the system in a natural and intuitive manner [Oviatt and Cohen, 2000].

In designing an embodied agent capable of spatial reference, an important consideration is that of deictic believability. This requires that the agent consider the physical properties of the world it inhabits, and effectively uses its knowledge of (1) the positions of objects in the world, (2) its relative location with respect to these objects, and (3) its prior explanations regarding these objects. This knowledge must be applied in creating deictic gestures, motions, and utterances that are both natural and unambiguous. [Lester et al., 2000]

II. iv. Route Directions

Much research has been devoted to route directions and their underlying structure and semantics. Michon and Denis examined the use of landmarks in direction-giving. They found that landmarks are used most frequently at specific points on the route, especially at reorientation points. Landmarks were also found useful to direction-receivers in helping them to construct mental representations of unfamiliar environments

in which they are preparing to move [Michon and Denis, 2001]. The use of spatial referents (such as landmarks) has received considerable research attention, including work on how spatial referent use is affected by gender and regional differences among direction-givers [Lawton, 2001].

Assessing the quality of route directions has also been studied. Lovelace, Hegarty, and Montello looked at the elements of good route directions in both familiar and unfamiliar environments. Once again, landmarks played a key role in the study. They found that higher quality route directions made use of more route elements (landmarks, segments, turns, etc.). They also found that the type of landmarks used differed for familiar versus unfamiliar routes [Lovelace, Hegarty, and Montello, 1999].

Most relevant to my thesis, Tversky and Lee described how people use speech and maps to convey routes. In examining external representations of spatial intelligence, they have begun to reveal the underlying structure and semantics for route directions:

> The first step is to put the listener at the point of departure. In the field, this is typically apparent to both interlocutors and need not be specified. The second step, beginning the progression, may also be implicit. The next three steps are used iteratively until the goal is reached: designate a landmark; reorient the listener; start the progression again by prescribing an action. Actions may be changes of orientation or continuations in the same direction. [Tversky and Lee, 1999]

Later work with Emmorey and Taylor distinguished two distinct perspectives for giving directions: *route* and *survey*. These perspectives were found to differ with respect to (1) point of view (moving within the scene vs. fixed above the scene), (2) reference object (the direction-receiver vs. some landmark), and (3) reference terms (right-left-front-back vs. north-south-east-west). Route and survey perspectives also correspond to

two distinct ways of experiencing one's environment. With a route perspective, the direction-giver experiences the environment from within, describing one's navigation through it. In contrast, a survey perspective corresponds to viewing the environment from the outside, essentially looking on it as an object with parts. [Emmorey, Tversky, and Taylor, 2000]

Taylor and Tversky found that which perspective is adopted depends in part on features of the environment. Their research indicates that people tend to use the survey perspective when the environment has features on several size scales and when there are several routes through the environment. The route perspective, on the other hand, tends to occur when environmental features are on a single size scale and when there is only one natural route through the environment [Taylor and Tversky, 1996]. This indicates that while the route perspective is used in cases where the directions are straightforward, the survey perspective serves as a method of disambiguation.

III. My Research

Building on the related research, my goal was to create an accurate model for how people give directions, and then implement that model in an ECA. To create such a model, two human subject direction-giving studies were completed. The first study was in an unconstrained environment, an observation of people giving directions in the Lobby of the MIT Media Lab, where a map was easily accessible on the wall between the elevators. The second study was in a more controlled environment, and focused on how people give directions using speech and gesture.

III. i. First Study

The goal of the first study was to observe human-to-human direction-giving in as natural a setting as possible. With this in mind, subjects were told to find their way to two distinct locations in the MIT Media Lab from the lab's first floor lobby. They were asked to stand by the elevators, where there was a map of the building on the wall, and

Figure 1: Sample interaction from the First Study

ask for help from passersby, requesting clarification – for example, "I'm not sure I understand …" – after the second set of directions. The order of the two locations was varied among the subjects, and as the direction-requests were unscripted, the phrasing of the requests varied as well. Figure 1 shows a sample interaction.

The study ran with eight unique direction-receivers and eleven unique direction-givers, for a total of twenty direction-giving instances. (In two interactions, the direction-receiver did not request directions to a second location.) Of the eight unique direction-receivers, three were female and five were male. The direction-givers were composed of five females and six males.

Direction-givers employed three methods for direction-giving: (1) relative descriptions – i.e., a description of the destination relative to an implied or established context, such as "it's near the freight elevator" or "it's on the third floor" – (2) explanations with speech and gesture, and (3) map-based directions. Recalling the distinction between route and survey perspectives [Emmorey, Tversky, and Taylor, 2000], there was a strong correlation between the direction method and the perspective taken. As Table 1 indicates, speech and gesture (SG) directions coincided with the route perspective, while map-based (MB) directions coincided with survey.

	Route Perspective	Survey Perspective
Speech and Gesture	100% (9/9)	0% (0/9)
Map-Based	11% (1/9)	89% (8/9)

Table 1: Correlation between perspective and direction method

There was also a correlation between direction requests (the way the question was phrased) and the resulting direction method. Direction requests were segmented into three templates:

Q1: How would I get to Room 320?
Q2: How do you get to Room 320?
Q3: Where is Room 320?

These templates are not strict, e.g. "Can you tell me how I would get to Room 320?" and "Do you know how I could get to Room 320?" would both fall into Q1. As Table 2 indicates, Q1 tended to prompt MB directions, while Q2 prompted SG.

	Q1	Q2
Speech and Gesture	0% (0/2)	80% (4/5)
Map-Based	100% (2/2)	20% (1/5)

Table 2: Direction methods for different question types

The results show no correlation between the requested destination and the direction method (MB or SG), as well as no correlation between the destination and the perspective taken (survey vs. route). Since direction-givers' routes varied little for a given destination, this perhaps contradicts past research that suggested a direction-giver's perspective depends on whether there are several possible routes to the destination or only one natural route [Taylor and Tversky, 1996]. However, as the thought processes of the study's direction-givers are unknown, it is difficult to say how many potential routes

they considered, and so it cannot be said for certain whether or not the data coincides with Taylor and Tversky's findings.

With these initial results in mind, the second study was administered to examine the content of SG directions in more detail.

III. ii. Second Study

The second study focused on SG directions. Six subjects were used, three female and three male. Each subject was placed in the Media Lab's lower level lobby and asked to give directions to three distinct locations. The direction-requests were scripted so that each subject responded to all three question types previously described. To control for order effects, each subject received the question types in a different order, e.g. Q1-Q2-Q3 as opposed to Q3-Q1-Q2. All six orderings were used among the six subjects. Unrelated questions, such as "What are some of the research groups in the Digital Life consortium?" were asked between direction-requests to serve as a buffer.

The results showed that people tend to gesture when using direction words ("up," "down," "left," "right," "straight," "through"). Of the forty-four direction word occurrences, twenty-eight (64%) were accompanied by a pointing gesture. And 82% of those gestures were relative to the direction-giver's perspective. Essentially, people use gestures redundantly to emphasize the direction, and when, for example, they say, "Take a right," they gesture to their own right rather than gesturing to the listener's right. This somewhat contradicts their speech, however, in that 95% of the directions were given in the second person narrative ("you go") rather than first person ("I go").

When giving directions, people tended not to switch the gesturing hand. Of the twelve direction-giving instances where hand gestures occurred, 67% exhibited the same gesturing hand throughout the entire instance. In the other 33% the direction-giver went from a single hand to both hands. There was only one case (8%) where the direction-giver switched gesturing hands, and this was after switching from a single hand to both – i.e., right hand to both hands to left hand.

The results also indicated that direction-givers made use of previously mentioned direction segments when possible. Rather than generate an entirely new set of directions for the third direction request, five of the six subjects (84%) made use of past segments. An example from the data (note that "The Garden" and "The Pond" are rooms in the Media Lab):

> The Garden, it's just going the same way as The Pond,
> except you carry on cuz that's how you get to the entrance

As both rooms are off the same hallway, direction-givers tended to reuse segments in this manner. The one subject (16%) who gave entirely new directions for the third location started them with, "ok so you also go to the elevator," indicating with the use of "also" that she was repeating earlier direction segments.

Regarding question types, Q3 prompted a relative description (RD) every time. In four of those six cases, the direction-giver either prompted the listener with a question – e.g., "Do you want directions?" – or stopped talking, apparently considering the RD to be an entirely satisfactory response. This emphasizes the correlation between Q3 and RD.

This study provided a more thorough understanding of SG directions, as explained in more detail in the following chapter.

IV. Direction-Giving Model

IV. i. Relative Descriptions (RD)

In the case of RD, direction-givers generally seek the simplest, most concise description of the destination in terms of an implied or established common ground. Combining the results of the two studies, there were twenty-four instances of RD. Of those, twenty-two (92%) were of the form, "Room 320 is on the third floor." As Media Lab visitors tend to be unfamiliar with the building's key landmarks, it seems the only implicit common ground that direction-givers put to use was that the Media Lab was a building broken up into multiple floors.

IV. ii. Speech and Gesture (SG)

Similar to Tversky and Lee's findings [Tversky and Lee, 1999], people tend to give SG directions with segments to (1) designate a landmark and then (2) prescribe an action with a direction word. As the second study indicated, a deictic gesture coincides with the direction word in most cases, and that gesture is relative to the direction-giver's perspective. Once the destination is reached, some direction-givers then describe nearby landmarks, such as, "You'll see a large television set with two black couches." This happened in 22% of the second study's direction-giving interactions. In many cases (38% of the instances), direction-givers provided iconic gestures that coincided with these landmarks.

As mentioned in the previous chapter, an interesting discovery was that direction-givers gestured relative to themselves, regardless of question type, and despite the fact that they gave directions in second person narrative. This was an unexpected result, as it

was originally hypothesized that the direction-giver's perspective would likely depend on the way the question was phrased. It seemed reasonable that "how would I get to" direction-requests might elicit directions relative to the listener, while "how do you get" direction-requests would receive directions relative to the direction-giver. The results, however, clearly indicate otherwise.

IV. iii. Map-Based (MB)

Recalling related work describing the survey perspective [Emmorey, Tversky, and Taylor, 2000] and route maps [Tversky and Lee, 1999], MB directions closely resembled earlier findings. However, since earlier work focused on longer-distance directions (as opposed to within-building) and used maps drawn by direction-givers (instead of pre-existing maps), this MB model represents a different domain and is more concerned with map reference than map creation.

While some direction-givers (22%) start by showing the starting point on the map, in most cases (59%) the direction-giver begins by referencing the destination. (Others do neither and simply proceed with the next step.) Direction-givers then show a same-floor reference point on the map, usually the elevators on the destination's floor. This happened 78% of the time, with the others going directly into path segments.

Directions are broken into path segments from the starting point to the destination. Segments consist of (1) designating a landmark and pointing to it on the map, then (2) drawing a path from the previous landmark to this one. An example from the data:

> [points to area on map] elevators are here
> go down this hallway [traces path on map]

the kitchen is somewhere around there [points to area on map]

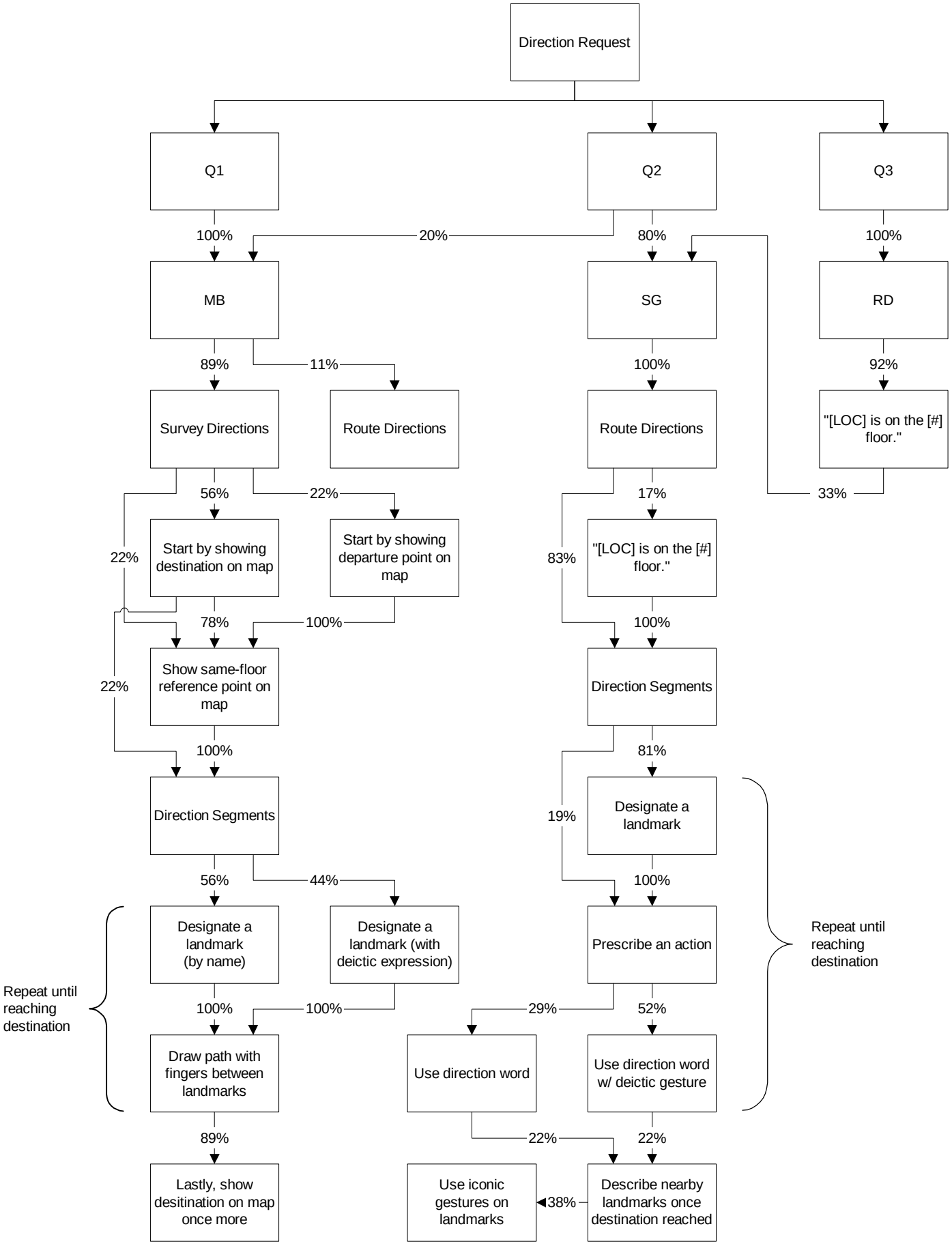
Landmarks were referenced by name 56% of the time, and with a deictic expression such as “here” in the other instances. There is no clear explanation for these different methods of reference. In looking for contextual influences on whether an objective name or deictic expression was used, potential correlations were explored – such as the type of landmark being referenced, whether the landmark had been referenced earlier, and whether certain subjects has a tendency toward one or the other – but none proved significant.

Landmarks took many forms, from elevators and doorways, to sections of the hallways (e.g., “the end of the hall”), to distinct objects, such as “black couches” and “a big TV.” Despite examples, however, there does not seem to be a clear method for predicting what constitutes a landmark, which is a subject that deserves future research.

Segments are continued until the destination is reached. Once reached, direction-givers tend to reference the destination once more by pointing to it on the map and saying something to the effect of, “And this is Room 320.” This happened 89% of the time, in eight of the nine instances of MB directions.

Figure 2, on the following page, illustrates the overall direction-giving model.

Figure 2: Direction-Giving Model



V. Implementation

From the direction-giving model, a set of probabilistic rules was extracted and implemented into the existing MACK (**M**edia Lab **A**utonomous **C**onversational **K**iosk) system [Stocky and Cassell, 2002].

V. i. Introduction to MACK

MACK is an interactive public information ECA Kiosk. He was designed with three primary goals in mind:

(1) Real-time multimodal input as a basis for natural face-to-face interaction,
(2) Coordinated natural language and gesture generation, and
(3) The ability to reference a shared physical space with the user.

The motivation was to take an existing virtual reality paradigm – immersing the user in a computer system's virtual world – and flip it around, to instead immerse the virtual agent into the user's physical world.

On the input side, MACK currently recognizes two input modalities: (1) speech recognition via the MIT Laboratory for Computer Science's SpeechBuilder technology [Glass and Weinstein, 2001] and (2) deictic map input via a paper map atop a table with an embedded Wacom tablet. The inputs operate as parallel threads, to be used individually or in combination. For example, a user can say, "Tell me about this" while pointing to a specific research group on the map, and MACK will respond with information about that group.

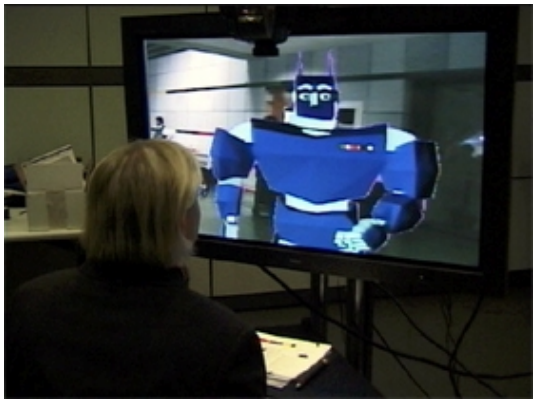


Figure 3: User interacting with MACK

MACK uses multimodal output as well, with (1) speech synthesis using the Microsoft Whistler Text-to-Speech (TTS) API, (2) an LCD projector output directed at the physical map allowing MACK to reference it, and (3) on-screen graphical output including synchronized hand and arm gestures, and head and eye movements. Pantomime [Chang, 1998] animates MACK's VRML-defined humanoid character using a variety of motor skill modules, and resolves any remaining conflicts in character degrees-of-freedom.

From the user's perspective, MACK is a life-sized on-screen blue robot seemingly located in their shared physical environment. (See Figure 3.) This is achieved with a video mixer and camera mounted atop the plasma screen display. On the screen behind MACK appears the video input, a direct feed of the physical background.

MACK is implemented on a Java 2 platform (J2SE 1.4), and both the map input and projector output modules make use of the Java 2D API. MACK's knowledgebase is stored in a MySQL database and is capable of updating itself directly from the Media Lab's internal database system. As the knowledgebase was designed to be modular, the Media Lab data could easily be substituted with another domain.

V. ii. Map Representation

The existing MACK system gave directions by calling on static scripts stored in a database. To implement a new direction-generation module, it was first necessary to develop an internal map representation.

As shown in Figure 4, the physical maps are typical building maps, and depict a two-dimensional "bird's eye view" of the various floors. The internal representation was designed to coincide with this, so each floor is represented on a 2D Cartesian coordinate plane, with the origin (0,0) in the upper left-hand corner and the X and Y axes increasing to the right and down, respectively. Rooms are represented as rectangles, hallways as line segments, and landmarks as points. This representation allows for easy data entry, as

Figure 4: Sample physical map for use with MACK

MACK's database can be quickly populated with coordinates to define the various map objects.

MACK's list of landmarks is based on the data from both studies. Most of the mentioned landmarks were elevators, doorways, or sections of the hallways (e.g. "halfway down the hall" or "at the end of the hall"). The other landmarks referenced were unique and apparent objects near to the path. These landmarks were unique in that they were one-of-a-kind within the building, and apparent in that they could be described easily with a few words and without need for further clarification. These two criteria were used as a guideline for determining the landmarks to populate MACK's database.

In addition to its location, a landmark's "visible area" is also represented. Its visible area is the area that includes all the locations from which the landmark can be seen. This becomes important when determining, for example, what landmarks can be seen from a particular room or hallway. In this way, it is possible for the system to know that people walking down a certain hallway will pass some landmark on their left or right, even though the hallway does not intersect the landmark itself.

Some types of landmarks have special properties. For example, a doorway is landmark that also serves as the way to get from a hallway to a room. This means that before MACK's directions reach a destination room, the path's endpoint must be a doorway. Similarly, an elevator is a special type of doorway that allows travel between floors. (This representation is facilitated by the use of subclasses in Java.) These special properties become important in the design of the path calculator.

V. iii. Path Calculation

Paths are determined by going from landmark to landmark until the destination is reached. Travel may take place either through rooms or via hallways, but not anywhere else. This is an important restriction since walls are not explicitly defined. Corners are defined at each intersection of two or more hallways, and corners are defined as the only way to get from one hallway to another.

The best path between two points is calculated by performing a breadth-first search through all possible paths, until a complete path is found. The "best path" is therefore the one containing the fewest number of segments. Each potential path consists of a series of segments, and each segment is defined by (1) the landmark to which it goes, (2) the two points that identify the line segment (the previous point and the landmark's location), and (3) the prescribed action at that landmark. The prescribed action uses one of seven direction words: up, down (as in, "go down the elevator"), right, left, straight, through, and down (as in, "down the hall"). Figure 5 shows a representation of sample output from the path calculator.

```
Path [(333,460), (255,66)]
 - Segment [(333,460), (154,365)], RIGHT @ the door
 - Segment [(154,365), (154,169)], STRAIGHT @ the door
 - Segment [(154,169), (154,140)], RIGHT @ corner
 - Segment [(154,140), (169,140)], LEFT @ the glass doors
 - Segment [(169,140), (255,66)], -|- @ end
 = totDistance: 555

Path [(333,460), (532,445)]
 - Segment [(333,460), (477,365)], RIGHT @ the door
 - Segment [(477,365), (477,574)], LEFT @ the door
 - Segment [(477,574), (532,445)], -|- @ end
 = totDistance: 521
```

Figure 5: Sample text output from the path calculator

Paths between multiple floors are determined by finding a path to the nearest elevator, and then concatenating a path from the destination floor's elevator to the destination. Elevators are defined as the only method for getting from one floor to another.

V. iv. Direction Generation

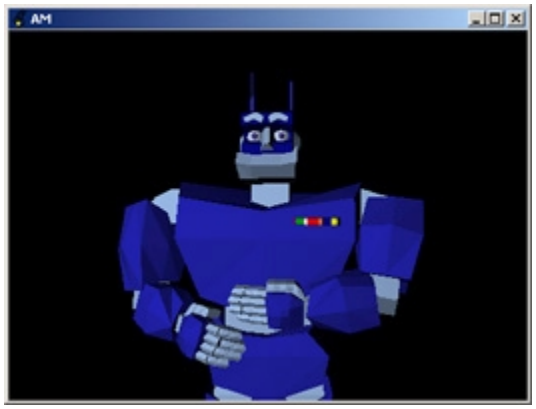
Once a path is calculated, that path serves as the basis for MACK's direction generation. From the direction-giving model (Figure 2), a probabilistic rule-based system was designed to govern MACK's direction-giving behaviors. Currently, the type of directions he gives depends on the phrasing of the user's request (Q1, Q2, or Q3). Q1 results in MB directions. Q2 leads to either SG directions (80%) or MB (20%). Q3 prompts RD, followed by SG directions 33% of the time. The generation of each of those direction methods mirrors the model defined in Chapter 4, which was based on the results of the two empirical studies.

For SG direction generation, a number of improvements were required. First, a new gesture library was created for the various new gestures. The data from both studies served as a basis for this gesture library. Eighteen gestures were added to coincide with direction words – four for "right" (left or right hand, pointing or using a flat hand gesture), four for "left," two for "up," two for "down," four for "straight," and two for "through" – as well as additional iconic gestures for various landmarks. In addition to arm and hand gestures, torso movement was also added, as the data showed that subjects turned their bodies in the direction they were referencing. New pointing gestures were also added to ensure unambiguous reference to landmarks within MACK's field of

vision, motivated by earlier research on deictic believability [Lester et al., 2000]. Figure 6 shows samples from the new gesture library.

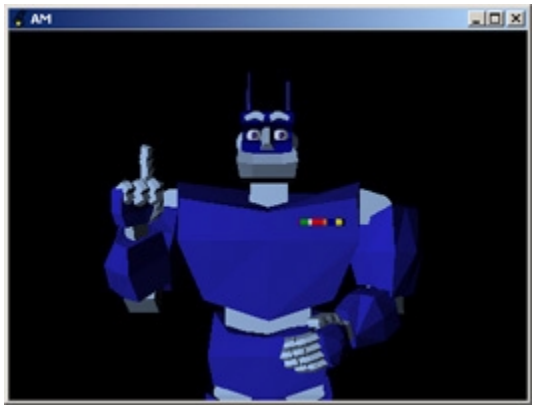
MB direction generation required significant additions to MACK's projector output. The projector map output was redesigned to coincide with the system's internal map representation. This meant that path and segment information could be directly expressed through the projector output. Pointing translated into a small circle highlighted at a specific point. To trace a path between points, MACK highlighted a line segment over time as though drawing a line from point to point. Rectangular areas could also be highlighted so that MACK could reference a specific destination (usually a room). With these additions, MACK was able to generate MB directions as subjects did, according to model described earlier.

Figure 6: Samples from new gesture library

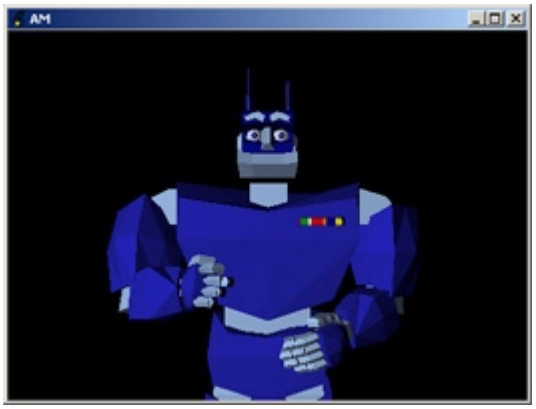


“make a right”

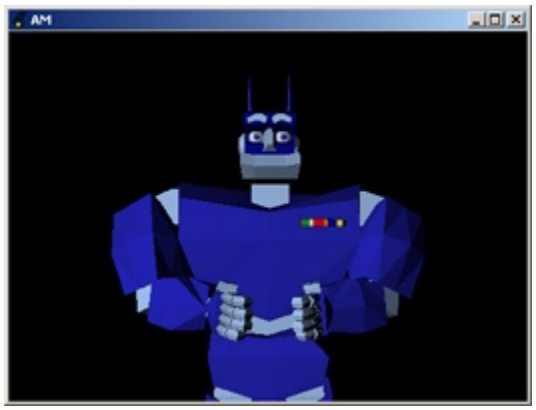
“make a right”



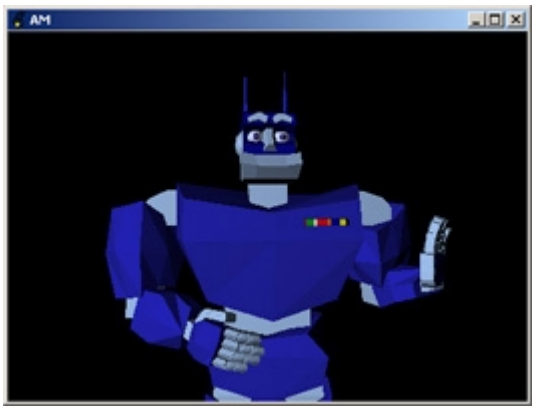
“go up”



“go straight”

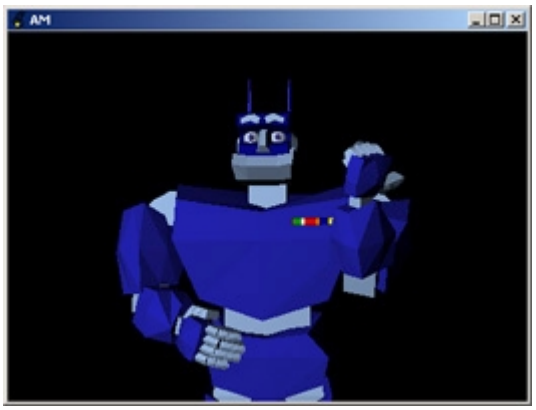


“walk through”



“you’ll see glass doors on your left”

“go to that door”



“go to this door behind me”

VI. Future Work

In evaluating MACK and observing user interaction with him, I found that people were very comfortable using the system. The use of an ECA made the kiosk approachable and welcoming, and users were able to interact with MACK without any prior instruction. Using natural language to communicate face-to-face with MACK was intuitive, however MACK did little to instruct users in cases of partial understanding, which made interactions somewhat strained at times.

Generally, users' behaviors appeared natural, as though they are interacting with another person. And users acted as though MACK demonstrated agency and was trustworthy in the information he conveyed. When MACK said something like, "Turn to your right and go to that door," while pointing to the door, users turned around and then turned back to MACK with a nod. MACK was also successful in engaging and entertaining users, as people would invent new questions simply to hear MACK's responses.

Regarding MACK's direction-giving capabilities, users were engaged by MACK's use of multiple modalities. His directions were generally clear and understandable, and in most cases users would nod and say "uh-huh" in recognition, demonstrating their understanding along the way. The system, however, was not monitoring these behaviors, so in cases where users expressed confusion, MACK did not stop to clarify. MACK's gestures appeared smooth and natural, and users commented that he gave directions well. However, MACK did a poor job directing the user's focus of attention. Users would sometimes look at the map after requesting directions, not

realizing that MACK was giving speech and gesture directions until he was almost halfway done.

With these observations in mind, there remain significant improvements to be made to the MACK system. A useful addition would be a dialogue manager that explicitly modeled turn-taking and the user's focus of attention. Such a dialogue manager would allow MACK to be more pro-active in engaging the user's attention and ensuring the user's understanding. An input device that could monitor users' gaze and nod behaviors would enhance the system further, and allow MACK to more precisely monitor the user's attention and understanding. With that information, MACK could tailor his direction-giving method to where the user's attention is at the time, or instead direct the user's attention to the proper place. And in cases of user confusion, for example, MACK could then stop to clarify the directions.

Better use of a comprehensive discourse history is also an important next step. With a discourse history, MACK could make use of previously mentioned locations when giving directions. For example, he could say, "To get to Room 368, you go the same way as to get to Room 320, except you make a left at the end instead of a right." In the same way, MACK could make use of previous path segments. Once MACK has given directions via the elevators, for example, he could then start subsequent sets of directions from the elevators, rather than each time explaining the entire path from start to finish.

In addition, speech recognition errors remain prevalent, especially with non-native English speakers and with any significant background noise. Future research will address some of these issues by relating more closely the speech recognition engine and

the dialogue planner. Shared knowledge between these systems would result in significant recognition improvements, as information from the dialogue planner could help limit the range of potential utterances to recognize. For example, if the speech recognition engine were to know that MACK just asked a yes/no question, its accuracy in recognizing the user's response would significantly improve.

Past research has focused on multimodal error recovery, but there remains significant progress to be made. Even simple feedback regarding partial understanding would be helpful to manage user input. For example, paraphrasers such as the one in McKeown's CO-OP, which formulates paraphrases based on given and new information [McKeown, 1983], could allow MACK to provide feedback as to the source of his confusion, which in turn would lead to improved follow-up interactions. Another possibility is requesting a specific modality in cases of partial understanding. Rather than responding, "Could you please repeat that," MACK can instead suggest, "I don't quite understand. Perhaps using the map would help clear up my confusion." Multimodal fusion, often suggested as a method for error correction, typically refers to synchronized redundancy across multiple modalities [Oviatt, 2000]. However, requests for asynchronous redundancy might provide similar results while appearing more natural to the user.

In addition to the MACK system, this thesis serves as a basis for interesting future work in the area of spatial intelligence presentation, and specifically direction generation. While my work is a first step in describing human direction-giving behavior, the subject deserves additional research. As described earlier, people give directions in different ways depending on how the question is phrased. There are surely other variables that

affect how directions are generated, such as the state of the common ground – e.g., what other locations have been discussed recently – or perhaps even direction-giver personality traits, and it would be interesting to explore these possibilities.

Shared reality is another area that deserves continued attention. As the next step for interfaces that currently employ virtual reality models to bridge the gap between user and interface, the concept of shared physical space provides a perhaps more effective approach toward the same end. The MACK system shares a physical map with its users, and other research has explored similar ways to share objects between users and virtual agents [Ryokai, Vaucelle, and Cassell, 2002; Cassell et al., 2000; Ishii and Ullmer, 1997]. This concept of a shared reality offers myriad opportunities for interesting future work exploring similar ways to blur the line between the real and virtual worlds.

VII. Conclusions

Presenting spatial information is an important issue in the design of a public information kiosk. This thesis serves as a first step in approaching some of the challenges involved in bridging the gap between the real and virtual worlds to move toward the ideal of an ECA seemingly immersed in the user's physical world. The described direction-giving model is motivated by human-to-human interactions and details the coordination of speech, gestures, and map-based reference.

The direction-giving framework has been implemented in MACK and uses the domain of the MIT Media Lab. As the framework is modular, it could easily be implemented in other interactive systems. Similarly, MACK's knowledgebase could be substituted for another locale and the direction-giving framework would still apply. With continued developments in the areas of spatial intelligence and embodied agents, I envision ECA Kiosks such as MACK offering a new level of service in public spaces.

VIII. References

[Billinghurst et al., 1996] Billinghurst, M., et al. *The Expert Surgical Assistant: An Intelligent Virtual Environment with Multimodal Input*. In *Medicine Meets Virtual Reality IV*. 1996. Amsterdam: IOS Press.

[Cassell, Bickmore, Vilhjálmsson, and Yan, 2000] Cassell, J., T. Bickmore, H. Vilhálmsson, and H. Yan. *More Than Just a Pretty Face: Affordances of Embodiment*. In *IUI 2000*. 2000. New Orleans, LA. pp. 52-59.

[Cassell et al., 1999] Cassell, J., T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. *Embodiment in Conversational Interfaces: Rea*. In *CHI '99*. 1999. Pittsburgh, PA: ACM. pp. 520-527.

[Cassell et al., 2000] Cassell, J., M. Ananny, A. Basu, T. Bickmore, P. Chong, D. Mellis, K. Ryokai, H. Vilhjálmsson, J. Smith, H. Yan. *Shared Reality: Physical Collaboration with a Virtual Peer*. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. April 4-9, 2000. Amsterdam, NL. pp. 259-260.

[Cassell et al., 2001] Cassell, J., T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan. *More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment*. In *Knowledge-Based Systems* 14. 2001: pp. 55-64.

[Cassell and Stocky et al., 2002] Cassell, J., T. Stocky, T. Bickmore, Y. Gao, Y. Nakano, K. Ryokai, D. Tversky, C. Vaucelle, H. Vilhjálmsson. *MACK: Media lab Autonomous Conversational Kiosk*. In *Imagina '02*. February 12-15, 2002. Monte Carlo.

[Chang, 1998] Chang, J. *Action Scheduling in Humanoid Conversational Agents*. M.S. Thesis in Electrical Engineering and Computer Science. Cambridge, MA: MIT.

[Christian and Avery, 2000] Christian, A.D. and B.L. Avery. *Speak Out and Annoy Someone: Experience with Intelligent Kiosks*. In *CHI '00*. 2000. The Hague, Netherlands: ACM.

[Emmorey, Tversky, and Taylor, 2000] Emmorey, K., B. Tversky, and H. A. Taylor. *Using Space to Describe Space: Perspective in Speech, Sign, and Gesture*. In *Spatial Cognition and Computation*. Ed. S. Hirtle. 2000. Vol. 2, Number 3. Kluwer Academic Publishers: The Netherlands. pp. 157-180.

[Feiner and McKeown, 1998] Feiner, S. and K. McKeown. *Automating the Generation of Coordinated Multimedia Explanations*. In *Readings in Intelligent User Interfaces*. Ed. M. Maybury and W. Wahlster. 1998. San Francisco: Morgan Kaufmann. pp. 89-97.

[Glass and Weinstein, 2001] Glass, J. and E. Weinstein. *SpeechBuilder: Facilitating Spoken Dialogue System Development*. In *EuroSpeech '01*. September, 2001. Aarlborg, Denmark

[Ishii and Ullmer, 1997] Ishii, H. and B. Ullmer. *Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms*. In *Conference on Human Factors in Computing Systems (CHI '97)*. March, 1997. Atlanta, GA: ACM. pp. 234-241.

[Kerpedjiev et al., 1998] Kerpedjiev, S., G. Carenini, N. Green, J. Moore and S. Roth. *Saying It in Graphics: from Intentions to Visualizations*. In *IEEE Symposium on Information Visualization*. October, 1998. Research Triangle Park, NC: pp. 97-101.

[Koda and Maes, 1996] Koda, T. and P. Maes. *Agents with Faces: The Effects of Personification of Agents*. In *IEEE Robot-Human Communication '96*. 1996. Tsukuba, Japan.

[Lawton, 2001] Lawton, C.A. *Gender and Regional Differences in Spatial Referents Used in Direction Giving.* Sex Roles, 2001. **44**(5/6): pp. 321-337.

[Lester et al., 2000] Lester, J. C., S. G. Towns, C. B. Callaway, J. L. Voerman, and P. J. FitzGerald. *Deictic and Emotive Communication in Animated Pedagogical Agents*. In *Embodied Conversational Agents*. Ed. J. Cassell, S. Prevost, J. Sullivan, and E. Churchill. 2002. Cambridge: MIT Press. pp. 123-154.

[Lovelace, Hegarty, and Montello, 1999] Lovelace, K.L., M. Hegarty, and D.R. Montello. *Elements of Good Route Directions in Familiar and Unfamiliar Environments*. In *Conference on Spatial Information Theory*. 1999. Stade, Germany: Springer-Verlag.

[Maybury, 1998] Maybury, M. *Planning Multimedia Explanations Using Communicative Acts*. In *Readings in Intelligent User Interfaces*. Ed. M. Maybury and W. Wahlster. 1998. San Francisco: Morgan Kaufmann. pp. 99-106.

[McKeown, 1983] McKeown, K. *Paraphrasing Questions Using Given and New Information*. In *American Journal of Computational Linguistics* 9.1. Jan-Mar, 1983. pp. 1-10.

[Michon and Denis, 2001] Michon, P.-E. and M. Denis. *When and Why Are Visual Landmarks Used in Giving Directions*. In *Conference on Spatial Information Theory*. 2001. Morro Bay, CA: Springer-Verlag.

[Oviatt, 2000] Oviatt, S. *Taming Recognition Errors With a Multimodal Interface*. In *Communications of the ACM* 43.9. September, 2000. pp. 45-51.

[Oviatt and Cohen, 2000] Oviatt, S. and P. Cohen, *Multimodal Interfaces That Process What Comes Naturally.* Communications of the ACM, 2000. **43**(3): pp. 45-53.

[Raisamo, 1999] Raisamo, R. *Evaluating Different Touch-based Interaction Techniques in a Public Information Kiosk.* In *Conference of the Computer Human Interaction Special Interest Group of the Ergonomics Society of Australia*. 1999. Charles Stuart University: pp. 169-171.

[Reeves and Nass, 1996] Reeves, B. and C. Nass, *The Media Equation: how people treat computers, televisions and new media like real people and places*. 1996. Cambridge: Cambridge University Press.

[Ryokai, Vaucelle, and Cassell, 2002] Ryokai, K., C. Vaucelle, and J. Cassell. *Literacy Learning by Storytelling with a Virtual Peer*. In *Computer Support for Collaborative Learning*. 2002.

[Steiger and Suter, 1994] Steiger, P and B. Ansel Suter. *MINELLI – Experiences with an Interactive Information Kiosk for Casual Users.* In *UBILAB '94*. 1994. Zurich: pp. 124-133.

[Stocky and Cassell, 2002] Stocky, T. and J. Cassell. *Shared Reality: Spatial Intelligence in Intuitive User Interfaces*. In *Intelligent User Interfaces*. Jan 13-16, 2002. San Francisco, CA: pp. 224-225.

[Sumi and Mase, 2000] Sumi, Y. and K. Mase. *Supporting Awareness of Shared Interests and Experiences in Community*. In *ACM CSCW*. 2000. Philadelphia, PA.

[Taylor and Tversky, 1996] Taylor, H. A. and B. Tversky. *Perspective in Spatial Descriptions*. In *Journal of Memory and Language*, 1996. **35**(3): pp. 371-391.

[Tversky and Lee, 1999] Tversky, B. and P.U. Lee. *Pictorial and Verbal Tools for Conveying Routes*. In *Conference on Spatial Information Theory*. 1999. Stade, Germany.

[Waters and Levergood, 1993] Waters, K. and T. M. Levergood, *DECface: An Automatic Lip-Synchronization Algorithm for Synthetic Faces*. 1993, Digital Equipment Corporation Cambridge Research Laboratory.

[Wahlster et al., 1993] Wahlster, W., E. André, W. Finkler, H.-J. Profitlich and T. Rist. *Plan-based integration of natural language and graphics generation*. In *Artificial Intelligence* 63. 1993. pp. 387-427.