# "Move the Couch Where?" : Developing an Augmented Reality Multimodal Interface

Sylvia Irawati [a, c], Scott Green [b, d] , Mark Billinghurst [b], Andreas Duenser [b], Heedong Ko [a]

[a] Imaging Media Research Center, Korea Institute of Science and Technology

[b] Human Interface Technology Laboratory New Zealand, University of Canterbury

[c] Department of Human Computer Interaction and Robotics, University of Science and Technology

[d] Department of Mechanical Engineering, University of Canterbury

## Abstract

This paper describes an augmented reality (AR) multimodal interface that uses speech and paddle gestures for interaction. The application allows users to intuitively arrange virtual furniture in a virtual room using a combination of speech and gestures from a real paddle. Unlike other multimodal AR applications, the multimodal fusion is based on the combination of time-based and semantic techniques to disambiguate a users speech and gesture input. We describe our AR multimodal interface architecture and discuss how the multimodal inputs are semantically integrated into a single interpretation by considering the input time stamps, the object properties, and the user context.

**CR Categories and Subject Descriptors:** H.5.1 (Multimedia Information Systems): Artificial, augmented, and virtual realities; H.5.2 (User Interfaces): Auditory (non-speech) feedback, Graphical user interfaces (GUI), Interaction styles, Natural language, Voice I/O.

**Additional Keywords:** multimodal interaction, paddle gestures, augmented reality, speech input, gesture input.

## 1 Introduction

Augmented Reality (AR) is an interface technology that uses head worn or handheld displays to allow users to see virtual images superimposed over the real world. In AR applications three-dimensional computer graphics appear to be fixed in space or attached to real objects.

AR techniques have been shown to be useful in many application areas such as engineering [1], education [2] and entertainment [3]. However, there is still research that needs to be conducted on the best way to interact with AR content. Researchers have explored a wide variety of interaction methods including using mouse input [4], pen and tablet [5], magnetic tracking [6], real objects [7] and even natural gesture input [8].

In this paper we present an example of how speech and gesture input can be used to manipulate virtual content in an AR application. Combined speech and pen or mouse input has been studied extensively for desktop computing, and there are a number of examples of gesture and speech interfaces for immersive virtual environments. Our work differs from these because it involves speech and gesture interaction with virtual objects in the real world. It is also different from the few previous multimodal AR interfaces because it uses a combination of time-based and domain semantics to disambiguate the users speech and gesture input and respond to the command.

In this paper, we first review related work and then describe our multimodal architecture. We then discuss the implementation of our multimodal system, with a main focus on the fusion strategies. Finally we conclude and provide possible directions for future research.

## 2 Related work

Our work is motivated by earlier research work in multimodal interfaces, VR and AR interfaces. From this research we can learn important lessons that can inform the design of our system.

One of the first interfaces to combine speech and gesture recognition was the Media Room [9]. Designed by Richard Bolt, the Media Room allowed the user to sit inside the computer interface and interact with the computer through voice, gesture and gaze.

Since Bolt's work, there have been many two dimensional desktop interfaces developed that show the value of combining speech and gesture input. For example, Boeing's "Talk and Draw" [10] application allowed users to draw with a mouse and use speech input to change interface modes. Similarly Cohen's QuickSet [11] combined speech and pen input for drawing on maps in command and control applications.

Multimodal interaction has also been used in Virtual Reality (VR) environments. Weimer and Ganapathy [12] developed a prototype virtual environment that incorporated a data glove and simple speech recognizer. Laviola [13] investigated the use of whole-hand gestures and speech to create, place, modify, and manipulate furniture and interior decorations. Ciger et al. [14] presented a multimodal user interface that combined a magic wand with spell casting. The user could navigate in the virtual environment, grab and manipulate objects using a combination of speech and the magic wand.

More recent works have used ontology definition for complementing the semantic models. Latoschik [15] presented a framework for modeling multimodal interactions, which enriched the virtual scene with linguistic and functional knowledge about the objects to allow the interpretation of complex multimodal utterances. Holzapfel et al. [16] presented multimodal fusion for

[a, c] 39-1 Hawolgok-dong, Seoul, 136-791, South Korea
{sylvi, ko} @ imrc.kist.re.kr

[b, d] Private Bag 4800, Christchurch, New Zealand
{mark.billinghurst, scott.green, andreas.duenser}@hitlabnz.org

natural interaction with a humanoid robot. Their multimodal fusion is based on an information-based approach by comparing object types defined in the ontology.

Although AR interfaces are closely related to immersive VR environments, there are relatively few examples of Augmented Reality applications that use multimodal input. McGee and Cohen [17] created a tangible augmented reality environment that digitally enhanced the existing paper-based command and control capability in a military command post. Heidemann et al. [18] presented an AR system designed for online acquisition of visual knowledge and retrieval of memorized objects. Olwal et al. [19] introduced a set of statistical geometric tools, SenseShapes, which use volumetric regions of interest that can be attached to the user, providing valuable information about the user interaction with the AR system. Kaiser et al. [20] extended Olwal's SenseShapes work by focusing on mutual disambiguation between input channels (speech and gesture) to improve interpretation robustness. This research shows that statistical modelling and knowledge about user actions in the real world can be used to resolve language ambiguities in multimodal AR applications.

Our work differs from previous AR multimodal interfaces in that it uses a combination of time-based and domain semantics to interpret the users speech and gesture input.

## 3 Developing a Multimodal Interface

Our system is a modified version of the VOMAR application [21] for tangible manipulation of virtual furniture in an AR setting. The goal of this application is to allow people to easily arrange AR content using a natural mixture of speech and gesture inputs.

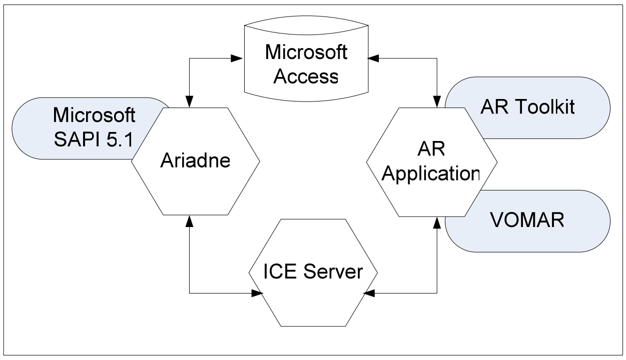


Figure 1 System architecture

Figure 1 illustrates the components of our multimodal system architecture. Ariadne [22] is a spoken dialog system that uses the Microsoft Speech API 5.1 as the speech recognition engine. AR Application is an application that allows a user to interact with the system using paddle gestures and speech. It is responsible for receiving the speech commands from Ariadne, recognizing paddle gestures, and fusing the speech and paddle gesture input into a single interpretation.

The AR Application is based on the ARToolkit [23] library and the VOMAR paddle gesture library. Ariadne and the AR Application communicate with each other using the middleware ICE (Internet Communication Engine) [24]. A Microsoft Access database is used to store the object descriptions. This database is used by Ariadne to facilitate rapid prototyping of speech grammar, as well as by AR Application to enable multimodal fusion.

The AR application involves the manipulation of virtual furniture in a virtual room, although the multimodal interface can be applied to various domains. When the user looks at each of the menu pages through a head mounted display with a camera attached to it (Figure 2), they see different types of virtual furniture on the pages (Figure 3), such as a set of chairs or tables. Looking at the workspace, a large piece of paper with specific fiducial markers, they see the virtual room. The user can then pick objects from the menu pages and place them in the workspace using paddle and speech commands.

The system also provides visual and audio feedback to the user by showing the speech recognition result and the object bounding box when the paddle touches the object. In addition, audio feedback is given after the speech and paddle gesture command, so the user can immediately recognize if there is an incorrect result from the speech or gesture recognition system.

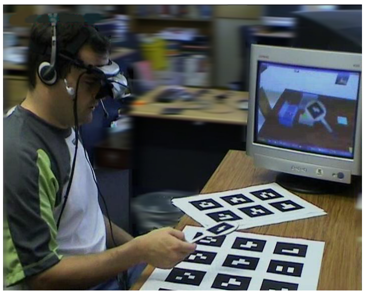


Figure 2 A user wears a HMD and holds a paddle

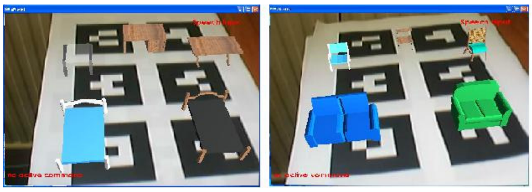
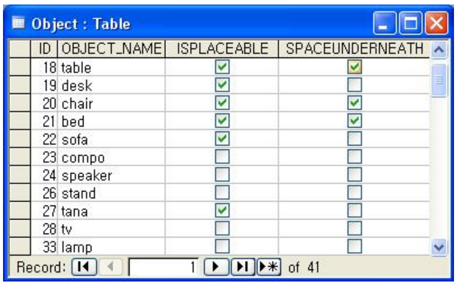


Figure 3 Virtual menus that contain a set of virtual furniture

### 3.1 Ariadne + speech input

A critical component of using speech in a computer interface is the speech recognition engine. We use the Ariadne spoken dialog system, which contains all the necessary components to build an effective spoken dialogue system. To enable rapid prototyping of a grammar structure, objects are imported from an SQL database, such as the one shown in Figure 4.



Object : Table

| ID | OBJECT_NAME | ISPLACEABLE | SPACEUNDERNEATH |
|---|---|---|---|
| 18 | table | ☑ | ☑ |
| 19 | desk | ☑ | ☐ |
| 20 | chair | ☑ | ☑ |
| 21 | bed | ☑ | ☑ |
| 22 | sofa | ☑ | ☐ |
| 23 | compo | ☐ | ☐ |
| 24 | speaker | ☐ | ☐ |
| 26 | stand | ☐ | ☐ |
| 27 | tana | ☑ | ☐ |
| 28 | tv | ☐ | ☐ |
| 33 | lamp | ☐ | ☐ |

Record: 1 of 41

Figure 4 Object descriptions stored in the SQL database

The database table *Object* stores objects that the user can interact with. The `OBJECT_NAME` column corresponds to the virtual objects that can be manipulated by the user and is used to generate the speech grammar. The `ISPLACEABLE` and `SPACEUNDERNEATH` fields are used to perform multimodal fusion by determining the possible location referred to by deictic referencing, such as 'here' or 'there'. An object with the property

ISPLACEABLE set to *true* means that another object can be put on top of it. An object with the property SPACEUNDERNEATH *true* means that an object can be placed under it. By knowing these properties, the ambiguity contained in deictic referencing can be easily resolved.

The following are some commands recognized by the system:

- Grab Command "Select a desk": to select a virtual object from the menu or workspace, and place it on the paddle.
- Place Command "Place here": to place the attached object at the paddle location in the workspace.
- Move Command "Move the couch": to attach a virtual object in the workspace to the paddle so that it follows the paddle movement.

### 3.2 VOMAR - Paddle Gestures

In the VOMAR application the paddle is the main interaction device. This is a real object with an attached tracking marker whose position and orientation is tracked using the ARToolKit software. The paddle allows the user to make gestures to interact with the virtual objects. A range of static and dynamic gestures is recognized by tracking the motion of the paddle. The gestures correspond to virtual content manipulation (Table 1).

Table 1 The VOMAR Paddle Gestures

| | |
|---|---|
| **Static Gestures** | Paddle proximity to object<br>Paddle tilt/inclination |
| **Dynamic Gestures** | Shaking: side to side motion of paddle<br>Hitting: up and down motion of paddle<br>Pushing object |

### 3.3 Fusion strategies

To understand the combined speech and gesture, the system must fuse inputs from both streams into a single understandable command. This is accomplished by combining the time each input occurs. The paddle and speech input can only be considered for fusion if the input time stamps from both input streams are within a certain time threshold of each other. Object properties can further be used to disambiguate speech and gesture input.

When a speech interpretation result is received from Ariadne, the AR Application checks whether the paddle is in view. The speech command is only active when the paddle is in sight. Next, depending on the speech command type and the paddle pose, a specific action is taken by the system. For example, consider the case when the user says "grab this" while the paddle is placed over the menu page to grab a virtual object. The system will test the paddle proximity to the virtual objects. If the paddle is close enough to an object, the object will be selected and attached to the paddle. If the paddle is not close enough, the object will not be selected. However, the grab command is still active for five seconds after the user finishes speaking, allowing the user to move the paddle closer to the desired object to select it. After five seconds, if the user wants to reactivate the previous command, the user must repeat the speech command.

As mentioned previously, the object properties described in the database are used to fuse multimodal inputs. The properties ISPLACEABLE and SPACEUNDERNEATH are used to resolve deictic terms contained in the speech commands given by the user. For example, if the user says "put here" while touching a virtual rack with the paddle, the possible locations referred to by 'here' are 'on the rack' or 'under the rack'. By checking the object properties of the rack, e.g. SPACEUNDERNEATH is false and ISPLACEABLE is true, the system understands that 'here' refers to the position 'on top of the rack'.

In case the object properties cannot disambiguate user input, the position of the paddle is used by the system. For example, when 'here' could mean 'on the table' or 'under the table' and the properties SPACEUNDERNEATH and ISPLACEABLE of the table both are true, the system checks the paddle in the z (up-down) direction. If the z position of the paddle is less than a threshold value (in this case the height of the table), the system understands 'here' as 'under the table'. On the other hand, if the z position of the paddle is greater than the height of the table, the system understands 'here' as 'on the table'.

Another factor that needs to be considered in multimodal fusion is the user interaction context, current user viewpoint and paddle position. If there is more than one object near the paddle, the system will select the object closest to the paddle. The current user viewpoint is also used when the user command contains spatial predicates, such as, behind, in front of, beside, etc. For example, given the command "put it behind the table", there are four possible regions relative to the object table, referred to by the spatial predicate 'behind' (see Figure 5). If the user viewpoint is in region one, the phrase 'behind the table' refers to region three, but if the user viewpoint is in region two, the phrase 'behind the table' refers to region four. Therefore, issuing the same command from different user viewpoints can provide different results. For example, in Figure 6, the user viewpoint relative position to the table is in region one. Given the input "put the chair behind the table", the coordinates for 'chair' are calculated as follows:

$$x_{chair} = x_{table}$$
$$y_{chair} = y_{table} + y\_size_{table} + y\_size_{chair}$$
$$z_{chair} = z_{table}$$

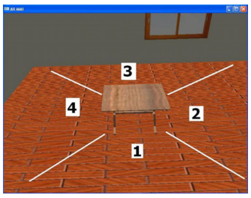


Figure 5 The possible regions relative to the object (table)

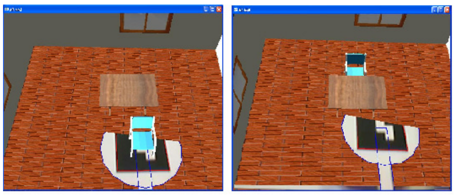


Figure 6 Given the speech input: "put the chair behind the table"

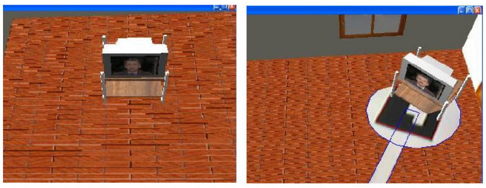


Figure 7 Given speech input: "put on the rack" and "move the rack"

When one object is on top of another object, the relationship between those objects is maintained by the system. Therefore, when the user moves the base object, objects on top of that object must move with the base object. For example, if the user places a TV on the table, when the user moves the table, the TV moves with the table (see Figure 7).

## 4 CONCLUSION

In this paper, we described an augmented reality multimodal interface that uses the combination of speech and paddle gestures to interact with the system. The interface is designed to effectively and easily arrange AR content using a natural mixture of speech and gesture input. Our work is different from previous AR multimodal interfaces in several important ways. Unlike other AR multimodal interfaces, our multimodal fusion uses a combination of time-based and domain semantics to disambiguate the users speech and gesture input. In combining the multimodal inputs, the system considers the object properties, the user interaction context, the paddle position and the user viewpoint. The object properties, which are defined in a database, are shared by the spoken dialogue system and the AR application. Our system also maintains object relationships in the workspace. Therefore, moving one object may cause the movement of other objects.

Early results from a pilot user study, not discussed in this paper, found that combining speech and paddle gestures can improve performance when manipulating virtual objects. Speech is suitable for system control and gestures are suitable for spatial input such as direct interaction with the virtual objects. Contextual knowledge may resolve ambiguous input. The proper location referred to by deictic terms, such as 'here', and 'there', can be resolved by knowing the object properties, the current user viewpoint and current paddle position.

There are several areas of future research that we intend to explore. First, new paddle gestures could be introduced to optimize the intuitiveness of interaction. Second, the speech grammar could be extended to support a dialogue making the system more interactive. Finally, the multimodal interface could be extended to other AR domains to determine how effectively the benefits we have seen can be extended to other fields.

## 5 ACKNOWLEDGEMENT

The first and last authors are funded by the Ubiquitous Autonomic Computing and Network Project, the Ministry of Information and Communication (MIC) 21st Century Frontier R&D Program in Korea.

## REFERENCES

[1] Anthony Webster, Steven Feiner, Blair MacIntyre,William Massie, and Theodore Krueger. Augmented reality in architectural construction, inspection and renovation. In *Proceedings of .ASCE Third Congress on Computing in Civil Engineering*, Anaheim, CA, June 1996, pp. 913-919.

[2] Hannes Kaufmann. Collaborative Augmented Reality in Education. Keynote Speech at Imagina Conference, 2003.

[3] Istvan Barakonyi and Dieter Schmalstieg. Augmented Reality Agents in the Development Pipeline of Computer Entertainment. In *Proceedings of the 4th International Conference on Entertainment Compute*, September 2005.

[4] Christian Geiger, Leif Oppermann, and Christian Reimann. 3D-Registered Interaction-Surfaces in Augmented Reality Space. *In Proceedings of 2nd IEEE International Augmented Reality Toolkit Workshop*, 2003.

[5] Zsolt Szalavári and Michael Gervautz. The Personal Interaction Panel – A Two-Handed Interface for Augmented Reality. In *Proceedings of EUROGRAPHICS, Computer Graphics Forum*, volume 16, issue 3, 1997, pp. 335-346.

[6] Kiyoshi Kiyokawa, Haruo Takemura and Naokazu Yokoya. A Collaboration Support Technique by Integrating a Shared Virtual Reality and a Shared Augmented Reality. *In Proceedings of IEEE International Conference on Systems Man and Cybernetics*, 1999, pp. 48-53.

[7] H. Kato, M. Billinghurst, I. Poupyrev, N. Tetsutani, and K. Tachibana. Tangible Augmented Reality for Human Computer Interaction. In Proceedings of Nicograph 2001, Nagoya, Japan.

[8] V. Buchmann, S. Violich, M. Billinghurst and A. Cockburn. FingARtips. Gesture Based Direct Manipulation in Augmented Reality. In *Proceedings of 2nd International Conference on Computer Graphics and Interactive Techniques*, 2004, pp. 212-221.

[9] Richard A. Bolt. Put-That-There: Voice and Gesture at the Graphics Interface. *In Proceedings of ACM SIGGRAPH*, 1980, *Computer Graphics*, volume 14, pp 262-270.

[10] M. W. Salisbury, J. H. Hendrickson, T. L. Lammers, C. Fu, S. A. Moody. Talk and Draw: Bundling Speech and Graphics. *IEEE Compute,* volume 23, issue 8, August, 1990, pp. 59-65.

[11] P.R. Cohen, M. Johnston, D.R. McGee, S.L. Oviatt, J.A. Pittman, I. Smith, L. Chen, and J. Clow. Quickset: Multimodal Interaction for Distributed Applications. In *Proceedings of the Fifth Annual International Multimodal Conference*, 1997, pp. 31-40.

[12] D. Weimer and S.K. Ganapathy. A Synthetic Visual Environment with Hand Gesturing and Voice Input. In *Proceedings of ACM Conference on Human Factors in Computing Systems*, 1989, pp. 235-240.

[13] Joseph J. Laviola Jr. Whole-Hand and Speech Input in Virtual Environments. Master Thesis, Brown University, 1996.

[14] Jan Ciger, Mario Gutierrez, Frederic Vexo, Daniel Thalmann. The Magic Wand. In Proceedings of the 19th Spring Conference on Computer Graphics, 2003, pp. 119-124.

[15] M.E. Latoschik, and M. Schilling. Incorporating VR Databases into AI Knowledge Representations: A Framework for Intelligent Graphics Applications. In *Proceedings of the 6th International Conference on Computer Graphics and Imaging*, 2003.

[16] Hartwig Holzapfel, Kai Nickel, and Rainer Stiefelhagen. Implementation and Evaluation of a Constraint-based Multimodal Fusion System for Speech and 3D Pointing Gestures. In Proceedings of the 6th International Conference on Multimodal Interfaces, 2004, ACM Press, pp. 175-182

[17] David R. McGee and Philip R. Cohen. Creating Tangible Interfaces by Augmenting Physical Objects with Multimodal Language. In *Proceedings of the 6th International Conference on Intelligent User* Interfaces, 2001, pp. 113-119.

[18] Gunther Heidemann, Ingo Bax, Holger Bekel, Multimodal Interaction in an Augmented Reality Scenario. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, 2004, pp. 53-60.

[19] Alex Olwal, Hrvoje Benko, and Steven Feiner. SenseShapes: Using Statistical Geometry for Object Selection in a Multimodal Augmented Reality System. In *Proceedings of The Second IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2003)*. October 2003, pp. 300–301.

[20] Ed Kaiser, Alex Olwal, David McGee, Hrvoje Benko, Andrea Corradini, Li Xiaoguang, Phil Cohen and Steven Feiner. Mutual Dissambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality. In *Proceedings of The Fifth International Conference on Multimodal Interfaces (ICMI 2003)*, 2003, pp. 12–19.

[21] H. Kato, M. Billinghurst, I. Poupyrev, K. Imamoto and K. Tachibana. Virtual Object Manipulation on a Table-Top AR Environment. In *Proceedings of the International Symposium on Augmented Reality (ISAR 2000)*, October 2000, pp. 111-119.

[22] Matthias Denecke. Rapid Prototyping for Spoken Dialogue Systems. In *Proceedings of the 19th international conference on Computational Linguistics*, volume 1, 2002, pp. 1-7.

[23] ARToolKit, http://www.hitl.washington.edu/artoolkit

[24] ICE, http://www.zeroc.com/ice.html