Using Contextual Knowledge to Inform a Multimodal Interactive System

Elizabeth Harte and **Prof. Raymond Jarvis**
Monash University, Australia
elizabeth.harte@monash.edu.au, raymond.jarvis@monash.edu.au

Abstract

An essential feature for personal robots in domestic environments is the quality of its interaction. This work describes a Contextually Informed MultiModal Integrator, CIMMI, that fuses speech and symbolic gestures probabilistically and is informed by contextual knowledge in an assistive technology robotic application. Symbolic gestures are gestures that have semantic meaning, such as a wave gesture meaning 'hello'.

Contextual knowledge is used to resolve ambiguities associated with object and location words in a command, where it is defined as both conversational and situational. Conversational contextual knowledge uses the dialogue history to resolve ambiguities in the selected command. Situational contextual knowledge consists of the last known locations of the user, the robot and a list of objects that exist in the environment.

The accuracy of the speech recognition system alone (53%) was compared to using the speech with contextual knowledge, both conversational and situational, increasing the accuracy of the system to 72%. The accuracy of the system increased again to 76% with the addition of gestures as well.

1 Introduction

The design and development of robots to assist the elderly and infirm has been broadly researched in recent years, as demographics indicate there will be a greater proportion of elderly in the future [Pollack, 2005]. A robot carer could assist an elderly person in a domestic environment, allowing that person to be independent and remain in their home while still having some basic support and security. A crucial requirement of a robot that interacts with humans is that it correctly recognizes what is communicated to it, interprets the user's intention and acts on it. To effectively do this, a robot should demonstrate both spatial and transactional intelligence. Spatial intelligence is being able to understand and navigate a space, as well as recognizing objects and how to manipulate them. Transactional intelligence is the capacity to communicate with a person in a natural way. Natural communication is essential for a robot dealing with untrained people, but unstructured environments make recognizing speech and gestures ambiguous [Hüwel and Wrede, 2006] [Stiefelhagen *et al.*, 2007]. Some of these ambiguities can be resolved using mutual disambiguation, where errors in one mode of input can be resolved by using partial information of another input [Oviatt, 2003]. Contextual knowledge, such as conversational and situational context, can also resolve ambiguities in a similar way. Conversational context is built by previous dialogue and the situational context includes spatial information.

In this work, the transactional intelligence consists of speech and gesture recognition as well as the use of dynamic contextual knowledge. These components are combined probabilistically in CIMMI, a Contextually Informed MultiModal Integrator. The spatial intelligence in this work is a simple navigation and obstacle avoidance system and an object recognition system. These were implemented to demonstrate the successes, and failures, of CIMMI on Wakamaru, a humanoid robot.

In the next section, we will outline work related to multimodal communication and the use of contextual knowledge, followed by outlining our own contributions in Section 3. We will describe and discuss our experiments and demonstrations in Section 4. Lastly in Section 5, we will outline our conclusions and our future work for this project.

2 Related Work

Multimodal integration or fusion is the challenge of taking many different modes of input, such as speech, lip tracking, gestures, posture, head pose etc, and merg-

ing all this information to create a richer interpretation than any single mode may have provided [Oviatt, 1999]. It has been proved that this fusion of inputs can resolve errors that may occur in a single input by using partial information of another input - a concept called mutual disambiguation, as defined and proved by Sharon Oviatt in [Oviatt, 1999].

Speech is often used as the primary mode of communication for people so this is often the case for both human-computer (HCI) and human-robot interactions (HRI) [Gullberg, 1999] [Hüwel and Wrede, 2006] [Stiefelhagen *et al.*, 2007]. People also use different types of gestures in conjunction with speech to communicate their intentions [Gullberg, 1999]. Symbolic gestures have a direct meaning and can be understood without context from additional speech [Koons and Sparrell, 1994].

In addition to gestures, contextual knowledge is instinctively used by people to communicate with each other. Some of the key contexts used in HRI systems are conversational and situational. Conversational contextual knowledge is where the previously recognised commands contribute to selecting the current command or to resolve ambiguous commands by combining information provided by the speaker to form an unambiguous command [Bischoff and Graefe, 2004] [Hüwel and Wrede, 2006] [Stiefelhagen *et al.*, 2007]. Situational contextual knowledge is where the location of the agents involved contributes to selecting a current command [Iba *et al.*, 2002]. This is called 'situatedness' in speech recognition applications [Gorniak and Roy, 2005] and has been defined a number of ways in robotics, including by plan recognition [Gorniak and Roy, 2005] and information about objects in the environment [Stiefelhagen *et al.*, 2007].

In addition to the types of modes and knowledge that are fused multimodally, how they are fused is also an important field of research. QuickSet was the first system to use a unification based system - a logic based method for combining the meanings from two modes into a single common interpretation [Cohen *et al.*, 1997]. It was extended to include an associative map to represent legal semantic combinations between all the defined speech and pen based inputs, as well as formalizing the integration process probabilistically [Wu *et al.*, 1999]. More recently, the approaches developed for multimodal fusion have varied greatly, including Russ and others' [Russ *et al.*, 2005] semantic network fusion technique and a new hierarchical semantic representation as defined by [Ammicht *et al.*, 2007] and [Potamianos *et al.*, 2007].

In the field of robotics, multimodal integration has featured only recently in HRI. Some multimodal robotic systems include Albert [Rogalla *et al.*, 2002], Iba and others' vacuum cleaner [Iba *et al.*, 2002], BIRON [Hüwel and Wrede, 2006] and ARMAR III [Stiefelhagen *et al.*, 2007]. Each of these systems used pointing gestures, with and without gloves, to support their speech recognition which were then unified semantically. The multimodal fusion in ARMAR III, one of the most recent HRI systems, only references the gesture recognition system if there is a need of disambiguation because of the low accuracy [Stiefelhagen *et al.*, 2007], an idea supported by [Eisenstein and Christoudias., 2004]. They also used contextual knowledge to identify which object was pointed at and have plans to include dialogue history for user identification.

Figure 1: Wakamaru with Bumblebee camera on head, HokuyoURG laser range finder at middle and laptop on back.

Based on these implementations of multimodal technology for robotic platforms, there are many areas to explore.

3 System Overview

Wakamaru, a humanoid robot made by Mitsubishi Heavy Industries, was selected as our robotic platform so it can interact with humans in a human way. Wakamaru has a wheeled base, a pair of articulated arms, each with four degrees of freedom, and a head with three degrees of freedom. It also has an array of onboard infrared and ultrasonic sensors for local obstacle avoidance and touch sensors on the shoulders and in the hands. The robot is able to localize itself with a set of infrared reflectors positioned at known locations on the ceiling using the panoramic camera on the top of its head. This robot has unarticulated hands preventing retrieval of objects in our demonstrations.

Processing of the speech, vision and laser depth data was carried out on a separate 2.66GHz laptop computer, which is attached Wakamaru's back and communicates with Wakamaru via a wireless router. Speech is detected through a microphone that the user wears. A Point Grey Research BumbleBee stereo camera was attached to the front of the robot's head and could produce disparity and colour images at 15fps. Because it is difficult to find

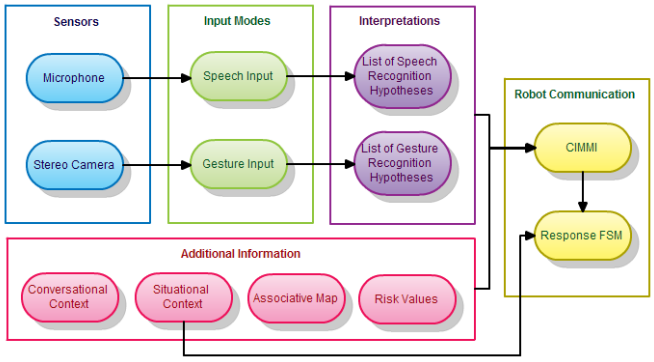


Figure 2: Overview of Multimodal Communication.

stereo matches on visually bland regions, we also used a HokuyoURG Laser Range Finder to return a single stripe of depth data for up to a 4m range (Fig. 1).

A user would communicate with Wakamaru by speech and dynamic symbolic gestures to complete specific tasks. The speech recognizer detects a single utterance, triggering the gesture recognizer to return the last detected gesture. Retrieving the last recognized gesture, rather than detecting the current one, is done because previous studies showed that when any meaningful gestures were made, 85% of the time it was accompanied by a spoken keyword temporally aligned during and after the gesture [Kettebekov and Sharma, 2001] and that gestures could precede the key word by as much as four seconds [Oviatt, 2003]. The speech and gesture recognizers generated n-best lists of recognition hypotheses, where $n = 5$ for speech (the 5 most likely recognized phrases) and $n = 3$ for gestures, as there are only three gestures currently implemented. Each parsed input on the lists would have a probability of correctness value which would be sent to the multimodal integrator.

CIMMI selects the best speech/gesture pair of inputs based on mutual disambiguation, the probability of correctness values, the spatial and contextual knowledge and a predefined risk value associated with each word. This pair is checked for ambiguities according to some predefined rules. If an ambiguity is found then the robot will ask the user for more information. If there is no ambiguity, then the speech/gesture pair are passed into the responses finite state machine so the robot can appropriately respond to the input (Fig. 2).

3.1 Interaction Modes

Speech Recognition

Speech is often used as the main mode of HRI in multimodal systems because it is the primary form of communication between people. This speech system would capture a single utterance and generate at most five possible recognitions. The confidence values for each phrase, as generated by the speech engine, were scaled so all the values sum to 1.0. The speech accuracy of this system was 79% for the first ranked recognition [Harte, 2008].

Each hypothesis is parsed by the developed system using a domain dependent vocabulary, where each keyword has a risk value, representing the value of certain words semantically, where the risk values are higher for words where misinterpretation could have seriously negative consequence, such as if the words 'help' and 'stop' were misrecognised, and a list of associative map values relating the word to each of the gestures, similar to [Wu *et al.*, 1999]. For example, the gesture 'Come' is contradictory to the word 'Go', so the corresponding associative map value is 0 (See Table 1). Also, the utterance 'What's the time?', represented as 'Time' in Table 1, is not complimentary or contradictory to any gesture, so its associative map values are always 1. This associative map is a key factor of mutual disambiguation in CIMMI [Harte and Jarvis, 2007].

Once all the recognition hypotheses have been parsed, the parsed recognition structures and probabilities are sent to CIMMI, which a hypothesis has the form *[action, colour, object, location, probability]* [Harte, 2008].

Symbolic Gesture Recognition

Our gesture recognition system recognizes symbolic dynamic gestures using OpenCV's Camshift tracking and 3D geometry. Symbolic gestures were used instead of the pointing deictic gestures because symbolic gestures can substitute, and reinforce, an action word semantically. The three symbolic dynamic gestures recognized were 'wave', 'come here' and 'go away'. Each of these gestures have a few semantic meanings,

- **Come** - bring or come
- **Go** - get, go, call or answer
- **Wave** - hello, goodbye or help

To perform gesture recognition, a frontal face must be detected in the scene as it is assumed a person looks at who they're gesturing to. The gestures were segmented by skin colour and motion (see Fig. **??**), tracked using OpenCV's Camshift tracking and classified using a previously trained k-Nearest Neighbour classifier [Harte and Jarvis, 2007]. The distances from the captured gesture to the eight nearest neighbours were summed for each class (come, go and wave), normalized so they sum to 1.0, then inverted so that the class with the shortest distance has the highest probability. The camera server then sends the probability for each class through to CIMMI in response to its request. This information is sent as a list of recognition hypotheses, where a hypothesis has the form *[class, probability]* [Harte, 2008].

Table 1: Sample Values from the Associative Map

Words	Time	Help	Come	Go	Stop	Hello	Cup	Desk
Come Gesture	1	2	2	0	0	1	2	2
Go Gesture	1	2	0	2	0	1	2	2
Wave Gesture	1	2	1	1	1	2	1	1

(a) A binary image of the skin segmentation

(b) A greyscale image of the motion segmentation

(c) The conjunction of (a) and (b)

(d) The colour capture with the detected regions identified

Figure 3: A single frame from a 'wave' gesture using skin and motion segmentation

3.2 Contextual Knowledge

Having described both the speech and gesture recognition systems to be combined, the contextual knowledge needs to be discussed. As people use contextual knowledge to help interpret what was communicated to them, so do the transactional and spatial intelligences implemented in this system. The implemented contextual knowledge consists of conversational and situational knowledge.

The conversational context is based on a simple assumption - in conversation, if a person had been talking about a 'yellow cup' then conversational context allows them to simplify additional references to the same object by saying 'the cup' without having to specify the implied 'yellow' again [Harte, 2008]. Additionally, this knowledge can be used to solve recognition errors where a piece of information may be missing or misrecognized in a command hypothesis, so the robot must request more information from the user. The combination of the previous input and the user's response to the robot's question should form an unambiguous interpretation (See example 1 and 2 in Table 2) [Harte, 2008].

In CIMMI, the conversational context contributes to the correlation value of the command hypothesis, where $wtConv_s$ is the weighting of the speech recognition hypothesis s with respect to the conversational context and is defined as:

$$wtConv_s = \sum_{c=1}^{C} conv_{sc}$$

where $conv_{sc}$ is a value between 0 and 1 if a word in s matches a previous word c. Action words are not considered in this weight, as instruction words are considered unique from one command to the next. By this method, older command hypotheses may not contribute to the correlation values for the new command hypotheses much, but their contents can still be used to substitute missing information in an ambiguous new command hypothesis if needed [Harte, 2008].

In human interactions, the situational context can influence how a dialogue is interpreted. Situational context can include speaker identification, the speaker and listener's locations, objects that may be discussed and their locations and any other item that could be relevant to the dialogue. The situational contextual knowledge implemented in this work consists of Wakamaru's and the user's last known locations and a list of known objects and their last known locations [Harte, 2008].

The known object database is a extendable list of objects, where an object has a colour, an object type and a location. The locations can be updated and new objects added as the robot, or the user, change the environment. Knowing the colour or location of an object allows the user to simplify their spoken requests so they can simply ask 'bring my the blue cup' without stating the implied 'from the table'. Additionally, this context can be used to resolve misrecognitions where colour or location information could be missing from a command hypothesis [Harte, 2008].

In CIMMI, the object knowledge is used to contribute to the correlation values of the command hypotheses, where $wtObj$ is the weighting with respect to the known object database and is defined as:

$$wtObj_s = \sum_{o=1}^{O} obj_{so}$$

Table 2: Table showing some Sample Command Hypotheses from Speech Input. These utterances were spoken in the order listed

Ex.	Spoken Utterance	n-Best List		Command Hypotheses
		Parsed Input	Prob.	
1	Fetch the yellow cup	Get Yellow Cup _ Get Yellow _ Get Yellow Cut _	0.42 0.28 0.3	Get Yellow Cup _
2	On the desk	_ _ _ Desk	1.0	Get Yellow Cup Desk
3	Take the cup to the kitchen	Take _ Cup Kitchen Take _ _ Kitchen _ _ Cut Kitchen	0.38 0.27 0.23	Take Yellow Cup Kitchen

where obj_{so} is incremented if a word of speech recognition hypothesis s and an object in the database o match by type, and additionally incremented if the colour and location of the object matches other words in s also.

The situational context also consists of Wakamaru's and the user's locations. In this system, this knowledge is only used by the spatial intelligence. By knowing the robot's location, the response process can be simplified if Wakamaru is already at the desired location. Knowing the user's location was used so the robot could return there as part of fetching an object or to provide information about an object. When an object is returned to the user, its new location is updated in the known object database to the user's current location [Harte, 2008].

3.3 Multimodal Integration

A multimodal system should be able to handle uncertainties in the modes of input and generate an interpretation, the confidence of which varies in relation to the input [Jaimes and Sebe, 2007]. CIMMI receives up to five speech and up to three gesture recognition hypotheses, each with their corresponding probabilities of correctness. These hypotheses can be semantically combined to generate a list of command hypotheses, where a command hypothesis consists of a speech recognition hypothesis, s, and gesture recognition hypothesis, g, with a correlation value, $C(hyp_m)$, which represents CIMMI's confidence in this command hypothesis, in the form $[s, g, C(hyp_m)]$. The speech and gesture recognition hypotheses are only unified if there is an ambiguity in the speech hypothesis [Harte, 2008].

For each command hypothesis (hyp_m) in the command hypotheses list, a correlation value ($C(hyp_m)$) is calculated by summing CIMMI's confidence in the command hypothesis (wt) with the probabilities of correctness, $P(s)$ and $P(g)$, for the speech and gesture recognition hypotheses, s and g, respectively. This calculation can be shown by:

$$C(hyp_m) = wt + P(s) + P(g) \quad (1)$$

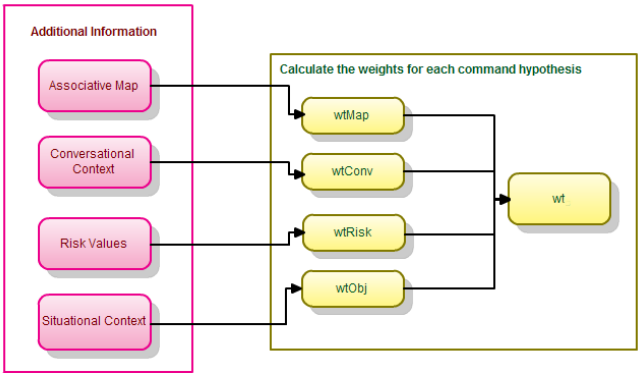


Figure 3: Combining contextual knowledge and speech information to calculate a command hypothesis weight.

where hyp_m is command hypothesis m of the list of M command hypotheses. $P(s)$ and $P(g)$ are the probabilities of correctness for the speech recognition hypothesis s and gesture recognition hypothesis g, where there are S speech hypotheses and G gesture hypotheses as received from the recognition systems. wt is the weight for the command hypothesis, hyp_m.

The weights wt from Eq. 1 are defined by summing the contribution of the associative map, $wtMap_s$, the risk values, $wtRisk_s$, the conversational context, $wtConv_s$ and the known object database in the situational context, $wtObj_s$ as illustrated in Fig 3 and shown in:

$$wt = wtMap_{sg} + wtRisk_s + wtConv_s + wtObj_s \quad (2)$$

The associative map, q, is an S x G matrix of values where S is the number of speech recognition hypotheses and G is the number of gesture recognition hypotheses. $wtMap_{sg}$ is the value of the sth x gth value in the associative map.

$wtRisk_s$ is the weighting of all the risk values, r, for the words in the speech recognition hypothesis s, where particular words, such as 'stop' or 'help', have a higher

risk of an adverse action occurring if there was a misrecognition.

$$wtRisk_s = \sum_{r=1}^{R} risk_{sr}$$

$wtConv_s$ is the weighting of the speech recognition hypothesis s with respect to the conversational context, where $conv_{sc}$ is a value between 0 and 1 if a word in s matches a previous word c. Action words are not considered in this weight, as instructions words are considered unique from utterance to utterance.

$$wtConv_s = \sum_{c=1}^{C} conv_{sc}$$

$wtObj_s$ is the weighting with respect to the known object database, where obj_{so} is incremented if a word of speech recognition hypothesis s and an object in the database o match by type, and additionally incremented if the colour and location of the object matches other words in s also.

$$wtObj_s = \sum_{o=1}^{O} obj_{so}$$

The list of command hypotheses are ranked according to the calculated correlation values, $C(hyp_m)$, and the best hypothesis selected. This hypothesis is then checked for ambiguities, possibly clarified and, if unambiguous, responded to by the robot.

A hypothesis is ambiguious either because of missing information (such as no object colour or location or because of a conflicting relationship between the speech/gesture pair according to the associative map. If there was a conflict between the speech and gesture, such as the word 'go' versus the gesture 'come', then the next hypothesis was selected from the list, as we assumed the conflict arose from our poor gesture recognizer [Harte and Jarvis, 2007]. However, if the ambiguity was an absence of information to complete a task, the robot would try to clarify the hypothesis using its knowledge and if unsuccessful, request a clarification from the user. If there was no ambiguity, then the robot will attempt to complete the task set it.

3.4 Wakamaru's Responses

Wakamaru was developed as a fetch-and-carry platform to demonstrate the successes, or failures, of our multimodal system, though it was unable to retrieve objects because of its unarticulated hands. The responses were defined using a developed finite state machine of all Wakamaru's functionalities. Wakamaru could answer questions that reference known information, such as 'where is the blue cup?', navigate to locations with obstacle avoidance, recognize objects and touch them. It would also return itself to its charger if that battery power was too low.

Localization was done onboard Wakamaru and a distance-transform based algorithm was developed allowing Wakamaru to navigate and avoid dynamic obstacles using the Hokuyo laser ranger finder data [Gupta and Jarvis, 2007]. Simple object recognition was also implemented using ShapeMatcher, a recognizer that creates a skeleton of a 2D binary shape and then stores indexed segments of it into a hierarchical, directed acyclic graph. This graph was stored in a database for later comparison [Macrini, 2003]. Wakamaru could respond by onboard speech synthesis (with Japanese phonetics) gestures, and task performance. Many of these motions were predefined by the Mitsubishi engineers, and are visually smooth.

4 Experiments

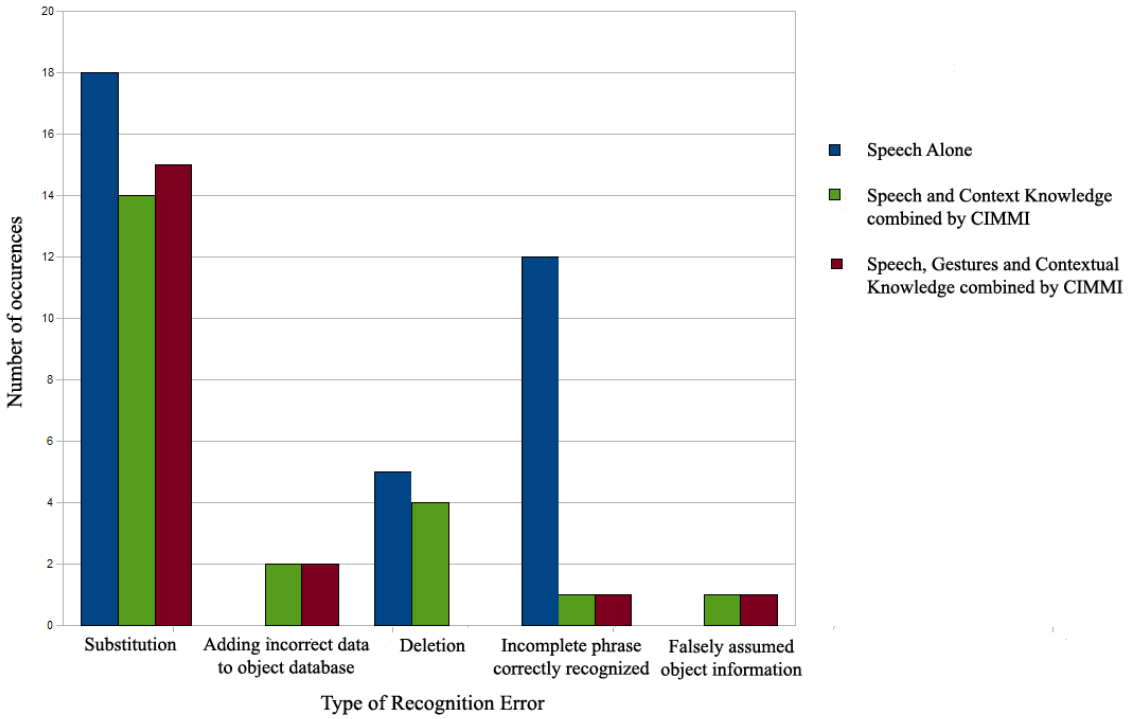
The experiments were designed to compare the strength of using additional information with the speech and gesture inputs over not using it. Speech and gesture interactions were captured for testing, where a user stood two or three meters from the robot wearing a microphone. The robot was placed in several positions in the room and predefined conversations were performed, without interaction on its part. Eight unique conversations were performed to the robot twice through giving 151 user commands, 79 of which were multimodal.

4.1 Experiment 1: Accuracy of the Individual Modes

While the speech recognition accuracy was previously reported as 79% in [Harte, 2008], this measure of accuracy was calculated as the percentage of parsed speech recognitions where the correct recognition hypothesis was in the list of recognition hypotheses. The gesture recognition accuracy was calculated as the percentage of inputs where the correct gesture class was listed in the list of recognition hypotheses. The gesture recognition accuracy for this dataset was 64% [Harte, 2008]. The correct recognition hypothesis was selected by comparing it with the spoken utterance or the performed gesture.

However, just because a spoken command was correctly recognized does not mean it could be responded to by a robot. Because of this, a second measure of accuracy was needed. This measure of accuracy was defined as the percentage of parsed recognition hypotheses for which the hypothesis could be correctly responded to by the robot. That is whether there is enough information for the robot to correctly perform the intended task without asking for clarification from the user [Harte, 2008]. This measure became the primary one used for

Graph 1: Comparing the Different Misrecognition Errors for the Different Recognition Results



comparing the accuracies of the different unimodal and multimodal systems implemented.

Under this measure and using the 79 multimodal commands, the speech recognition accuracy was 53% if no contextual knowledge or gestures were used. Without the contextual knowledge there was no way for the speech recognition to recover from its errors [Harte, 2008].

Of these errors, 51% (18 cases) were due to substitution errors, where a word such as 'cut' was recognized instead of 'cup', and 11% (5 cases) were deletion errors where a spoken word was misrecognized and the alternative was not in the vocabulary file leaving a blank in the parsed input. The remaining 38% of errors (12 cases) were due to incomplete spoken requests where the requests were correctly recognized by were intentionally missing information the robot would need to accomplish a task. Without contextual knowledge, the missing information cannot be resolved, such as 'Bring the yellow cup' [Harte, 2008]. These error values are illustrated in Graph 1.

The gestures could not be responded to by the robot alone, so no gesture recognition accuracy under the second measure was calculated.

Experiment 2: CIMMI with Speech and Contextual Knowledge

Having calculated the accuracy of the speech recognition system for the 79 multimodal commands, this needs to be compared to using the contextual knowledge as well, and using the gesture recognition system. The accuracy of CIMMI using the speech recognition hypotheses and the contextual knowledge was 72% - 19% higher than without using contextual knowledge [Harte, 2008]. Two percent of this improvement was because new objects were added to the known object database (See Ex 1 of Table 3). Adding objects, though, came with its own errors. In two cases, misrecognized object information was added to the object database (See Ex 2 of Table 3).

Of the errors, 67% (16 cases) were substitution errors - two cases less than without the contextual knowledge. These cases were resolved by selecting a more correct hypothesis from the list of recognition hypotheses, such as in Ex 3 of Table 3. An error occured where the colour of an object was falsely assumed because the first 'Cup' object in the known object database was the blue one, but the user was in fact interested in the green one (See Ex 4 in Table 3).

Of all the errors, four cases (17%) were deletion errors - the same number as without the contextual knowledge. One deletion was created by resolving a substitution error, but a deletion error was resolved to balance this. The error was resolved because a 'call' action was assumed since the 'doctor' word was recognized. This rule also applied to other personas implemented in the system such as police, ambulance and neighbour because the only action a robot would perform with regard to a

Table 3: Table showing some Sample Speech Hypotheses Selected by CIMMI

Ex.	Spoken Utterance	n-Best List		Selected hypothesis
		Parsed Input	Prob.	
1	Bring the yellow cup	Get Yellow Cut	0.42	Get Yellow Cup Desk
		Get Yellow Cup	0.29	
		Get Yellow _	0.29	
2	Bring the green cup from the kitchen	Get _ Cut Kitchen	0.35	Get Green Cut Kitchen
		Get Green Cut Kitchen	0.35	
		Get _ _ Kitchen	0.3	
3	Bring my cup	Get _ Cut _	0.52	Get Blue Cup Table
		Get _ Cup _	0.48	
4	Bring my cup	Get _ Cut	0.42	Get Blue Cup Table
		Get _ Cup	0.29	
		Get _ _	0.29	

persona was to call them.

An additional 12.5% (3 cases) of cases were errors due to incomplete spoken requests by the user - nine less cases than by the speech system alone. These nine cases were resolved by referencing the known object database for the missing information needed to complete a command hypothesis. However, the three unresolved cases could not be resolved by this knowledge because the objects mentioned were not in the object database. The robot could only request the missing information from the user to create a command hypothesis that could be responded to.

Experiment 3: CIMMI with Speech, Gestures and Contextual Knowledge

The multimodal recognition accuracy of CIMMI with the contextual knowledge was 76%, 4% higher than using the speech only with the contextual knowledge.

CIMMI resolved the four deletion errors for the dataset. These cases, where an action word was missing in at least the first speech recognition hypothesis, if not all of them, could only be resolved by another mode of input, such as gestures. The missing action words were substituted by the action of the symbolic gestures based on the recognized words in the command hypothesis.

In two cases, an action word was correctly assumed based on the selected gesture recognition hypothesis, such as a 'Go' gesture recognized and the speech recognition hypothesis ' _ Door' was resolved to 'Answer Door'. For a third cases, a 'Go' gesture strengthened the 'Go Charger' speech recognition hypothesis over the incomplete alternatives. These three cases represent the ideal fusion of the speech and gesture recognition hypotheses in this multimodal system, and were the reason for the 4% improvement in the multimodal recognition accuracy.

The fourth deletion error became a substitution error as the incorrect action word, 'Get', was assumed instead of the spoken action word, 'Where'. This occurred because of a gesture was recognized where none was performed. With the current system, this error could not be resolved. However, extending the gesture vocabulary so there was a gesture for 'Where' could resolve this error in the future.

The only 3 other cases in error of this system were:

- Assuming the wrong action for a command from the misrecognized gesture, as outlined above.
- Assuming the wrong colour for an object from the known object database information as described in Experiment 2 of this section.
- A low probability for a selected speech recognition hypothesis

Resolving these errors would create a system that could respond to all the commands recognized in this dataset without requesting clarification from the user.

4.2 Robotic Demonstration

The key difference between human-computer interaction and human-robot interaction is that the robot interacts in the unstructured world with the human participant. To demonstrate our multimodal integration, the system was implemented on a humanoid robot, Wakamaru, using localization, navigation, obstacle avoidance and object recognition, in addition to the speech and gesture recognition. Wakamaru localized itself using reflective beacons at known locations on the ceiling, and navigation and obstacle avoidance were achieved using a distance transform based algorithm and requesting depth data from the laser range finder [Harte and Jarvis, 2007].

Twenty-seven videos were captured of a user interacting with the robot in the new laboratory environment. The user would make requests verbally and gesture where it would reinforce the spoken utterance. in

Figure 4: Wakamaru touching a recognized object.

these demonstrations, a dataset of 50 spoken utterances were recorded, 43 of which were performed multimodally.

Overall, only 56% of the interactions were recognized without clarification - a much poorer result than the 76 for the offline captured dataset testing the algorithm. Of the errors, 5 cases (10% of cases) were easily avoidable because the user had spoken before the software had finished its initialization. To avoid this, Wakamaru could advise the user when its ready to receive instructions. Ignoring these errors gives a more accurate representation of the task performance of the robotic system, which was then 66%.

In another 4% of cases, the name 'Waka' was misrecognized resulting in no response from Wakamaru. In the demonstrations it was apparent if the user had spoken and no response had been made - a frustrating scenario. Despite these errors for the first spoken commands, with clarifications from the user 96% of the commands were recognized correctly by the robot. The robot would ask the user clarification questions, such as 'What did you want me to find?' and 'What did you want me to get?'. Using the user's answers to these questions, complete command hypotheses were formed, resolving the misrecognitions.

For the fetch-and-carry requests 'bring me the green cup', 56% of the cases were recognized without clarification. The main reason for this accuracy was the poor recognition of the word 'cup'. 'Cut', the common misrecognition, was not in the vocabulary file so the parsed speech recognition would have a blank in the object slot and Wakamaru would ask 'What did you want me to get?' With clarifications, the correct request was resolved to 'Bring Green Cup Kitchen' in 89% of cases, resolving the location of the green cup from the known object database.

5 Conclusions and Future Works

We have described an implementation of CIMMI, a contextually informed multimodal system fusing speech and symbolic gestures probabilistically. Symbolic gestures were used because they could substitute or reinforce action words in the speech recognition. As a complementary source of information, the implemented contextual knowledge was used to substitute or reinforce object or location words in the speech recognition. Experiments were performed to test the effectiveness of using symbolic gestures and the contextual information to support the interpretation of the primary speech mode.

Using CIMMI, the speech recognition accuracy was 53%, where the accuracy was defined as the percentage of commands which the robot could respond to. This was increased to 72% with the use of contextual knowledge, which consisted of the conversational history and situational information. The contextual knowledge helped select the best command hypothesis from the generated list or resolved missing information in the selected command hypotheses.

Only three symbolic gestures were recognized by the gesture recognition system, so the accuracy of CIMMI only improved to 76% with the addition of the symbolic gestures. This increase is only small because only four test cases required the gestures to resolve an error, whereas the contextual knowledge could resolve 12 test cases. This reflects the proportion of test cases that were missing an action word compared to test cases where there was missing object or location information.

Our system can be extended by including an implementation of user models and extending the gesture recognition vocabulary. Because some of the gestures have more than one meaning, parsing both the speech and gesture recognitions together could also be beneficial to the ranks of the parsed n-best lists, perhaps increasing the accuracy of the multimodal integration further. We also plan to improve the various systems used by the robotic system to increase its task performance accuracy.

Acknowledgments

The authors would like to thank all the members of the Intelligent Robotics Research Centre at Monash University for their support and Mitsubishi Heavy Industries for the loan of the Wakamaru robots for this project. The authors also acknowledge the reviewers' comments.

References

[Ammicht *et al.*, 2007] Egbert Ammicht, Eric Fosler-Lussier, and Alexandros Potamianos. Information seeking spoken dialogue systemspart i: Semantics and pragmatics. *IEEE TRANSACTIONS ON MULTIMEDIA*, 9(3):532–549, April 2007.

[Bischoff and Graefe, 2004] R. Bischoff and V. Graefe. Hermes - a versatile personal robotic assistant. In

Proceedings of the IEEE, volume 92, pages 1759– 1779, 2004.

[Cohen *et al.*, 1997] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. Quickset: Multimodal interaction for distributed applications. In *Proceedings of the Fifth Annual International Multimodal Conference*, New York, US, 1997. ACM Press.

[Eisenstein and Christoudias., 2004] Jacob Eisenstein and C. Mario Christoudias. A salience-based approach to gesture-speech alignment. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages .25–32, East Stroudsburg, Pennsylvania, 2004.

[Gorniak and Roy, 2005] Peter Gorniak and Deb Roy. Probabilistic grounding of situated speech using plan recognition and reference resolution. In *Seventh International Conference on Multimodal Interfaces (ICMI'05)*, Trento, Italy, October 2005.

[Gullberg, 1999] M Gullberg. Gestures in spatial descriptions. In *Working Papers 47*, page 8797. Lund University, Department of Linguistics, 1999.

[Gupta and Jarvis, 2007] O. K. Gupta and R. A. Jarvis. Multi-sensory fusion and understanding of human-robot interaction in assistive robotic technology environments: Navigation and hand-eye coordination. Transfer Report, ECSE, IRRC, Monash University, Australia, 2007.

[Harte and Jarvis, 2007] E. Harte and R.A. Jarvis. Multimodal human-robot interaction in an assistive technology context. In M. Dunbabin and M. Srinivasan, editors, *Proceedings of the 2007 Australasian Conference on Robotics & Automation*, 2007.

[Harte, 2008] E. Harte. Cimmi - a contextually informed multimodal integration method for a humanoid robot in an assistive technology context. Master's thesis, Intelligent Robotics Research Centre, Department of Electrical Computer Systems Engineering, Monash University, Clayton, VIC 3800, Australia, July 2008 2008.

[Hüwel and Wrede, 2006] S. Hüwel and B. Wrede. Robust speech understanding for multi-modal human-robot communication. In *Proceedings of the The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06)*, Hatfield, UK, Sept 2006.

[Iba *et al.*, 2002] Soshi Iba, Chris Paredis, and Pradeep Khosla. Interactive multimodal robot programming. In *International Conference on Robotics and Automation (ICRA) 2002*, pages 161–168, 2002.

[Jaimes and Sebe, 2007] Alejandro Jaimes and Nicu Sebe. Multimodal human-computer interaction: A survey. *Comput. Vis. Image Underst.*, 108(1-2):116–134, 2007.

[Kettebekov and Sharma, 2001] Sanshzar Kettebekov and Rajeev Sharma. Toward natural gesture/speech control of a large display. In *Engineering for Human Computer Interaction*, pages 221–234, 2001.

[Koons and Sparrell, 1994] David B. Koons and Carlton J. Sparrell. Iconic: speech and depictive gestures at the human-machine interface. In *CHI '94: Conference companion on Human factors in computing systems*, pages 453–454, New York, NY, USA, 1994. ACM.

[Macrini, 2003] D. Macrini. Indexing and matching for view-based 3-d object recognition using shock graphs. Master's thesis, Graduate Department of Computer Science, University of Toronto, 2003.

[Oviatt, 1999] S. Oviatt. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the Conference on Human Factors in Computing Systems: CHI'99*, pages 576–583, Pittsburgh, PA, 1999. ACM Press.

[Oviatt, 2003] S. Oviatt. User-centered modeling and evaluation of multimodal interfaces. In *Proceedings of the IEEE*, volume 91, pages 1457– 1468, Sept 2003.

[Pollack, 2005] M. Pollack. Intelligent technology for an aging population. *AI Magazine*, Summer:9–24, 2005.

[Potamianos *et al.*, 2007] A. Potamianos, E. Fosler-Lussier, E. Ammicht, and M. Perakakis. Information seeking spoken dialogue systems part ii: Multimodal dialogue. *IEEE Transactions on Multimedia*, 9(3):550 – 566, April 2007.

[Rogalla *et al.*, 2002] O. Rogalla, M. Ehrenmann, R. Zollner, R. Becher, and R. Dillmann. Using gesture and speech control for commanding a robot assistant. In *Proceedings of the 11th IEEE Int. Workshop on Robot and Human Interactive Communication*, pages 454–459, 2002.

[Russ *et al.*, 2005] G. Russ, B. Sallans, and H Hareter. Semantic based information fusion in a multimodal interface. In *International Conference on human-computer interaction, HCI'05*, pages 94–100, 2005.

[Stiefelhagen *et al.*, 2007] R. Stiefelhagen, H.K. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A Waibel. Enabling multimodal humanrobot interaction for the karlsruhe humanoid robot. *Robotics, IEEE Transactions on*, 23(5):840 – 851, Oct 2007.

[Wu *et al.*, 1999] L. Wu, S.L. Oviatt, and P.R. Cohen. Statistical multimodal integration for intelligent hci.

In *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, Aug 1999.