
Multimodal Interaction with Mobile Devices: Fusing a Broad Spectrum of Modality Combinations

Dissertation
zur Erlangung des Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)
der Naturwissenschaftlich-Technischen Fakultät I der Universität des Saarlandes

vorgelegt von

Rainer Wasinger

Saarbrücken
20. November 2006

Datum des Kolloquiums:

Montag den 20. November 2006

Dekan und Vorsitzender:

Prof. Dr. Thorsten Herfet

Gutachter:

1. Prof. Dr. Dr. h.c. mult. Wolfgang Wahlster

2. Prof. Dr. Elisabeth André

Akademischer Beisitzer:

Dr. Dominik Heckmann

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Diese Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in anderen Prüfungsverfahren vorgelegt.

Saarbrücken, den 20. November 2006

Acknowledgements

Most of all, I would like to thank Professor Wahlster for providing the funds, the time, and the environment, for any and all of this to have been possible. I also thank you for permitting me to work on such an interesting topic at the DFKI. I value your advice, which is always relevant, and your ability to see the interesting aspect of practically any research field. I also thank Elisabeth André for being the second corrector of this work, and the German Federal Ministry for Education and Research (BMBF), which has funded this work under the contract number 01 IN C02 as part of the projects COLLATE and COLLATE II.

I would like to dearly thank my wife Emma and my son Xavier for making everything worthwhile. The experiences that we have shared together in Europe are unforgettable, and I thank you very much for sticking it out with me till the end.

I would particularly like to thank the stunning long-distance support and endless amounts of reading (and rereading) that all of my family has invested into this work. I am greatly appreciative for the valued support and motivational drive that my father, Walter Wasinger, always provides, and for his ability to help me find my way each time I travel into untried waters. I thank my mother for being able to share complement views and for reminding me that some things just aren't the be all and end all. Appreciation is furthermore extended to Valerie, Glenn, Karina, and Tristan, to Chris and Debbie for sharing Emma with me, and to Heinz, Margit, Magdalena, and Fritz.

Special thanks go to all my colleagues at the AI chair. I thank Antonio Krüger for his good mix of work ethics and care free attitudes, and his ability to relentlessly pour interesting discussion into lunchtime. I thank Dominik Heckmann for his humour and for his positive outlook on life, and for also being the Akademischer Beisitzer for this work. I thank our secretary, Doris Borchers, for taking such good care of my family and myself over the course of this dissertation, and for creating and maintaining the family-like structure of this university chair. I thank Boris Brandherm and Ilhan Aslan for contributing to the social requirements of this dissertation, and in particular, for their great company at all the good pubs in Saarbrücken.

Additional thanks go to Jörg Baus, Michael Feld, Ralf Jung, Michael Kruppa, Michael Schmitz, Michael Schneider, Tim Schwartz, Ljubomira Spassova, Christoph Stahl, Anthony Jameson, and Kerstin Klöckner. Thanks are also extended to several former colleagues that I had the pleasure to work with at the chair, namely Thorsten Bohnenberger, Christian Kray, Christian Müller, and Frank Wittig. Also to Christian Kersten, Andreas Maier, Christian Schmitz, and Gerrit Kahl, who all contributed to the MSA/BPN system during their HiWi years, thank you.

Rainer Wasinger, September 2006

Diese Dissertation präsentiert eine multimodale Architektur zum Gebrauch in mobilen Umständen wie z. B. Einkaufen und Navigation. Außerdem wird ein großes Gebiet von möglichen modalen Eingabekombinationen zu diesen Umständen analysiert. Um das in praktischer Weise zu demonstrieren, wurden zwei teilweise gekoppelte Vorführungsprogramme zum ‘stand-alone’ Gebrauch auf mobilen Geräten entworfen. Von spezieller Wichtigkeit war der Entwurf und die Ausführung eines Modalitäts-fusion Modul, das die Kombination einer Reihe von Kommunikationsarten wie Sprache, Handschrift und Gesten ermöglicht. Die Ausführung erlaubt die Veränderung von Zuverlässigkeitswerten innerhalb einzelner Modalitäten und außerdem ermöglicht eine Methode um die semantisch überlappten Eingaben auszuwerten. Wirklichkeitsnaher Dialog mit aktuellen Objekten und symmetrische Multimodalität sind zwei weitere Themen die in dieser Arbeit behandelt werden. Die Arbeit schließt mit Resultaten von zwei Feldstudien, die weitere Einsicht erlauben über die bevorzugte Art verschiedener Modalitätskombinationen, sowie auch über die Akzeptanz von anthropomorphisierten Objekten.

This dissertation presents a multimodal architecture for use in mobile scenarios such as shopping and navigation. It also analyses a wide range of feasible modality input combinations for these contexts. For this purpose, two interlinked demonstrators were designed for stand-alone use on mobile devices. Of particular importance was the design and implementation of a modality fusion module capable of combining input from a range of communication modes like speech, handwriting, and gesture. The implementation is able to account for confidence value biases arising within and between modalities and also provides a method for resolving semantically overlapped input. Tangible interaction with real-world objects and symmetric multimodality are two further themes addressed in this work. The work concludes with the results from two usability field studies that provide insight on user preference and modality intuition for different modality combinations, as well as user acceptance for anthropomorphized objects.

Diese Dissertation untersucht die Anwendung von multimodalem Dialog zwischen Mensch und Computer im alltäglichen Kontext und insbesondere in solchen Umständen, bei denen der Anwender mobil ist. Zwei vollkommen ausgeführte und verbundene Anwendungsbeispiele mit Namen BMW Personal Navigator (BPN) und Mobile ShopAssist (MSA) wurden im Rahmen dieser Arbeit entwickelt um eine solide Basis zu dieser Arbeit zu schaffen. Die speziellen mobilen Umstände für diese Arbeit beziehen sich auf Fußgängernavigation und Einkaufen. Die ausgeführten Verständigungsarten umfassen Sprache, Handschrift und Gesten. Beide Anwendungsbeispiele wurden 'stand-alone' und für mobile Geräte, nämlich 'Personal Digital Assistants' entwickelt. Um das eingebettete Programm zu erweitern, erlaubt es außerdem eine 'Always Best Connected' Methodologie, wie Dialogkomponenten, die sich in einer öffentlich instrumentierten Umgebung befinden. Z.B. verteilte Erkenner für Sprache und Gesten können zum Zweck der erweiterten Funktion wie verbesserte Erkennungsgenauigkeit und zur Unterstützung eines größeren Wortschatzes benutzt werden.

Die Dissertation trägt in einer Anzahl von Richtungen zur Grundlagenforschung im Feld der intelligenten Anwenderschnittstellen des mobilen multimodalen Dialoges zwischen Computer und Anwender bei. Ein System zur Einstufung von multimodalen Eingaben wurde entwickelt, um die Synchronie von Eingaben auf Zeit, Semantik und Herkunft zu kategorisieren. Eine solche Klassifikation ist besonders angebracht für gleichzeitige Eingabesignale die einen gemeinsamen temporalen und/oder semantischen Platz teilen, in welchem Fall die Signale überlappt genannt werden. Überlappte Eingaben formen eine wichtige Grundlage für jegliche Arbeit, die sich mit multimodalem Dialog und Modalitäts-fusion befasst.

Eine primäre Aufgabe ist, zu erstellen wie semantisch überlappte Konflikte in multimodalen Eingaben gelöst werden können. Es ist zu erwarten, dass semantisch überlappte Eingaben häufig vorkommen können in gewissen Anwendungsgebieten wie z. B. in den instrumentierten Umgebungen, wo mehrere gleichartige und verschiedenartige Erkenner leicht angewandt werden können um Anwendereingaben gleichzeitig zu erkennen. Die angewandte Fusionstechnik zur Erkennung von semantisch überlappten Eingaben haben den Gebrauch von 'certainty factors' und Zuverlässigkeitswerten zur Grundlage. Letztere sind in der 'N-best' Resultatsliste des jeweiligen Erkenners gespeichert. Zeitgebundene Aspekte sind auch im Lauf des Modalitätsfusionsprozesses mit einbezogen sowie auch der Gebrauch von Salienz um die betroffenen Referenten zu identifizieren.

Weiters ist zu untersuchen, wie die Zuverlässigkeitswerte einer breiten Serie von Kommunikationsarten wie Sprache, Handschrift und Geste umgewichtet werden können, sodass die Werte 'unbiased' zwischen Modalitäten sind. Das wird mit Hilfe einer statistischen Datenbank der Er-

kennergenauigkeit demonstriert, die während einer Feldstudie an hand von gesammelten Daten erstellt wurde. Methoden, die Erfahrungsdaten des Anwenders zur Erkennungsgenauigkeit zu erhalten, ist ein anderer Gesichtspunkt, den diese Arbeit kurz behandelt. Dieses bahnt den Weg, um maschinelle Lernverfahren der KI einzuschließen um Zuverlässigkeitswerte von gesammelten Daten in Laufzeit auszuwerten.

Greifbarer Dialog ist eine Kommunikationsform, die sich zur mobilen Anwendung gut eignet, weil es dem Anwender erlaubt, direkt mit computertechnisch vernetzten Objekten in einer realen physischen Welt zu kommunizieren. Ein erwähnenswertes Merkmal dieser Arbeit ist der Einbezug des greifbaren Dialoges mit der wirklichen Welt. Mehrere Arten von Kommunikation in der wirklichen Welt einschließlich aufheben, zurücklegen und Zeigegesten, werden ausgewertet.

Es ergeben sich eine umfangreiche Zahl von benutzbaren Modalitätskombinationen zur Eingabe. Insgesamt 23 Kombinationen wurden in dieser Arbeit behandelt. Von diesen sind 12 Kombinationen nicht überlappt während die restlichen 11 Kombinationen semantisch überlappt sind. Diese Kombinationen erlauben dem Anwender den Dialog zu harmonisieren, so dass individuelle semantische Eigenschaften innerhalb des Dialoges für verschiedene Modalitäten spezifiziert werden können. Alle Kommunikationsarten, die in dieser Arbeit benutzt wurden, d.h. Sprache, Handschrift und Geste, wurden so ausgelegt, dass sie gleichwertig sind. Das erlaubt mobilen Anwendern die Eingaben den wechselnden Umgebungsbedingungen anzupassen, wie z.B. die gewünschte Privatsphäre bzw. Hintergrund Geräusche.

Das Konzept der symmetrischen Multimodalität hilft zur Erweiterung des multimodalen Dialoges. Symmetrische Multimodalität bezieht sich auf die Fähigkeit eines Systems und seinen Anwendern mit derselben Modalität zu kommunizieren, d. h. dass das System zusätzlich zur multimodalen Eingabe in allen Kommunikationsarten einschließlich mit Geste antworten kann. Dieses Konzept ist für diese Arbeit wichtig, weil es zeigt, dass die Anwendereingabe nur eine Seite einer zweiseitigen Münze ist, während die andere Seite die Antwort des Systems repräsentiert.

Ein Höhepunkt dieser Arbeit sind die Ergebnisse von zwei Feldstudien, die in öffentlichen und privaten Umgebungen durchgeführt wurden. Die Ergebnisse zeigen die generelle Anforderung der Anwender für eine Reihe von Kombinationen zur Modalitätseingabe und berichten auch über Anwenderakzeptanz für anthropomorphisierte Gegenstände. Anthropomorphisation ist ein Konzept das in der Vergangenheit nicht ernst genommen wurde, ohne dass dafür ausreichende Begründung existierte. Ein Vorteil, der von anthropomorphischen Verbindungen erwartet werden darf, ist die Möglichkeit den Dialog auf einen spezifischen Anwender zu personalisieren.

Zusätzlich zu der theoretischen Forschung dieser Arbeit wurden auch praktische, kommerzielle und finanzielle Beiträge geleistet. Ein solcher Beitrag ist im Projekt COMPASS enthalten wo multimodale Komponenten und Ergebnisse der Benutzerstudien der MSA/BPN eingebunden sind in einen mehrsprachigen und multimodalen Fremdenführer. Der Fremdenführer wird für die Besucher zu den olympischen Spielen 2008 in Beijing entwickelt. Die MSA/BPN wurde auch die ausgewählte Plattform für die Untersuchungen am Konzept genannt 'Personal Journals' im Rahmen des SPECTER Projektes. Außerdem wird sie für Forschungszwecke auf dem Gebiet der instrumentierten Umgebungen und produkt-bezogenen Anzeigen unter dem Projekt FLUIDUM benutzt. Letztlich hat die Arbeit auch zur Förderung des öffentlichen Bewusstseins vom Stand der Technik in Sache der Sprachtechnologie beigetragen. Insbesondere, das MSA/BPN System wurde in mehreren Ausstellungszentren innerhalb Deutschlands demonstriert. Mobiler multimodaler Dialog im Zusammenhang mit dieser Arbeit wurde ebenso aufgenommen in Zeitungen und Fernsehen. Im Rahmen dieser Forschung wurden bis heute bereits 16 Artikel in führenden internationalen Journals, Konferenzen und Workshops veröffentlicht.

This dissertation investigates the use of multimodal human-computer interaction in everyday contexts, and particularly those contexts in which a user is mobile. Forming a solid basis for this research are two entirely implemented and interlinked demonstrators called the BMW Personal Navigator (BPN) and the Mobile ShopAssist (MSA), both of which were developed over the course of this work. The mobile contexts that these applications focus on are that of pedestrian navigation and shopping, and the communication modes that are catered for include speech, handwriting, and gesture. Both of these demonstrators have been designed for stand-alone use on mobile devices, namely Personal Digital Assistants. Extending on this embedded design, the applications additionally support an Always Best Connected methodology such that interaction components located in a publicly instrumented environment, for example distributed speech recognizers and gesture recognizers, can be made accessible to a user for the purpose of enhanced application functionality like improved recognition accuracy and support for larger vocabularies.

The dissertation contributes in a number of ways to leading-edge research in the field of intelligent user interface design and mobile multimodal interaction. A formal classification of multimodal input is established to categorize the synchrony of inputs based on time, semantics, and origin. Such a classification is particularly relevant for input streams that share a common temporal and/or semantic space, in which case the input is said to be overlapped. Overlapped input forms an important foundation for any work dealing with multimodal interaction and modality fusion.

A primary objective is to outline how semantically overlapped conflicts in multimodal input can be resolved. It is anticipated that semantically overlapped input will occur frequently in certain application domains like instrumented environments, where multiple same-type and multiple different-type recognizers can easily be deployed to simultaneously recognize user input. The fusion techniques used for resolving semantically overlapped input in this work are based on the use of certainty factors and the use of confidence values stored in each recognizer's N-best list of results. Timing aspects are also taken into account during the process of modality fusion, as too is the use of salience to identify relevant referents.

Another objective is to examine how confidence values returned by a broad range of communication modes like speech, handwriting, and gesture, can be re-weighted such that the values are unbiased within and between modalities. This is demonstrated by means of a statistical database on recognizer accuracy that has been created based on data collected during a field study. A requirement of this field study was that it accurately reflect mobile user interaction in a public environment setting. Methods for capturing user feedback on recognizer accuracy is another aspect covered briefly in this work, and this paves the way for incorporating AI machine learning techniques that can be used to re-weight confidence values based on data collected at runtime.

Tangible interaction is one form of communication that is proving to be well-suited to mobile applications because it permits users to interact directly with computationally augmented artifacts in the real physical world. A notable feature of this work is the incorporation of tangible interaction. Several types of real-world interaction are harnessed, including pickup, putdown, and point gestures. These interaction types, called off-device interaction, can be seen to augment their on-device interaction counterparts like handwriting, and also give rise to multimodal combinations comprised of both on-device and off-device interaction.

Off-device interactions are classified as extra-gestures and form part of the rich set of modality input combinations available. A total of 23 combinations are studied in this work, of which 12 combinations are non-overlapped and the remaining 11 are semantically overlapped. These combinations allow a user to fine-tune interactions so that even individual semantic constituents within an interaction can be specified using different modalities. All of the communication modes utilized in this work, i.e. speech, handwriting, and gesture, have been designed to be equally expressive. As a result, the modes cater for true supplementary input rather than just a limited set of complementary modality combinations, and this has the effect that mobile users can tailor their interaction according to the changing requirements of a particular environment such as the level of required privacy and background noise.

Extending on the use of multimodal interaction is the concept of symmetric multimodality. Symmetric multimodality refers to the ability for a system and its users to communicate via the same set of modalities, meaning that in addition to multimodal input, a system can provide output using all communication modes including gesture. This concept is important for this work because it effectively illustrates that user input is only one side of a two-sided coin, the other side representing system output.

A highlight of this work is the results of two usability field studies conducted under public and private environment contexts. The results outline user preference and modality intuition for a range of modality input combinations, and also report on user acceptance for anthropomorphized objects. Anthropomorphization is a concept that has been ridiculed in the past despite a lack of solid foundation existing for such ridicule. One benefit that can be awaited from anthropomorphic interfaces is the ability to personalize interaction to a specific user or user group such as children, adults, or the elderly.

In addition to the theoretical research of this work, there have also been practical and commercial contributions. One such contribution has been to the project COMPASS, where multimodal components and usability study results from the MSA/BPN have been integrated into a multilingual and multimodal tourist guide being developed for visitors to the 2008 Olympic Games in Beijing. The MSA/BPN has also become the platform of choice for research being conducted on a concept called 'personal journals' under the SPECTER project, and for research being conducted on instrumented environments and product associated displays under the FLUIDUM project. Finally, the work has also contributed to public awareness for state-of-the-art language technology. In particular, public system demonstrations have been held at a number of exhibition centres within Germany, and mobile multimodal interaction relating to this work has also made its way into the newspaper and television. At the time of writing, this work has also generated 16 peer-reviewed articles published in leading international journals, conferences, and workshops.

1	Introduction	1
1.1	Aims and Methods	1
1.2	Chapter Outline	3
2	Basic Concepts	5
2.1	The Human Senses and Communication	5
2.1.1	Basic Model for Human-Computer Interaction	5
2.1.2	Perception and the Human Senses	6
2.1.3	Verbal and Nonverbal Communication	8
2.1.4	Computer Input and Output Media	10
2.2	Multimodality and Modality Fusion	12
2.2.1	Benefits of Multimodal Interaction	12
2.2.2	Multimodality Defined	14
2.2.3	W3C Classification of Multimodal Input	15
2.2.4	Modality Fusion	16
2.2.4.1	Early and Late Fusion	17
2.2.4.2	Generalized Architecture of a Multimodal System	18
2.3	Reference Resolution	18
2.3.1	Referents, Referential Terms, and Referring Modes	19
2.3.2	Multimodal Discourse Phenomena	22
2.4	Mobile Users and Instrumented Environments	24
2.4.1	Progression from Stationary Computing to Mobile Computing	24
2.4.1.1	Mobile Device Limitations	25
2.4.1.2	Current Trends for Mobile Computing	28
2.4.2	MSA/BPN System Descriptions and Scenarios	30
2.4.2.1	BMW Personal Navigator Scenario	31
2.4.2.2	Mobile ShopAssist Scenario	37
2.4.3	MSA/BPN System Architecture	42
3	Related Work	45
3.1	Projects with a Focus on Multimodal Interaction	45
3.1.1	Put-That-There	46
3.1.2	XTRA	47
3.1.3	QuickSet	48

3.1.3.1	QuickSet-Rasa	50
3.1.3.2	QuickSet-3D Hand: 3D Hand Gesture Extension to QuickSet	51
3.1.3.3	QuickSet-ExertEnv: Mobile System for Exerted Conditions	51
3.1.4	MATCH	52
3.1.5	MUST	53
3.1.6	EMBASSI	54
3.1.7	SmartKom	55
3.1.7.1	SmartKom-Mobile	59
3.1.7.2	SmartWeb	60
3.1.8	COMIC	60
3.1.9	MIAMM	61
3.1.10	COMPASS	62
3.1.11	Project Short-form Comparisons	63
3.2	Mobile Users and Instrumented Environments	66
3.2.1	Shopping Assistance	67
3.2.1.1	Commercial Shopping Assistance Systems	68
3.2.1.2	Research Shopping Assistance Systems	69
3.2.2	Map-based Guides	71
4	Modal and Multimodal Interaction	75
4.1	Modal Interaction	75
4.1.1	MSA/BPN Communication Modes	76
4.1.1.1	Speech Recognition	78
4.1.1.2	Handwriting Recognition	82
4.1.1.3	Tangible Interaction and Gesture Recognition	84
4.1.2	Confidence Values and Confidence Scoring	89
4.1.3	Accuracy, Interaction Times, and Scalability	95
4.1.3.1	Modality Accuracy	97
4.1.3.2	Modality Interaction Times and Scalability	100
4.2	Multimodal Interaction	101
4.2.1	Temporal and Semantic Synchrony of Multimodal Interaction	101
4.2.1.1	Temporal Synchrony of Multimodal Interaction	102
4.2.1.2	Semantic Synchrony of Multimodal Interaction	104
4.2.2	Semantic Overlap	106
4.2.2.1	Linking Individual Modalities to Individual Semantic Constituents	107
4.2.2.2	Modality Input Combinations in the MSA	108
4.2.3	Multiple Recognizers and their Contribution to Semantically Overlapped Input	111
4.2.3.1	The For and Against on Semantically Overlapped Input	111
4.2.3.2	Passive Collection of Semantically Overlapped Input	113
4.2.3.3	Same-type Recognition as a Source of Semantically Overlapped Input	114
4.3	Direct and Indirect Interaction: Anthropomorphization	115
4.3.1	The Role of Anthropomorphization in the MSA	115
4.3.2	Adding Human-Like Characteristics	118
4.3.3	State-based Object Model	119

4.4	Symmetric Multimodality and Presentation Output Planning	120
4.4.1	Symmetric Multimodality in the MSA/BPN	121
4.4.2	Presentation Output Planning in the MSA/BPN	123
4.5	Multiple Users and Multiple Devices in the MSA Instrumented Environment . . .	126
4.5.1	Situated Interfaces	126
4.5.2	Multiple Users	127
4.5.3	Device Taxonomy	128
4.5.4	Device Control	130
5	Modality Fusion Strategies	133
5.1	Multimodal Input Modelling and Knowledge Representation	133
5.1.1	Background to Input Modelling and Knowledge Representation	133
5.1.1.1	Mobile Device Limitations to Modelling Input	133
5.1.1.2	Syntax, Semantics, and World Knowledge	134
5.1.1.3	Knowledge Sources and their Use in the MSA/BPN	134
5.1.1.4	Modelling of User Input at different Processing Stages	136
5.1.2	Multimodal Input and Knowledge Representation Standards	138
5.1.2.1	Ontology Modelling Protocols	138
5.1.2.2	Multimodal Communication Protocols	140
5.1.3	Multimodal Input and Knowledge Representation in the MSA/BPN . . .	145
5.1.3.1	Data and Method Attributes in the MSA/BPN	146
5.1.3.2	Communication Acts in the MSA/BPN	150
5.2	Timing, Timeframes, and Saliency in the MSA/BPN	152
5.2.1	Activating the Modality Fusion Component	152
5.2.2	Allocating an Appropriate Timeframe for Terminating a Current User-turn	153
5.2.3	Saliency in the MSA/BPN	156
5.3	Modality Fusion in the MSA/BPN	157
5.3.1	Previous Work on Defining Modality Fusion Strategies	157
5.3.2	Processing Multimodal Input in the MSA	159
5.3.2.1	Modality Fusion Architecture and Blackboard Design	160
5.3.2.2	Conflict Resolution in the MSA/BPN	162
5.3.3	Using Statistical Probabilities to Re-weight Confidence Values	162
5.3.3.1	User feedback on Recognition Accuracy	167
5.3.4	Conflict Resolution between Multiple Communication Acts	168
5.3.5	Multimodal Blackboard Event Filtering	169
5.3.6	Conflict Resolution between Multiple Semantic Elements	171
5.3.6.1	Uncertain Reasoning	172
5.3.6.2	Walkthrough of the Evaluation of semantically overlapped and Conflicting Input	174
6	Usability Studies on Multimodal Interaction	177
6.1	Previous Usability Studies on Multimodal Interaction	177
6.2	Modality Preference in Private (Laboratory) and Public (Real-world) Environments	182
6.2.1	Usability Study Descriptions	183
6.2.1.1	Study 1 - Private (Laboratory) Tests	183
6.2.1.2	Study 2 - Public (Real-world) Tests	184

6.2.2	Quantitative Analysis and Results	187
6.2.2.1	Effects of a Consolidated View	187
6.2.2.2	Preferred Modality Combinations Ranked by Feature Group	189
6.2.2.3	Preferred Modality Combinations Ranked by Preference	190
6.2.2.4	Modality Intuition	193
6.2.2.5	Modality Usage in Public and Private Environments	196
6.2.2.6	The Effects of Observability on Modality Combination Preference	198
6.2.2.7	General Results Regarding the MSA Demonstrator	199
6.2.3	Quantitative Comparisons Between the Studies	199
6.2.4	Anthropomorphized Objects	202
6.2.4.1	Direct and Indirect Interaction	203
6.2.4.2	User-Product Relationships	204
6.2.4.3	Direct Interaction with a Range of Products as a Buyer and as an Owner	204
6.2.5	Qualitative Observations from the Studies	205
6.2.5.1	Modalities and their Combinations	205
6.2.5.2	Characteristics Considered Important for Multimodal Interaction	207
6.2.6	Usability Study Conclusions	208
7	Conclusions	211
7.1	Scientific Contributions	211
7.2	Practical Contributions	214
7.3	Commercial Significance and Contributions to Public Awareness	214
7.4	Opportunities for Further Research	215
	Bibliography	219

List of Figures

1.1	Structural overview of the dissertation	3
2.1	Interaction loop consisting of four stages: human output, computer input, computer output, and human input	6
2.2	Classification of verbal and nonverbal communication	9
2.3	Code, media, and modalities	11
2.4	Two schematic architecture diagrams showing the incorporation of modality fusion and fission components	16
2.5	Architecture showing typical components required by multimodal systems	18
2.6	The two mobile demonstrators created under the scope of this dissertation: The BPN and the MSA	30
2.7	Environment contexts used during interaction with the BPN and the MSA	31
2.8	Different situations during a navigational task	33
2.9	Downloading a travel itinerary onto the PDA	33
2.10	BMW Connected Drive	34
2.11	BPN screen-shots demonstrating speech output and different 2D/3D visual perspectives ranging from birds-eye to egocentric	35
2.12	Interaction with the BPN system	35
2.13	Two typical indoor navigation images showing the user's path to the electronics shop	36
2.14	Using the MSA to select from a number of different product types situated on a shelf	38
2.15	Different product views in the MSA showing a 9x view, a 4x view, a 1x view, and a product comparison view	39
2.16	MSA functionality as defined in the interaction grammars for digital cameras	39
2.17	MSA interaction illustrating the use of speech, intra-gesture, and extra-gesture, during a product comparison query	40
2.18	MSA interaction illustrating handwriting+intra-gesture, and speech	41
2.19	MSA ties to the instrumented environment, showing a public display, the on-device shopping basket, and an RFID-instrumented shopping trolley	41
2.20	The MSA/BPN architecture showing the data flow between components	42
3.1	QUICKSET system architecture	48
3.2	Users collaborating with RASA	50
3.3	MUST mobile tourist guide and its distributed architecture	54
3.4	SMARTKOM architecture showing the multi-blackboard design	57

3.5	SMARTKOM-MOBILE: Hardware and software components used in the scenario and a picture of the interface used for navigation	59
3.6	COMPASS smart dining service	63
3.7	METRO's FUTURE STORE instrumented trolley and IBM's SHOPPING BUDDY	68
3.8	The SMART SHOPPING ASSISTANT with plan-recognition technology	70
3.9	CROSSTALK and an example of a mobile shopping assistant user interface designed on the basis of user studies	71
3.10	Example of the map interface used in REAL, showing also the components of the REAL system	72
4.1	MouseField, showing tangible interaction and the hardware consisting of two optical mice and an RFID reader	77
4.2	States and transitions of the BPN finite-state rule-grammars	80
4.3	Data sources in the MSA showing the data container representing a shelf of products and the accompanying digital camera product grammars	81
4.4	The use of speech-only input in the MSA for selecting an object, a feature, and both a feature and an object	82
4.5	The use of handwriting-only input in the MSA for selecting an object, a feature, and both a feature and an object	83
4.6	Selection gestures in the BPN application illustrating intra-point and intra-slide gestures, and an extra-point gesture	86
4.7	Intra-gestures in the MSA application showing intra-point object selection, intra-point feature selection, and an intra-slide gesture	86
4.8	Extra-gestures in the MSA application showing extra-pickup and extra-putdown gestures, and an extra-point gesture	88
4.9	An example of how handwriting input is recognized by the character recognizer, and then mapped to a valid grammar entry and given a confidence value	91
4.10	Confidence value generation for intra-point object resolution showing the intersect between AA and IR, and four example confidence values	93
4.11	The range of interaction combinations that were used as basis for the CeBIT 2006 study: 6 modality combinations x 12 features x 13 objects	97
4.12	Accuracy rates for the recorded confidence values generated by the recognizers, for speech, handwriting, and gesture	99
4.13	Temporal relationships of multimodal input	104
4.14	Semantic relationship of multimodal input	105
4.15	Supplementary and complementary modality combinations as applied to the MSA	109
4.16	MSA modality input combinations showing modality input combinations for two semantic constituents and three semantic constituents	110
4.17	IBM audio video speech recognition system	113
4.18	Architecture supporting same-type recognition with recognizers located both on the client PDA device and on a server	114
4.19	Different commercial instantiations of anthropomorphized objects, including the Gaultier perfume, M&M chocolates, and the Mr Proper cleaning product	117
4.20	Anthropomorphized object interaction during the MSA usability study conducted at Conrad Electronic in Saarbrücken, Germany	118
4.21	Direct and indirect interaction, shown as a modifier of multimodal interaction	119

4.22	Product states in the MSA that are used for object-initiated interaction	120
4.23	Symmetric multimodal input and output matching	122
4.24	Parameters that could affect the planning of output in mobile scenarios	122
4.25	The symmetric use of modalities in the MSA/BPN	123
4.26	Modifying the output communication modes in the MSA/BPN	124
4.27	Two examples demonstrating the configurability of the MSA/BPN's presentation output planner	125
4.28	Situated interfaces and categorization of different interface types	127
4.29	Characteristics of users	128
4.30	Public devices located in the instrumented environment, showing the PAD text and images, a virtual character, the spotlight, the shelf and physical products, a shopping trolley, plasma wall-display, and public speakers	129
4.31	Device taxonomy in the MSA instrumented environment	130
4.32	Assignment of device control	131
5.1	Understanding: Data, information, knowledge, and wisdom	135
5.2	Shelf synchronization in the MSA	136
5.3	Partial ontology of object types and an example use of the ontology during refer- ence resolution	140
5.4	Three communication acts typically used in the MSA for queries and commands .	152
5.5	Timeframe for a typical user interaction in the MSA/BPN	155
5.6	Processing multimodal input in the MSA/BPN	160
5.7	Modality fusion architecture	161
5.8	Modality fusion blackboard illustrating the main data and method attributes . . .	161
5.9	The tables show the statistical dataset and the re-weighted confidence values for each of the 7 different semantic-modality categories	164
5.10	Trend-lines for the semantic-modality categories	165
5.11	N-best list feedback for semantically non-overlapped input	167
5.12	Graph of the certainty factor's equation as used in the MSA/BPN	173
5.13	Conflict resolution between semantically overlapped object references	175
6.1	Usability demonstrators most similar to the MSA	181
6.2	MSA demonstrator installation situated at Conrad Electronic	185
6.3	Summary of the demographics for subjects that partook in the real-world usability study	186
6.4	The effects of user preference consolidation as shown in the laboratory study and the real-world study	188
6.5	Preferred modality combinations ranked by feature group as shown in the labora- tory study and the real-world study	190
6.6	Preferred modality combinations ranked by preference as shown in the laboratory study and the real-world study	191
6.7	Preferred male and female modality combinations	192
6.8	Intuitiveness of the 23 different modality combinations as rated during the written and the practical components of the laboratory study	194
6.9	Intuitiveness of the 23 different modality combinations as rated during the written and the practical components of the real-world study	194

6.10	The first 4 modality combinations selected by subjects in the laboratory study and the real-world study	195
6.11	Modality usage in public and private environments as shown in the laboratory study and the real-world study	197
6.12	Comparison of modality combinations for the laboratory and the real-world studies showing significant differences in user preference	200
6.13	Significant differences between modality combination intuition arising from the analysis of results across both studies	201
6.14	Differences between hypothesized and actual modality preference values in the real-world and laboratory studies	202
6.15	The effect that being the owner or a buyer can have on direct interaction with a range of different products	205

List of Tables

2.1	The five classical senses defined by Aristotle	7
2.2	The communication modes a user would require when sending and receiving information in different modalities	8
3.1	Multimodal project comparisons	65
3.2	Overview of map-based mobile guides	73
4.1	Selection gestures available in the MSA/BPN	85
4.2	An example of the sliding character match algorithm used for the input ‘Optinlzrein’	92
4.3	Minimum and maximum range of values for the intra-gesture N-best list of confidence values	93
4.4	Confidence value generation for intra-point feature resolution showing the N-best confidence values based on different given scrolling text speeds	94
4.5	The three modality combinations rated by users to be most preferred during laboratory and real-world usability studies	96
4.6	Confidence value and accuracy rate averages recorded during the study	98
4.7	Time statistics for the interactions recorded during the usability study	100
4.8	Temporal durations for modality input combinations in the MSA	103
4.9	Three classes of semantically overlapped input	106
4.10	The degree of semantic overlap based on referent expressiveness	106
4.11	The 23 modality input combinations analysed in the MSA	110
4.12	Different configurations illustrating how semantically overlapped input can arise in a user-friendly manner	115
4.13	Public devices in the MSA instrumented environment	129
5.1	Data and method attributes used by the MSA/BPN application	147
5.2	Summary of the communication acts	150
5.3	Examples of the use of the aforementioned communication acts	151
5.4	Additional time periods required for entering object information after a query+feature or command has been issued to the system	155
5.5	Trend-lines for the semantic-modality categories	165
5.6	Four examples of speech combined with different gesture inputs illustrating the affect that temporal placement has on semantically overlapped references	171

6.1	Male and female preferences for non-overlapped and overlapped modality combinations	193
6.2	Modality usage in public and private environments illustrating gender differences	197
6.3	Effects of observability on modality combination preferences during interaction in private and public environments	198
6.4	Summary of the differences in modality combination preference between the laboratory and the real-world studies	200
6.5	Modality combinations exhibiting a significant difference in preference between the laboratory and real-world studies	201
6.6	Male and female preferences for direct and indirect interaction	203

List of Acronyms

ABC	Always Best Connected
AI	Artificial Intelligence
AL	Always Listening
API	Application Programming Interface
AR	Augmented Reality
ARG	Attributed Relational Graph
BMBF	Bundesministerium für Bildung und Forschung (Federal Ministry for Education and Research)
BNF	Backus-Naur Form
BPN	BMW Personal Navigator
Cf	Confidence
CF	Certainty Factor
CFG	Context-Free Grammar
CHCC	Center for Human Computer Communication
DAML	DARPA Agent Markup Language
DARPA	Defense Advanced Research Projects Agency
DFKI	Deutsches Forschungszentrum für Künstliche Intelligenz (German Research Centre for Artificial Intelligence)
DOM	Document Object Model
DSR	Distributed Speech Recognizer
DTD	Data Type Definition
EMMA	Extended Multimodal Markup Annotation
EU	European Union
EVV	Embedded ViaVoice
FSA	Finite-State Automaton
FSG	Finite-State Grammar
GPRS	General Packet Radio Service
GPS	Global Positioning System
GSM	Global System for Mobile Communications
GUI	Graphical User Interface
HCI	Human-Computer Interaction
HSDPA	High-Speed Downlink Packet Access
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol

I/O	Input/Output
ID	Identification
IE	Intelligent Environments
IR	Infrared
IUI	Intelligent User Interfaces
JSGF	Java Speech Grammar Format
LAN	Local Area Network
LT	Language Technology
M3L	MultiModal Markup Language
MMIL	MultiModal Interface Language
MPI	Max Planck Institute
MSA	Mobile ShopAssist
MTI	Mensch-Technik Interaktion (Human-Computer Interaction)
NLSML	Natural Language Semantics Markup Language
NSR	Network Speech Recognizer
OCR	Optical Character Recognition
OIL	Ontology Inference Layer
OWL	Web Ontology Language
P2A	Push to Activate
P2T	Push to Talk
PAD	Public Associated Display
PDA	Personal Digital Assistant
POI	Point Of Interest
RDFS	Resource Description Framework Schema
RFID	Radio Frequency Identification
ROM	Read-Only Memory
SDK	Software Development Kit
SDRAM	Synchronous Dynamic Random Access Memory
SQL	Structured Query Language
SRCL	Speech Recognition Command Language
TCP/IP	Transmission Control Protocol/Internet Protocol
TFS	Typed Feature Structures
TUI	Tangible User Interface
UMTS	Universal Mobile Telecommunications System
URL	Uniform Resource Locator
USB	Universal Serial Bus
VNC	Virtual Network Computing
VRML	Virtual Reality Markup Language
W3C	World Wide Web Consortium
WCIS	What-Can-I-Say
WIMP	Windows Icon Menu Pointer
WLAN	Wireless Local Area Network
WoZ	Wizard of Oz
WWW	World Wide Web
XML	Extensible Markup Languages

This dissertation contributes to the fields of intelligent user interface design, mobile and pervasive computing, tangible interaction, and most importantly, the field of multimodal interaction.

1.1 Aims and Methods

Human-computer interaction is no longer limited to desktop computing, in which a user is typically sitting down at a table and looking at a stationary computer display. Modern-day applications based on mobile devices like Personal Digital Assistants (PDAs) and phones now afford their users an unprecedented degree of mobility, and they also afford their users the ability to interact in environments that are very much unlike traditional office spaces. These environments may span multiple and changing contexts, be that the outdoors, public shopping centres, or environments where the user is simply on-the-go. Interactions in such environments may be influenced by factors such as background noise and crowds, and these factors will quickly affect a user's ability to communicate effectively if not catered for with the utmost of care. Applications designed for use in such environments must provide communication modes that are as adaptable to change as the user is. Users will in addition have their own preferences for how they wish to communicate with a computer at any given time, and such requirements must also be taken into account when designing intelligent user interfaces. Some communication modes will, due to their inherent makeup, be better or worse suited to the capturing of input and the presentation of output, depending on the task at hand and the current environment context. Communication modes are also known to evolve, and the current era is witnessing the introduction of tangible and graspable interfaces in which users interact directly with real physical objects that are coupled to digital representations. To be useful, interaction in real-world environments must be natural, flexible, expressive, efficient, accurate, and robust. No single communication mode is ever likely to fulfil all of these requirements, and thus a hybrid solution based on multimodal interaction is required. Multimodal interaction is the central theme of this dissertation. Two additional research areas showing potential for human-computer interaction include anthropomorphization, where inanimate objects are given human-like characteristics, and symmetric multimodality, where both the system and the user are given equally powerful means to communicate with one another.

The environments that are referred to above provide a representative reflection of the environments in which people carry out their day-to-day activities. Two scenarios are referred to throughout this dissertation as a means of representing typical everyday activities, namely pedestrian navigation and shopping. In the pedestrian navigation scenario, a user can navigate indoors

and outdoors and can explore his or her surroundings via on- and off-device interaction with objects such as nearby buildings, rooms, and shelves within a room. In the shopping scenario, the user can interact with different types of products located on a particular shelf, such as digital cameras, mobile phones, grocery items, and in fact any product range imaginable. The communication modes used for interaction in these scenarios include that of speech, handwriting, and gesture, whereby gesture can be further categorized into the types point, pickup, and putdown. Interaction can take place with referents on the mobile device's display, as well as with real-world tangible objects, and may be unimodal or multimodal in nature. In addition, the systems cater for user input that is temporally and/or semantically overlapped, i.e. overlapped in time and with respect to semantic content.

The objectives of this work are manifold. On the one hand, this work sets out to tackle the issues involved in catering for a diverse range of communication modes for use in everyday tasks like pedestrian navigation and shopping. This is achieved from both a theoretical standpoint and from a practical standpoint. The work incorporates not just multimodal input for a range of flexible modalities like speech and handwriting, but also incorporates tangible interaction, and the concepts of anthropomorphized objects and symmetric multimodality. On the other-hand, no work is complete without thorough testing and to this end, the dissertation outlines results from three separate usability field studies covering mode characteristics like accuracy and scalability, user preference and the intuition of modality combinations in public and private environments, and user acceptance for interaction with anthropomorphized objects. A highlight of the work is that the implementations cater specifically for mobile devices.

The following is a series of research questions that this dissertation will address:

- **What communication modes might be appropriate for mobile users in everyday environments like shopping and navigation?** Different communication modes might be better suited to specific contexts. Such contexts may be dependent on aspects like the surrounding environment, the user, and the task at hand. Support for some communication modes like speech is already commercially available, even for mobile devices. Other communication modes like handwriting still require additional custom-made software to augment the limited ability of character recognition. Yet other modes like gesture and tangible interaction must be implemented from scratch due to the lack of existing commercial packages.
- **Can communication modes be combined to produce a superior outcome?** By combining multiple modalities, a range of rich interaction possibilities can be created in which individual modalities can be used to address not just individual tasks, but also different aspects of the same task.
- **What are the attributes of a good communication mode?** Each communication mode possesses a unique set of characteristics, which may change based on the context in which the mode is used. Attributes such as comfort, enjoyment, familiarity, speed, accuracy, scale, accessibility, privacy, intuition, and the complexity of a modality all affect the usability of a communication mode, and thus all need careful consideration if the modes are to be used to their potential.
- **How can these forms of communication be tested for practical use in real-world environments?** Defining a range of unimodal and multimodal interaction types is only the

first step in designing usable systems. Usability field studies conducted in real-world environments are required to identify the practical suitability for individual modalities and their combinations. Aspects considered to require attention include a user's preference for modality combinations, the intuitiveness of modality combinations, the accuracy and efficiency of modalities, and user acceptance for concepts such as anthropomorphization.

- **What infrastructure is required to support mobile users in dynamically changing environments?** Communication modes require supporting device infrastructure that is capable of capturing input from users and that is capable of presenting output back to users. Such devices may be either a part of an instrumented environment, in which case they are limited to a given geographical location, or they may be situated on the user, for example as part of a mobile PDA or phone, in which case they are always accessible by the user but often limited to a given set of resources.
- **What type of architecture is required to support natural and flexible interaction?** A flexible and modular architecture is required to support communication modes that are available at the time of implementation as well as those communication modes that become available in the future. Multimodal input also requires the use of modality fusion strategies in which possibly inaccurate results from different recognizers are analysed and fused and conflicts between multiple modalities are resolved.

The following chapters expand on these descriptions to provide more detail and explanations on the findings.

1.2 Chapter Outline

This dissertation is divided into a number of clearly structured chapters that provide the reader with the ability to start reading from various entry points. Figure 1.1 shows the sequential order of the seven chapters in this dissertation, categorized into three groups. The first group encompassing chapters 2 and 3 provides fundamental background information on the terminology used throughout this dissertation, and it also provides a summary of related work. Readers familiar with the concepts of multimodal interaction and modality fusion are invited to skip directly to the main chapters in this dissertation, namely chapters 4 and 5. The individual chapters are described below.

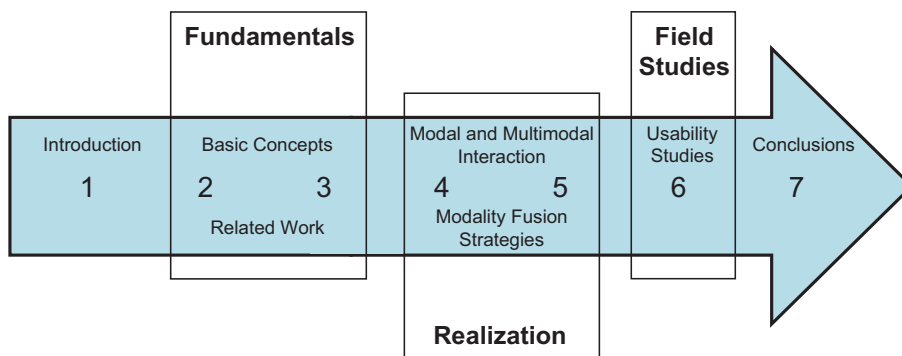


Figure 1.1: Structural overview of the dissertation.

Chapter 2 (Basic Concepts) introduces some basic terminology on human-computer and multimodal interaction, including the definition of terms like multimodality, modality fusion, and reference resolution. This is followed by a discussion on the environments in which multimodal interaction is likely to occur and a scenario walkthrough of the two systems implemented under the course of this dissertation, namely the Mobile ShopAssist (MSA) and the BMW Personal Navigator (BPN). Chapter 3 (Related Work) outlines related work in the fields of multimodal interaction and mobile computing. Particular emphasis is placed on projects that are multimodal and projects that cover the domains of shopping and navigation, two mobile contexts that stand to gain from the incorporation of multimodal interaction.

Chapter 4 (Modal and Multimodal Interaction) outlines the different communication modes used in the MSA/BPN, including speech, handwriting, and gesture. Particular focus is placed on the calculation of confidence values for each of these modes and the results from a field study on modality accuracy and modality efficiency. Following this, the chapter defines a formal classification for multimodal interaction in terms of its temporal and semantic synchrony and in terms of the degree of semantic overlap between different input modes. The concepts of direct and indirect interaction and anthropomorphization are discussed, and this is followed with an outline of symmetric multimodality in the MSA/BPN and a description of the encompassed presentation planning capabilities. The chapter closes with an extension to the MSA/BPN scenario, catering for interaction between multiple users and multiple devices. Chapter 5 (Modality Fusion Strategies) outlines how multimodal input is represented, and also discusses several timing issues relevant to the processing of multimodal input. The main focus of this chapter is however on the modality fusion strategies used in the MSA/BPN. Some of the research topics covered include the ability to re-weight possibly biased recognition results, the selection of relevant information based on timestamps, and the resolution of semantically overlapped and conflicting input.

Chapter 6 (Usability Studies) describes the results from two empirical usability studies designed to measure user preference for 23 different modality combinations in both private environments representative of one's home and public environments representative of a shopping centre. In addition to providing insight into user preference, the studies also detail aspects like modality intuition and user acceptance for conversing with anthropomorphized objects.

Chapter 7 (Conclusions) concludes with a summary of the scientific and commercial significance of this work, and the chapter also highlights several possibilities for future research.

The goal of this chapter is to provide an appropriate backdrop for the work outlined in subsequent chapters of this dissertation. A number of terms regarding multimodal interaction are defined, and this is followed with discussion on the mobile and instrumented environments in which such interactions may occur, including the description of two scenario walkthroughs based on the Mobile ShopAssist (MSA) and BMW Personal Navigator (BPN) applications developed as part of this dissertation. Section 2.1 introduces some basic terminology regarding human-computer interaction, human perception, computer perception, and verbal and nonverbal communication. Section 2.2 continues with a discussion of the benefits that multimodal interaction can provide and a definition of terms like multimodality, multimodal input, and modality fusion. In section 2.3 the term reference resolution is defined, as too are a broad range of discourse phenomena applicable to multimodal applications. Finally, in section 2.4, the setting for mobile users and instrumented environments in which multimodal interaction is likely to occur is described, and the chapter closes with two scenario walkthroughs based on the MSA and BPN applications.

2.1 The Human Senses and Communication

In this section, some basic terminology regarding user input and system output are introduced. The human senses for perception are described, as too are human forms of verbal and nonverbal communication. This is followed with discussion on computer input and output media used to support natural human-computer interaction.

2.1.1 Basic Model for Human-Computer Interaction

Multimodal human-computer interaction can be seen to consist of interacting agents. Similar to projects like MIAMI (Schomaker et al., 1995), the work in this dissertation assumes that there are minimally two interacting agents, a human (or multiple humans) and a machine (or more specifically a computer). For the purpose of this dissertation, the interacting computer agent entails both the computational software models underlying mobile applications like the MSA and BPN, as well as the physical hardware that these applications are built on. The physical hardware encompasses devices like mobile PDAs and any associated I/O peripherals like microphones and displays that may be physically connected to the processing device, or alternatively distributed throughout the environment.

Two basic processes that occur in human-computer interaction are that of control and perception. In (Schomaker et al., 1995), the term *perception* is used to describe the process in which communication from a machine to a human takes place, and *control* is taken to describe the process in which communication from a human to a machine takes place. The process of human-computer interaction can best be described by an interaction loop that can be viewed from the perspective of both the human and computer alike. As a result, the process of control encompasses ‘human output’ (alternatively viewed as ‘computer input’), and the process of perception encompasses ‘computer output’ (alternatively viewed as ‘human input’). In this section, human input is discussed with respect to the human senses and human output is discussed with respect to a human’s ability to communicate. Similarly, computer input and output are discussed with respect to the devices used to capture and present information to a user. The four-stage interaction loop is often simplified to entail just ‘user input’ and ‘system output’ as defined from the perspective of the computer (see figure 2.1). In this case, *user input* refers to information arriving at the computer’s input interface (e.g. speech and handwriting) and *system output* refers to interaction originating from the computer’s output interface (e.g. audio and graphics).

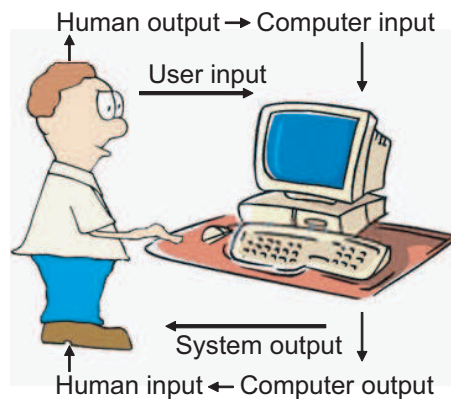


Figure 2.1: Interaction loop consisting of four stages: human output, computer input, computer output, and human input.

2.1.2 Perception and the Human Senses

The human senses allow for *perception*, which (Wikipedia, 2006d) defines in turn as “the process of acquiring, interpreting, selecting, and organizing sensory information”. The recognition of the importance of the human senses and human perception goes back at least as far as Aristotle (384 - 322 B.C.). Aristotle categorized senses into the groups: special, common, and incidental/inferential. The ‘special senses’ have since become known as the five ‘classical senses’: sight, hearing, touch, taste, and smell. Halliday (1998) outlines that sense organs are a person’s “window onto the world” and that (with respect to animals) the senses are “specially adapted to gather information that is of particular biological importance to the animal, for example that relating to its food, predators, and potential mates.” It is clear from this that our senses, their complexities, and their richness evolved foremost as a means for basic survival, but contributed also to specialized functions like communication. This dissertation concentrates on only three of the human senses - sight, hearing, and touch - all of which provide a solid basis upon which communication with humans can be built. These three senses can be categorized into the group of ‘distance senses’

like sight and hearing (Halliday, 1998) and ‘proximity senses’ like touch. The senses of taste and smell, although very important in the physiological makeup of humans, are not easily harnessed with respect to current communication paradigms, be that human-human interaction or human-computer interaction. To demonstrate, for the purpose of communication, the expressiveness of formal language in the auditive modality (e.g. spoken language), the visual modality (e.g. written or sign-language), and the tactile modality (e.g. Braille) is far more advanced than any communication set that currently exists for the modalities of olfaction (smell) and gustation (taste). Generally speaking, the generation and recognition of taste and smell stimuli is still experimental and not readily available in commercial off-the-shelf products, and will thus not be considered any further.

There are in fact at least nine human senses supported by the literature, with the four additional senses being thermoception (heat), nociception (pain), equilibrioception (balance), and proprioception (body awareness) (Wikipedia, 2006e). Table 2.1 (a derivation from that found in (Silbernagl & Despopoulos, 2003)) outlines the five classical senses grouped by: sense, modality, and sense organ. Other classifications of the human senses define up to 21 different senses, and senses are also often grouped under subcategories such as special senses (sight, hearing, taste, and smell) and somatic senses (tactile and haptics). With regard to multimodal systems and tangible user interfaces, the somatic senses have become an important focal point for the creation of new types of input device providing haptic and tactile feedback to the user. Both haptics and tactition relate to the sense of touch, but where ‘tactition’ is often used to describe touch sensations like smooth and rough, ‘haptics’ is used to describe touch sensations like resistance and vibration. Geiser (1990) goes as far as to make the distinction that tactition is a perception modality while haptic a form of output. A typical example of the use of the somatic senses can be seen when driving a car, where the driver is able to perceive the composition and contour of the steering wheel (e.g. leather) and the feedback on road conditions provided through the steering wheel (e.g. wheel jolts). Simple devices that are able to self-generate haptic sensations include computer joysticks, while more complex devices include the PHANToM¹, which allows a user to feel resistance when interacting with a mechanical arm in three degrees of freedom. Research projects like MIAMM (Reithinger et al., 2005) (see also chapter 3) focus specifically on research into tactile and haptic interaction and how a user’s perception of these senses can lead to better user interface designs.

Sense	Modality	Sense Organ
Sight	Vision/Visual	Eyes
Hearing	Audition/Auditory	Ears
Touch	Tactition/Tactile	Skin
Taste	Gustation/Gustatory	Tongue
Smell	Olfaction/Olfactory	Nose

Table 2.1: The five classical senses defined by Aristotle (384 - 322 B.C.).

The human sense organs (e.g. the ears) do not support input and output equally, and their respective senses describe only human methods of perception rather than control (e.g. there is no sense of speaking, only of hearing). During communication however, a compatible medium must exist in which mappings from generated output and received input can be conducted. Dance

¹Sensible PHANToM, <http://www.sensible.com>

(1982) defines this medium as the mode, and points out that for communication to be successful each sender mode must have a complementary receiver mode (e.g. oral/aural). Nigay and Coutaz (1993) take the definition one step further by making a distinction between the term *mode*, which is defined in a practical sense to be “the way information is interpreted to extract or convey meaning” (e.g. spoken language or speech, handwriting, and gesture), and the term *modality*, which is defined in a more theoretical sense to be “the type of communication channel used to convey or acquire information” (e.g. vision, audition, and tactition). In general however, the literature makes little distinction between the terms mode and modality, and indeed Nigay and Coutaz (1993) state that the terms are both derived from the same word ‘modal’. A consequence of this is that the term modality is often used both for describing communication modes like spoken language as well as modalities like audition.

The relationship between sender and receiver modes can be seen in table 2.2, which depicts the communication modes a user might use to send and receive information in different modalities. Table 2.2 is user-centric, meaning that only the perspective of the user (and not the computer) is taken into account, and thus more closely resembling human-human interaction rather than human-computer interaction. To illustrate, for the communication mode of gesture, which can be taken to refer to actions in the MSA like pointing at, picking up, or putting down physical objects, a user would require the modality of tactition when providing an interaction and the modality of vision when receiving such an interaction.

Control (User Sends)	Modality		
Communication Mode	Visual	Auditory	Tactile
Speech		X	
Handwriting			X
Gesture			X

Perception (User Receives)	Modality		
Communication Mode	Visual	Auditory	Tactile
Speech		X	
Text	X		
Gesture	X		

Table 2.2: The communication modes a user would require when sending and receiving information in different modalities.

2.1.3 Verbal and Nonverbal Communication

Communication is the process of exchanging information via a common system of symbols. Ogdan and Richards (1923) define it as “the use of symbols in such a way that acts of reference occur in a hearer which are similar in all relevant respects to those which are symbolized by them in the speaker”. Communication may be ‘one-to-one’ (e.g. communication between two people), ‘one-to-many’ (e.g. a public address), ‘many-to-many’ (e.g. social gatherings like parties), or ‘now-to-future’ (e.g. entries in a diary). In the mobile applications described in this dissertation, one-to-one and now-to-future communication modes are used. One-to-one communication takes place when a user queries the MSA or BPN application for information, and now-to-future com-

munication takes place as a result of the interaction logs that are created for the user to look back on at a later time through the use of applications like SharedLife (Wahlster, Kröner, & Heckmann, 2006) and SPECTER (Kröner, Heckmann, & Wahlster, 2006), which are used to share augmented personal memories and track a user's actions and affective states.

In (Leathers, 1997), a classification of communication types is defined incorporating not only the communication modes used in applications like the MSA and BPN (e.g. speech, handwriting, and gesture), but also a variety of other communication types that may in the future become relevant for multimodal systems as research in interaction-processing evolves. According to the classification shown in figure 2.2, communication is categorized as being either verbal, nonverbal, or a combination of both verbal and nonverbal (Leathers, 1997). While *verbal communication* refers to spoken and written language, both of which are a major source of meaning in applications like the MSA and BPN, *nonverbal communication* refers to the communication of information that is not conveyed by the literal meaning of words, i.e. "communication without words" (Dance, 1970). In the MSA and BPN, nonverbal communication takes place when users interact with referents via point, pickup, and putdown actions.

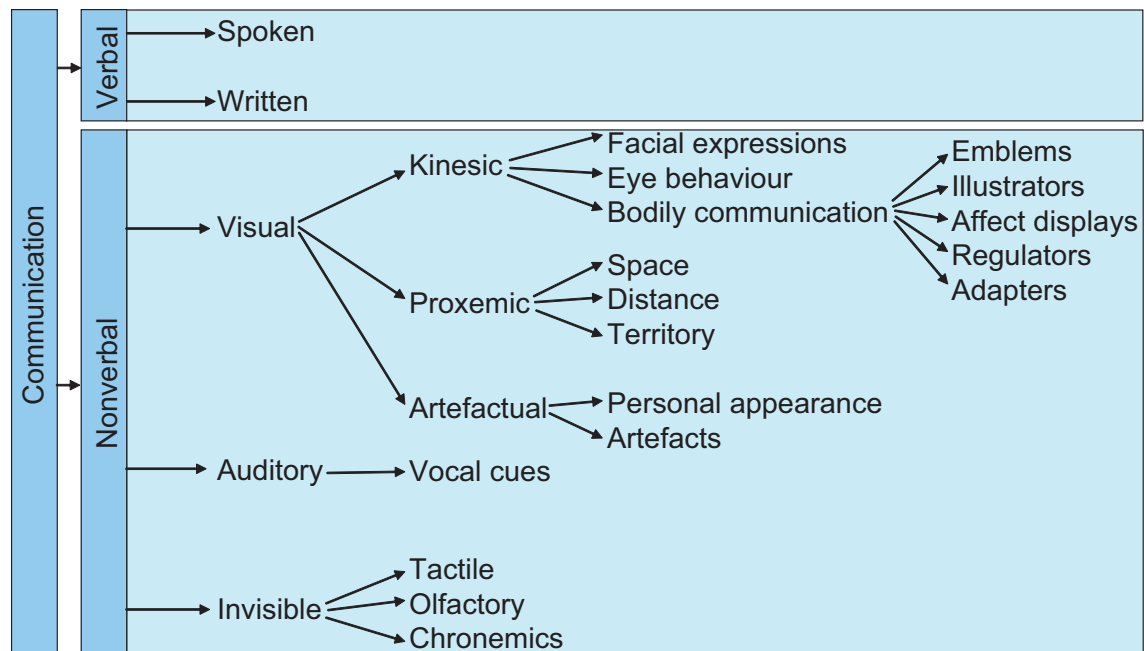


Figure 2.2: Classification of verbal and nonverbal communication.

To better understand the broad coverage of human communication, several of the significant subtypes are discussed. Nonverbal communication encompasses three major interacting systems: the visual-, auditory-, and invisible-communication system. While 'auditory communication' concerns itself with vocal cues and the meaning that sound can convey, 'invisible communication' encompasses the subsystems of tactile, olfactory (e.g. natural body odours), and chronemic (intercultural) communication. Relevant to the MSA and the BPN is however 'visual nonverbal communication', which can be grouped by the categories: kinesic, proxemic, and artefactual communication.

'Kinesic communication' refers to the communication subsystem consisting of facial expres-

sions, eye behaviours, and bodily communication, while ‘proxemic communication’ refers to the use of space, distance, and territory for communication purposes, and ‘artefactual communication’ concerns itself with personal appearances and the artefacts that people wear, e.g. clothing and cosmetics. The selection gestures used in the MSA and BPN applications like ‘point’, ‘pickup’, and ‘putdown’, belong to the class of ‘nonverbal visual kinesic bodily communication’.

‘Bodily communication’ used in the MSA takes the form of emblems and illustrators. ‘Emblems’ (Ekman & Friesen, 1969) refer to bodily cues that have a direct verbal translation consisting of a word or two, e.g. hand beckoning to imply “come here”, or in the case of the MSA pickup and putdown actions to imply product selection and product deselection. ‘Illustrators’ are much the same as emblems in that they are used with intentionality, but augment what is being said. Common types of illustrators include batons where movements emphasize a particular word or phrase, pictographs where a picture of a referent is drawn in the air, and particularly relevant for the MSA and BPN, deictic movements, where a referent such as an object, place, or event is pointed at. One beneficial use of emblems and illustrators is demonstrated in (Rogers, 1978) where it was found that gestural illustrators result in a significant increase in the comprehension of spoken words. These are increasingly useful as noise is introduced and when a spoken message becomes more complex. Some scenarios attract the use of gestural illustrators more than others, and in (Cohen, 1977) it was found that subjects giving directions on how to get from one place to another used significantly more hand illustrators in a face-to-face situation than when giving directions over an intercom. Subjective results from a usability study conducted on the MSA (see section 6.2.5) further indicate that aside from providing a simpler and more robust method of interaction, gestural illustrators also provide a method of interaction that users find fun and captivating to use.

2.1.4 Computer Input and Output Media

The flip side of human perception and human control is that of computer perception and computer control, or more specifically the capturing of information and presentation of information by the system. Natural human-computer interaction can be modelled on that of human-human interaction, but this requires computers to be able to perceive and express information through similar modalities to humans, e.g. vision, audition, and tactition. The ability for computers to reciprocate with users in the same modalities is referred to as ‘symmetric multimodality’, a concept defined in section 4.4.

Typical computer input devices include keyboard, mouse, pen, camera, and microphone. 3D input devices are also gaining in popularity (e.g. data glove²), and mobile users are seeing the emergence of input devices for tangible interfaces based on sensor technology such as RFID-instrumented products, shelves, and shopping trolleys. Typical computer output media examples include displays and projectors for vision, loud speakers for audition, and a range of emerging tactile/haptic devices for the somatic senses. Although a standard mouse and keyboard provide some haptic feedback when a key or button is pressed (known as the ‘breakaway force’), the term tactile/haptic is more commonly used to refer to devices specifically designed for a particular purpose, including pneumatic stimulation based on the control of air jets, vibrotactile stimulation based on vibrations generated by blunt pins, voice coils or piezoelectric crystals, and electro-tactile stimulation based on small electrodes attached to a user’s fingers to provide electrical pulses (Schomaker et al., 1995).

²Data glove, see <http://www.vrealities.com> and <http://www.fakespace.com>

The SmartKom project³ provides a solid foundation for many of the concepts outlined in this dissertation, and is thus often cited when comparisons are drawn between state-of-the-art multi-modal systems in general, and the MSA/BPN implementations created as part of this dissertation. In the SmartKom project, communication is defined to take place through the use of code, media, and modalities. In (Maybury & Wahlster, 1998; André, 2003), these terms are defined. The term ‘mode’ is used to refer to “different kinds of perceptible entities (e.g. visual, auditory, haptic, and olfactory)”, while the term ‘media’ relates to “the carrier of information (e.g. paper or CD-ROM), different kinds of physical devices (e.g. screens, loudspeakers, microphones, and printers), and information types (e.g. graphics, text, and video)”. Finally, the term ‘code’ is used to refer to “the particular means of encoding information (e.g. sign languages and pictorial languages)”. As shown in figure 2.3, human-computer interaction takes place through the use of the human senses, while computer-human interaction takes place through the use of physical information carriers (i.e. media), and supporting this interaction is a communally agreed-upon code system that incorporates language, graphics, gesture, mimics, and/or other.

For many systems, the agreed-upon code is in fact represented by a number of separate codes that are dependent on the individual modes and media, and may or may not map to the same semantics. One such system is WIP (André et al., 1993), which is a knowledge-based presentation system capable of generating coordinated graphics and text output for a variety of how-to-use applications (e.g. espresso machine, lawn-mower, and modem). To provide the ability to present instructions via graphics, text, or a combination of both, the system required multiple codes that each map to the same semantics. The WIP system points out that the connections between code, media, and modalities need not be 1:1, for example the language and graphic codes map to the use of a single computer display (i.e. two codes:one media).

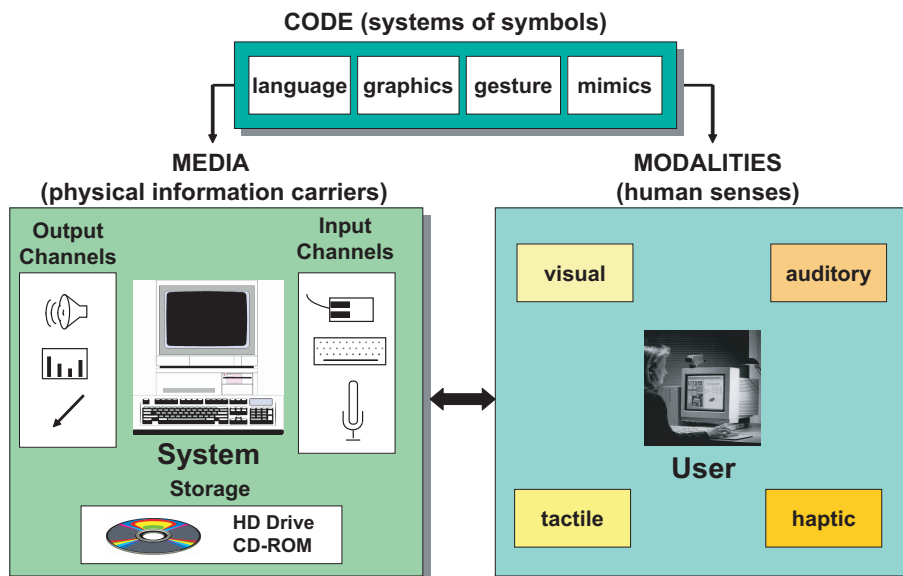


Figure 2.3: Code, media, and modalities (Maybury & Wahlster, 1998).

In the MSA/BPN, the code systems used by both human and computer include spoken and written language, gesture, and graphics. For user input, humans use the modalities of vision and

³SmartKom, <http://www.smartkom.org>

tactition, and the computer supports the perception of such input through the use of media types like a display, pointing device, microphone, and instrumented tangible objects. Similarly, for system output, the computer generates modal information via media types like a display and a speaker, and the human perceives such input via the modalities of vision and audition, as shown previously in table 2.2. Similar to the WIP system, m:n relationships also exist between the code, media, and modalities used in the MSA/BPN. For example, taction can be used to interface both the display and nearby real-world objects (one modality:multiple media), and a touch display caters for both point gesture and written language recognition (one media:multiple codes).

2.2 Multimodality and Modality Fusion

This section describes the benefits of systems that provide their users with the ability to interact multimodally, and the section also defines terminology such as multimodality, multimodal input, modality fusion, and mutual disambiguation.

2.2.1 Benefits of Multimodal Interaction

Multimodal applications provide a range of benefits over unimodal applications and traditional applications designed along the WIMP (Windows, Icon, Menu, Pointer) desktop computing paradigm. In (Oviatt & Wahlster, 1997; Oviatt, 2003), some benefits are outlined to be: naturalness, transparency, ease of use, ease of learning, flexibility, efficiency, and suitability for more challenging applications, for more adverse conditions, and by a broader spectrum of the population. Of significant effect is the statement “a single modality simply does not permit all users to interact effectively across all tasks and environments” (Oviatt, 2003). This is particularly relevant to mobile scenarios in which users often find themselves interacting with different tasks and under constantly changing environment contexts. Three benefits of multimodal interaction that are particularly relevant to mobile applications are that of flexibility, efficiency, and adaptability to context. Not all multimodal interfaces possess the same level of functionality, but from an ideal perspective, these three aspects can easily contribute to an interface that people will use and will enjoy using rather than an interface that people will not accept to use.

- **Flexibility:** Multimodal applications allow for the use of input modes in a complementary fashion (e.g. speech-gesture combined interaction), or in a unimodal fashion that allows users to switch among modes at different times (e.g. speech for some interactions, gesture for other interactions). In effect, this gives users the power to employ each mode for its strengths, thus leveraging a person’s natural ability to select modes that provide for accurate and efficient communication within a given context and modes that best suit their current needs. Delegation of this function to the user has even gained momentum in areas such as work injury prevention where media/modality overuse can result in repetitive stress injury (RSI) or such like.
- **Efficiency:** The ability for multimodal systems to process input from different modalities in parallel is one feature contributing to efficiency. In (Oviatt, 1997), speech-pen interaction was shown to yield 10% faster completion times over speech-only interaction during visual-spatial tasks. In studies conducted on the MSA in which subjects requested feature and object information on products in a shopping domain (via the modes speech, handwriting,

and gesture), it was shown that speech-gesture interaction was 1.11 times faster (i.e. 11% faster) than speech-only interaction and 2.11 times faster (or 111% faster) than both gesture-only and handwriting-only interaction. Multimodal communication is also often shorter than unimodal communication due to specific lexical content like names, e.g. 'PowerShot S1 IS', being better suited to one modality over another, in this case selection-gesture over speech. Subjective results from usability studies conducted on the MSA also show that in addition to multimodal interaction often being shorter, it is also often simpler. This is particularly the case with regards to the pronunciation of names like 'PowerShot S1 IS', where it may not be clear to a user how exactly to pronounce the particular camera name by speech, and thus simpler to use a modality like gesture instead.

Efficiency in multimodal systems also distinguishes itself from that of unimodal systems with regards to error handling and in particular error avoidance and error recovery. Studies have for example shown that users tend to select input modes that they judge to be least error prone, and input disfluencies (e.g. self-corrections, spontaneous repetitions, and false starts (Oviatt, 1997)) have also been shown to occur less frequently in multimodal interaction, thus leading to error avoidance. This is further reinforced by studies showing that users are more likely to switch between modalities when errors occur, thus facilitating error recovery (Oviatt, 1999).

- **Adaptability to context:** Multimodal applications can accommodate a wider range of users, tasks, and environments, all of which are aspects that were either poorly provided for or not at all provided for in the past. This allows users to engage computer systems with modality combinations best suited to their own preferences (e.g. speech for people that are keen on talking), to a particular task context (e.g. handwriting for personal information, speech for general information), or to a particular environment or situation context (e.g. handwriting in noisy environments, speech while on-the-go).

User demographics for a particular application are likely to differ in aspects like age, skill, native language, cognitive style, sensory impairments, temporary illness, and permanent handicaps (Oviatt, 2003). Thus, a visually impaired user or a user suffering from RSI will choose speech over other modalities, while a user with a hearing impairment or accented speech will choose to use handwriting as his or her preferred communication mode. One effect of not providing for multiple communication modes is outlined in (Archambault & Burger, 2001), where it is stated, with regards to the accessibility of data on the Internet, that "the use of the Internet first seemed very promising for visually impaired users due to its textual composition, but as graphics became more mainstream, the accessibility of the Web quickly became a problem".

Multimodal applications extend current computing capabilities by encompassing challenging scenarios including those in which the user is mobile or on-the-go. For such contexts, traditional desktop computing paradigms like WIMP are simply no longer applicable because the computing devices used in mobile scenarios lack even the most minimal features of their desktop counterparts like display space and the ability for a user to conveniently use a mouse and keyboard on a flat surface. For such mobile contexts, modalities like speech, handwriting, selection-gesture, and tangible interaction with the surrounding world are much more appropriate forms of human-computer interaction. Furthermore, through the selection of appropriate modalities, multimodal applications are better able to adapt to

adverse conditions and situations in which ambient conditions are constantly changing, as is often the case during the mobile use of applications in everyday contexts like indoor and outdoor navigation, exploration, and shopping.

2.2.2 Multimodality Defined

Multimodality is defined by Nigay and Coutaz (1993) in terms of ‘multi’ which literally means ‘more than one’ and ‘modal’ which covers the notion of ‘modality’ as well as that of ‘mode’. In (Charwat, 1992), a definition of the term modality limited to three of the human senses is given, i.e. “perception via one of the three perception channels ... visual, auditive, and tactile”, and in (Wahlster, 2006b), *modality* is said to refer to “the human senses which allow incoming information to be received and processed”. These definitions on multimodality are accepted for this work, and also provide a foundation from which one can differentiate between multimodal and multimedia systems. A commonality between these two types of systems is that they both use multiple communication channels but whereas the latter is focused on the actual medium or technology like audio, graphics, and video (Macdonald & Vince, 1994), multimodality is focused on the perception of the senses and the process in which user input (e.g. speech, handwriting, and gesture) is captured by a system. Although not the central focus of this dissertation, multiple output devices may also be incorporated under the banner of multimodality, e.g. audio and visual output produced by a display and a loudspeaker (Schomaker et al., 1995; Elting, 2002), and for these systems the basic distinction between multimedia and multimodality is that multimodal systems understand the semantics of what they capture or present, while multimedia systems do not.

Extending the concept of multimodality, *multimodal interaction* is defined to be “the means for a user to interact with an application using more than one mode of interaction, for instance offering the user the choice of speaking or typing, or in some cases allowing the user to provide a composite input involving multiple modes” (W3C-EMMA, 2005). In the MIAMI taxonomy (Schomaker et al., 1995), the authors differentiate between interaction that uses only one modality (unimodal interaction), exactly two modalities (bimodal interaction), and two or more modalities (multimodal interaction). The W3C Multimodal Interaction Requirements (W3C-MMIReqs, 2003) furthermore point out that whilst a system may be multimodal, user interaction with such a system need not always be multimodal because a user may at some times interact unimodally and at other times multimodally with the system.

In this dissertation, the reader may observe at times that different terminology is used synonymously. The word ‘interaction’ and the word ‘input’ are for example often used in a synonymous manner, i.e. ‘multimodal user input’ and ‘multimodal user interaction’. Some of the terminology used to describe multimodal interaction in this dissertation has also been borrowed from work on traditional spoken dialogue systems. This is seen by the use of terms like ‘utterance’ (e.g. input or dialogue utterance) and ‘speaker’ (e.g. “interaction from the speaker can take the form of...”). In a dissertation on multimodal interaction, such terms refer not just to ‘spoken language’, but rather to ‘multimodal language’ consisting of speech, handwriting, gesture, a combination thereof, or other.

2.2.3 W3C Classification of Multimodal Input

The W3C EMMA Working Draft (W3C-EMMA, 2005; W3C-MMIReqs, 2003) classifies three different types of multimodal input or interaction: sequential, simultaneous, and composite.

- **Sequential input:** *Sequential input* is input that is received on a single modality though that modality can change over time. In effect, this means that a user may interact with only one modality at a time, switching between modalities as needed, for example a user might fill personal details into a form on the Internet by first using a keyboard to enter their name into a field and by then using a pointing device like a mouse to select their gender from a drop-down box. The processing of sequential input does not require multimodal integration (e.g. natural language understanding or reference resolution).
- **Simultaneous input:** *Simultaneous input* is input received on multiple modalities but processed separately in the order in which they were received. For example, a user might use a force-feedback steering wheel together with an accelerator/brake pedal set to pilot a Formula 1 race car around a track. Although input is received on multiple modalities in this example, the input from the steering wheel is treated independently to that of the pedal set.
- **Composite input:** *Composite input* is input received on multiple modalities at the same time but processed as a single integrated compound input. Composite inputs have component parts in different modes, for example a user might say “zoom in here” in the speech mode while drawing an area on a graphical display in the ink mode. As the EMMA Working Draft points out, a central motivating factor for systems to allow for composite input is that different kinds of communicative content are best suited to different input modes. In the example above where a user draws an area on a map and says “zoom in here”, the zoom command is easiest to provide in speech while the spatial information is easiest to provide in ink.

These defined terms on multimodal input are relevant to this work because they are likely to be adopted by industry and the community at large if or when the W3C Working Draft in which they are defined succeeds in becoming a W3C Recommendation. These terms however only represent an abstract start to categorizing multimodal input, and lack depth of coverage, perhaps due to the still maturing nature of research into multimodal interaction, or perhaps due to the ‘draft’ state of the W3C EMMA working document. Another reason for the lack of coverage of these terms is the large topic area that EMMA covers, including for example applications ranging from unimodal to multimodal (where with respect to a single point in time, sequential input is analogous to unimodal input), and input devices ranging from those reminiscent of the stationary desktop computing era (e.g. QWERTY keyboard, DTMF handset, mouse, joystick) up to those representative of state-of-the-art language technology (e.g. speech, handwriting, gaze, and gesture recognition).

Multimodal input as it is discussed in this dissertation excludes the above described sequential and simultaneous inputs, taking only the category of composite input as its starting point, i.e. input that consists of multiple modalities which are unified to form a single integrated compound input. Sequential input is also possible in the MSA and the BPN but is discussed under what this dissertation classifies as unimodal interaction. Simultaneous input, which is also possible in the MSA and the BPN, is similarly not of relevance to this work as it is a field of study that is already mature, dating back many years to systems like SAGE⁴ (1963) where a pointing device could be

⁴SAGE Website, <http://www.mitre.org/about/sage.html>

used in conjunction with slide-switches similar to those on an electronic synthesizer, to interact with objects on a graphical display.

2.2.4 Modality Fusion

Having now defined some of the fundamental terminology behind multimodal interaction, it is important to describe how such interaction can be reliably processed. *Multimodal integration* is defined in (W3C-EMMA, 2005) as the process of combining inputs from different modes to create an interpretation of composite input. Multimodal integration is also often referred to as ‘multimodal fusion’, ‘modality fusion’, and ‘media fusion’. The term modality fusion, used henceforth in this dissertation, was first defined under the SmartKom project together with its reciprocal modality fission. Borrowed from terminology used in physics, the goal of *modality fusion* is to combine multiple modality input streams - for example provided by a user in the form of speech, handwriting, and gesture - into a single result that is modality-free but rich in semantic meaning. *Modality fission* on the other hand is responsible for splitting semantic meaning from within a modality-free utterance into different modality streams for presentation back to a user via media channels like a display and speaker. Figure 2.4 shows two schematic multimodal architectures from the EURESCOM MUST (Boves & Os, 2002) and SmartKom (Wahlster, 2006b) projects, both of which incorporate a modality fusion and a modality fission component. The terms modality fusion and modality fission can be seen to respectively relate to the interpretation of computer input and the generation of computer output, as shown in figure 2.1. In (Wahlster, 2003), the key function of modality fusion is outlined as being “the reduction of the overall uncertainty and the mutual disambiguation of the various analysis results”.

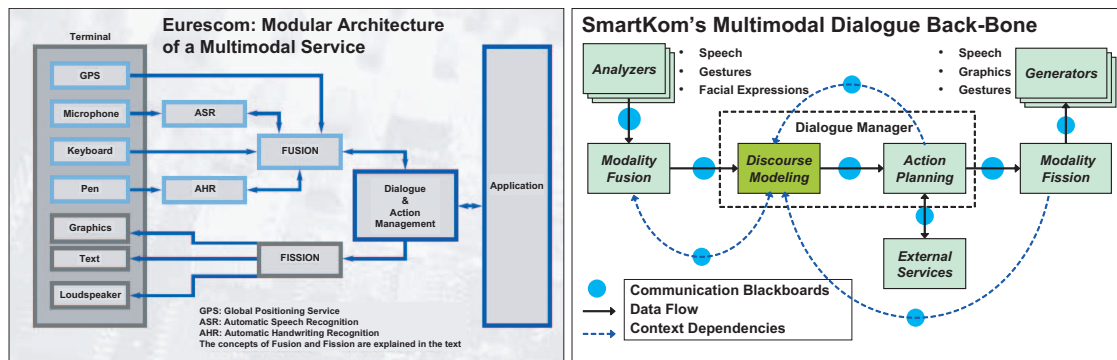


Figure 2.4: Two schematic architecture diagrams showing the incorporation of modality fusion and fission components in the EURESCOM MUST project on the left (Boves & Os, 2002) and in the SmartKom project on the right (Wahlster, 2002a).

Mutual disambiguation (Oviatt, 2000c) is another closely related term and refers to “the disambiguation of two input signals such that each mode provides partial information and dialogue context that aids in the interpretation of the other mode”. In (Kumar, Cohen, & Coulston, 2004), it is defined to occur when “the top-ranked multimodal command [in a list of interpretations] includes an interpretation from speech and/or from gesture that is not itself top-ranked for that modality”. Extending on this, the ‘rate of mutual disambiguation’ can be seen to be the percentage of correct multimodal commands in which mutual disambiguation between the input modalities takes place. In (Kumar et al., 2004) a distinction is made between calculations of this rate based

only on so-called ‘pull-ups’ that gain a top-ranking in the resulting N-best list, and calculations that incorporate all pull-ups including those that gain a higher-ranking but not necessarily a top-ranking. Only the more restrictive rate of mutual disambiguation allows significant parallels to be drawn with system accuracy because a high rate of mutual disambiguation need not necessarily influence the overall accuracy of the system if it is based on pull-ups that do not lead to best results. The term mutual disambiguation also bears close resemblance to the term super-additivity, which in the field of speech enhancement is used to describe the additivity effects occurring when an acoustic stream of phonemes (from speech) is supplemented with information from a visual stream of visemes (corresponding to lip movements).

2.2.4.1 Early and Late Fusion

The process of modality fusion can be applied at two different stages, giving rise to the terms early fusion and late fusion. *Early fusion* is said to occur when modality inputs are integrated at an early stage of processing such as before the input signals are sent to their respective recognizers. At this level of processing, the fusion is also commonly called subsymbolic fusion and is based on techniques like neural networks and hidden Markov models. Early fusion integration generally occurs through the analysis of feature vectors between multiple input signals and is considered appropriate for input modalities that have close temporal bonds such as speech and lip movement recognition (Rubin, Vatikiotis-Bateson, & Benoit, 1998; Stork & Hennecke, 1995), and emotion and facial expression recognition (Wahlster, 2003). In (Bregler, Manke, Hild, & Waibel, 1993; Pavlovic, Sharma, & Huang, 1997), further examples of systems employing an early fusion technique for combining multiple input streams are described. *Late fusion* in comparison occurs when modality inputs are integrated at a late stage of processing, such as after the signals have passed through their respective recognizers. Known also as symbolic fusion, the processing techniques employed for late fusion can include graph unification and Bayesian networks. This type of integration occurs at a semantic level, where utterances are first analysed for meaning and then fused. Late fusion techniques are often applied to the processing of modalities that do not have tight temporal bonds, for example the fusion of speech, handwriting, and/or selection-gesture.

In (Wahlster, 2003) one disadvantage of early fusion is stated to be that integration on a signal level makes back-tracking and the reinterpretation of a result more difficult. It is also difficult to pre-specify all varieties of crossmodal references at such a stage, thus making a system’s ability to cope with unusual or novel uses of multimodality complicated. The benefit of such an early fusion approach is however that potentially useful information can be gained that would otherwise be thrown away by the time late fusion takes place. The benefit of a late fusion approach in comparison is that the robust interpretation of incomplete and inconsistent multimodal input becomes more reliable at later stages due to more semantic knowledge becoming available from the different sources.

SmartKom is one system that combines both early and late fusion techniques when processing multimodal input, for example late fusion techniques are used to interpret combined speech and gesture interaction, while early fusion techniques are used to interpret combined emotional prosody and facial expression to compute a user’s affective state (Batliner et al., 2000). Systems like QuickSet (Cohen et al., 1997) and indeed the MSA/BPN described throughout this dissertation focus only on late fusion techniques.

2.2.4.2 Generalized Architecture of a Multimodal System

Figure 2.5 shows a generalized multimodal system architecture (Maybury & Wahlster, 1998). This architecture shows many of the central components required by multimodal systems for interaction processing and output generation, including modality specific analysers, multimodal interaction specific components, the application interface with its explicit application model, and the multimodal media design for the planning of output. As described in detail in chapter 5, the MSA/BPN has a similar architecture. Also seen in the figure are the knowledge-bases from which inferences can be made, including the user, discourse, domain, task, and media models.

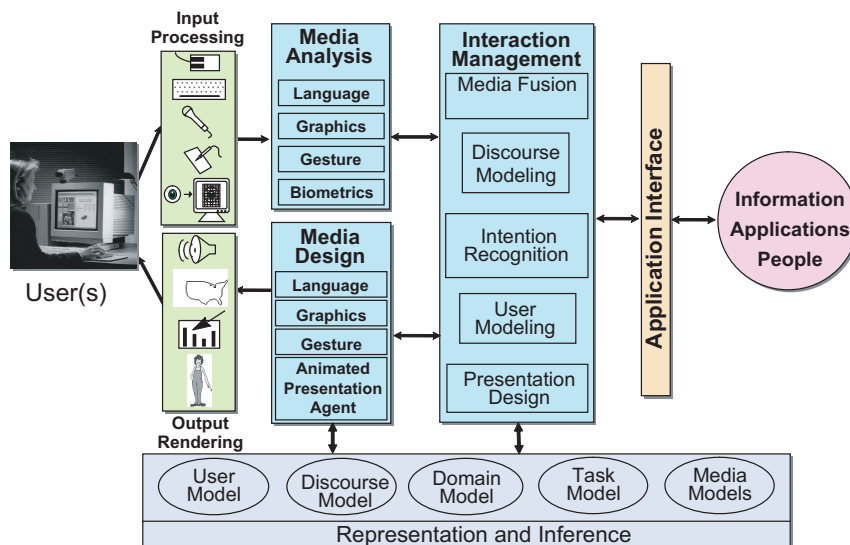


Figure 2.5: Architecture showing typical components required by multimodal systems (Maybury & Wahlster, 1998).

2.3 Reference Resolution

One important requirement of multimodal systems is their ability to resolve references that occur within the different modalities. Most research in this field has been linguistically motivated (i.e. based on verbal communication like spoken and written language), but the importance of non-linguistic information (i.e. non-verbal communication such as deictic gesture) is also becoming increasingly important as multimodal applications begin to focus on scenarios that are based on instrumented environments and mobile users. Many multimodal systems use speech or handwriting as the predominant modality and gesture as an aid to reference resolution, for example in determining missing lexical entities in the spoken or written input. Such systems (e.g. MUST (Almeida et al., 2002) as described in section 3.1.5) have the downfall that interaction although multimodal, always requires the predominant modality to be used. A prominent feature of the MSA is that interaction does not bias one modality over another, meaning that users are equally able to interact with the modality of gesture as they are in other modalities like speech and handwriting. This is achieved through novel interaction metaphors like a visual What-Can-I-Say (WCIS) scroll bar to access product features via pointing-gesture and the support for real-world tangible interaction like

pickup and putdown actions to access the actual products. In both cases, non-linguistic references are mapped directly to the underlying semantic language model.

Reference resolution is the process in which referring expressions within a (possibly multi-modal) dialogue are matched to individual referents. As an example, in the combined speech (S) and gesture (G) utterance: $\langle S = \text{"What is the price of this camera?"} \rangle \langle G = \text{"PowerShot S50"} \rangle$, the referring expression 'this camera' which occurs in the speech modality is matched to the referent 'PowerShot S50' occurring in the gesture modality. Reference resolution is a significant contributor to the process of modality fusion, and can occur within a single modality like speech or gesture/graphics, but also within multimodal contexts. For example, in the MSA/BPN, pointing at objects on the display requires the resolution of visual references, while referring to objects by their name during spoken and written communication requires the resolution of linguistic references, e.g. "What is the price of the PowerShot S50?", where 'PowerShot S50' is the reference to the concrete object in the product database. In the pointing-gesture example, selecting the correct referent involves calculating which particular object is closest to the screen coordinates pointed to on the display, while the speech example requires analysis on which object sounds phonetically most similar to the user's utterance. In (Chai, Hong, & Zhou, 2004), a probabilistic approach to reference resolution using a graph-matching algorithm is described, for which the author states that information from multimodal inputs, the interaction context such as the conversation history, visual feedback, and domain knowledge are all required to find the most probable referents. To resolve references, including complex interactions consisting of multiple referring expressions and multiple modalities, semantic constraints applicable to the referring expressions and contextual constraints from prior conversation need to be considered. The techniques used for reference resolution and also modality fusion and mutual disambiguation are left to chapter 5. The goal of this section is to outline three fundamental terms upon which reference resolution is based, i.e. the referents, referential terms, and referring modes. The section also analyses a multimodal checklist derived in the SmartKom project to identify a range of multimodal discourse phenomena requirements that symmetric multimodal systems should fulfil.

2.3.1 Referents, Referential Terms, and Referring Modes

Reference, as defined by Loos (2003) is the symbolic relationship that a linguistic expression has with the concrete object or abstraction that it represents. As a simple example, reference occurs when a user identifies an object like the 'PowerShot S50' during one interaction with the MSA application and then in subsequent interactions refers to the object via a linguistic expression like 'it' or 'the camera'. When this occurs, 'it' and 'the camera' become pointers or references to the concrete but more complex linguistic expression 'PowerShot S50'. One reason why reference is so common in language is that it makes communication more efficient. This is demonstrated for human-human communication by Grice's 'Maxim of Quantity' (Grice, 1975), which is defined as: 1. "Make your contribution to the conversation as informative as necessary" but 2. "Do not make your contribution to the conversation more informative than necessary". The field of linguistics defines a variety of different types of references that are used in language. The common challenge regarding the interpretation of such references, called *reference resolution*, is that the references and the referents that they refer to may be separated in discourse by an unknown period of time and may even span linguistic and non-linguistic contexts. This results in an increased degree of ambiguity that systems need to resolve. For example, a user might refer to an object during a spoken utterance but point to it in the real-world via a gesture: $\langle S = \text{"What is the price?"} \rangle \langle G = \text{"PowerShot"} \rangle$

S50”>. Fortunately for applications like the MSA/BPN, Kehler (2000) points out that in contrast to human-human interaction in which very little semantically overlapped input is given, in human-computer interaction users are generally “far less convinced of a computer’s ability to understand natural language, and are thus inclined to sacrifice some degree of conversational coherence in an effort to reduce ambiguity”. Nonetheless, reference resolution remains a vital requirement for all modality fusion modules.

Reference can occur either linguistically or non-linguistically. When reference occurs linguistically, it is based on verbal communication like spoken and written language, while when it occurs non-linguistically (which is also known as extra-linguistic reference), it is based on non-verbal communication like deictic gestures that can be used to select tangible objects in the surrounding real-world. The use of non-linguistic references during human-computer interaction is a field of study that is not encompassed by traditional spoken dialogue systems. For multimodal dialogue systems and tangible interfaces however, non-linguistic references are common practice and can contribute greatly to human-computer communication. It is for this reason that precisely these non-linguistic references form an important contribution to this dissertation.

Whereas referents represent the concrete objects or concepts (e.g. ‘PowerShot S50’), referential terms represent the class of linguistic elements that can be used to refer to referents (e.g. ‘it’, ‘this’, ‘that’), and referring modes represent the class of reference (e.g. endophora).

In (Allen, 1995), a range of different reference types are described including coreference, endophora and exophora. In (Landragin & Romary, 2003), these reference types are called *referring modes* and the linguistic constructs that are used to represent the referents (e.g. nouns like ‘PowerShot S50’, pronouns like ‘it’, and demonstratives like ‘this/that’) are classified as *referential terms*. Wahlster (2003) defines several additional referring modes that are particularly relevant to multimodal discourse, including ellipsis and crossmodal reference.

In this dissertation, a particular focus is placed on the use of references within multimodal dialogues, and a goal of the work is to satisfy “the full spectrum of dialogue phenomena that are associated with symmetric multimodality” as outlined in (Wahlster, 2003). The referring modes most relevant to this dissertation are defined below.

- **Endophora:** Endophora is reference within one expression that is directed to the same referent in another expression occurring either before or after it, i.e. reference to something that has already been or will soon be mentioned in the text. If the referent occurs before the endophoric reference, it is called anaphora, while if it occurs after the reference it is called cataphora. The term anaphora is in the literature however often used synonymously with the term endophora to include both anaphora and cataphora.

As an example, anaphora occurs in the MSA/BPN application when the user speaks out the following two sequential utterances: “Find me the PowerShot S50.” “What is its price?”. In this case, the reference ‘its’ refers back to the referent ‘PowerShot S50’. Such references are resolvable in the MSA/BPN application through the use of a small history context that contains the logged entries of discourse entities that have been evoked in the recent past, like features (e.g. ‘price’ and ‘megapixels’) and objects (e.g. ‘PowerShot S50’).

- **Exophora:** Exophora is reference that is made to something extra-linguistic, i.e. non-lingual. This is in contrast to the above definition of endophora in which reference is made to something intra-linguistic or lingual. One common form of exophora that is particularly relevant to tangible interfaces is that of deixis. ‘Deixis’ is defined as being “reference by

means of an expression whose interpretation is relative to the usually extra-linguistic context of the utterance, such as who is speaking, the time or place of speaking, the gestures of the speaker, or the current location of the discourse” (Loos, 2003). Deixis types that occur during typical interaction in the MSA/BPN include person deixis, place deixis, and pars-pro-toto deixis.

- **Person deixis:** Person deixis is deictic reference to the participant role of a referent such as the speaker, the addressee, or referents that are neither the speaker nor the addressee. In the MSA/BPN application, 2nd person deixis (e.g. U: “What is your price?”) and 1st person deixis (e.g. S: “I cost €599”) are used during human-computer interaction with anthropomorphized objects, while an example of 3rd person deixis is: U: “What is the price of this/that camera?” S: “The price of this/that camera is €599”.
- **Place deixis:** Place deixis (also known as spatial deixis) is deictic reference to a spatial location relative to the location of the speaker. It can be ‘proximal’ if it refers to a nearby location or ‘distal’ if it refers to a distant location, and it can also be either ‘bounded’, in which case it indicates a spatial region with a clearly defined boundary, or ‘unbounded’, in which case it indicates a spatial region without a clearly defined boundary. During map navigation and exploration in the MSA/BPN application, a user can for example say “Take me from here to there” while pointing to two distal spatial map locations on the mobile device’s display. Place deixis would also suit in a shopping scenario where users might ask “What is the name of the camera to the left of the PowerShot S50?”
- **Pars-pro-toto deixis:** Pars-pro-toto is defined as being a “part (taken) for the whole” (Merriam-Webster, 1998). This form of deixis commonly occurs in visual-spatial domains when a user refers to part of a referent but actually intends to be referring to the whole of a referent. This concept is defined in (Wahlster, 1991) with respect to the use of pointing gestures to select visual entities on an electronic form. In this case, pars-pro-toto is said to occur when the demonstratum (i.e. the region at which the user points) is geometrically embedded within the referent, and an extreme case is defined to occur when a user points at an arbitrary part (‘pars’ in Latin) of the form intending to refer to the form as a whole (‘pro toto’ in Latin). In the MSA/BPN, pars-pro-toto can be seen to occur during shelf synchronization, where a user might for example point at a particular product on the shelf, but really be referring to the entire shelf of products and the desire to synchronize his or her mobile device with this shelf.
- **Ellipsis:** Ellipsis is not a type of reference as much as it is the lack of a reference that should otherwise exist. Ellipsis occurs when a dialogue construction lacks an element that is recoverable or inferable from the context. Two different types of ellipsis are substitution ellipsis and expansion ellipsis. In the MSA/BPN application, substitution ellipsis occurs when a user first speaks the utterance “What is the price of this camera?” and then speaks the utterance “What is the optical zoom?”. In effect, what the user really means to say in the second spoken utterance is “What is the optical zoom [of this camera]?”. The missing information is in this case recoverable through the resolution of extra-linguistic information that would normally accompany the user’s utterance. Expansion ellipsis in comparison occurs when a user builds upon a dialogue construction over consecutive utterances, and a theoretical example of this in the context of the MSA would be “Do you have a camera with

5 megapixels?”, and then “and a 3x optical zoom?”. In this case, what the user really means to say in the second spoken utterance is “[Do you have a camera with 5 megapixels] and a 3x optical zoom?”

The above defined referring modes all represent different ways in which a referent can be referred to. Although there are many ways to reference a single referent, the referent itself is always a single entity. *Referent*, as defined in (Allen, 1995) is the concrete object or concept that is designated by a word or expression, and can for example be an object, action, state, relationship, or attribute. *Referential terms* in comparison represent the linguistic constructs such as nouns, pronouns, and demonstratives that linguistically represent the referents. From the four main classes of words - nouns, adjectives, verbs, and adverbs - nouns (e.g. ‘PowerShot S50’) and noun phrases (e.g. ‘the cheapest camera’) are two that are commonly used to represent referents in the MSA/BPN. Defined in (Loos, 2003), a noun is the name of a person, place, or thing and a noun phrase is a phrase whose head is a noun or pronoun and optionally accompanied by a set of modifiers. Noun classes used in the MSA/BPN include pronouns (e.g. ‘it’, as in “how many megapixels does it have?”), proper nouns (e.g. ‘PowerShot S50’, as in “What is the price of the PowerShot S50?”), and common nouns like count nouns (e.g. ‘cameras’, as in “I’d like to interact with the product set *cameras*”). Another linguistic class commonly used for referencing is that of specifiers. Specifier types include ordinals (e.g. ‘first’, ‘second’), cardinals (e.g. ‘one’, ‘two’), and determiners like demonstratives (e.g. ‘this’, ‘that’). Demonstratives deictically indicate a referent’s spatial, temporal, or discourse location. In the MSA/BPN, an example use of a demonstrative would be “What is the price of this?” where ‘this’ really stands for ‘this camera’. An example use of ordinals would be during product comparisons in the MSA where a user might ask for two products to be described and then ask the follow up question “What is the optical zoom of the first camera?”. The set of referents used in the MSA/BPN include not only objects like ‘PowerShot S50’ but also object features like ‘price’ and ‘megapixels’, which may be referenced through the use of extra-linguistic interactions like pointing at words on the visual-WCIS scroll bar displayed in the bottom section of the mobile device’s display.

2.3.2 Multimodal Discourse Phenomena

In (Wahlster, 2003), a list of multimodal discourse phenomena extending the aspects defined above is outlined. This list is of particular importance to multimodal systems because it defines a minimum set of requirements that state-of-the-art multimodal systems should be able to cater for with regards to multimodal interaction. The list of phenomena is discussed below in relation to the MSA/BPN application. Although the listed phenomena specifically targets both resolution and generation (which is a requirement for entirely symmetric multimodal systems), for the purpose of simplicity, examples will mostly be given only with respect to reference resolution rather than reference generation.

- **Mutual disambiguation of modalities:** Mutual disambiguation refers to the ability of a system to recover individual unimodal input streams from a temporally overlapped multimodal signal, in parallel, and for the purpose of creating a single modality-free semantic interpretation. In the MSA/BPN for example, a user’s multimodal speech-handwriting interaction is processed in parallel by one or more speech- and one or more handwriting-recognizers, first to recover the semantics contained in each modality, and then to fuse this information to form a single modality-free interpretation.

- **Multimodal deixis resolution and generation:** Deixis resolution refers to the interpretation of references in an expression that (usually) point to extra-linguistic context. In the MSA/BPN, spatial deixis commonly occurs when for example during a spoken dialogue utterance, the user refers to an object that is on the PDA's display or that is situated in the physical world around them, by any of a number of gestures including pointing, picking an object up, or putting an object down: <S="What is the price of this camera?"><G="PowerShot S50">. Deixis generation occurs in the MSA/BPN when for example in response to the user's utterance the system uses speech synthesis (S) to present the product's feature and value information (i.e. "The price of this camera is €599.") and uses a spotlight as a form of extra-gesture (G) to select the camera located on the shelf.
- **Crossmodal reference resolution and generation:** Crossmodal reference resolution refers to the interpretation of references in an expression that occur dispersed over multiple modalities such that the underlying semantics of the referent are only determinable on interpretation of some or all of the collaborating modalities. The example illustrated under the previous dialogue phenomena is one instance of crossmodal resolution and generation as the referent is partly defined by both the speech and gesture modalities. A further crossmodal reference example occurs when a user looking at four objects on the PDA's display, three of which are mobile phones and only one of which is a camera (see figure 5.3 for a similar example), enquires about the price of the single camera: "What is the price of the camera?". In this case, the system must resolve the reference by interpreting information contained in both the auditive and visual modalities.
- **Multimodal anaphora resolution and generation:** Anaphora resolution refers to the interpretation of references in an expression that point to previously mentioned referents. In the MSA/BPN, anaphora occurs when the user for example speaks out the following utterances over two separate user-turns: <S="Find me the PowerShot S50"> and then <S="What is its price?">. In this case, the word 'its' refers back to the 'PowerShot S50' that was identified in the previous interaction.
- **Multimodal ellipsis resolution and generation:** Ellipsis resolution refers to the interpretation of an expression in which a particular element is missing but nonetheless inferable from the context. For example, within a shopping context the user utterance "What is the price?", might be interpreted as referring to one or more product references, depending on the semantic interpretation of a previous utterance ("What is the price of this camera?", "What is the price of these cameras?").
- **Multimodal turn-taking and back-channelling:** Multimodal turn-taking refers to the issue of which communicative partner (e.g. human or computer) is the next to interact during a multimodal dialogue. Turn-taking is often required by multimodal dialogue systems as a means to collect additional information needed before a particular task can be performed. Such information is obtained by the user and the computer populating information slots in predefined communication acts, either all at the same time or over multiple turns. This was a focal point in the Smart Shopping Assistant (SSA) (Schneider, 2003), which is closely related to the MSA/BPN application. In the SSA, probabilistic relational models were used for object-oriented plan recognition, with the goal of identifying a user's plan and then helping to fulfil this plan. In the MSA/BPN shopping scenario, multimodal turn-taking

might entail a user and the system working together to find an ideal camera, based on a set of predefined constraints like price, megapixels, and optical zoom.

Back-channelling refers to the use of control signals (e.g. head nods and verbal confirmations like ‘yes’) to indicate that the intended message was correctly perceived by the communicative partner. In the MSA/BPN, back-channelling occurs when a user indicates to the system (via button presses) that an interaction was correctly recognized (see section 5.3.3.1).

2.4 Mobile Users and Instrumented Environments

This section establishes the setting for mobile users and instrumented environments, and discusses the application contexts covered in this dissertation. Section 2.4.1 starts with an outline of the progression that has been seen in recent times from the more traditional desktop computing scenarios to modern mobile computing scenarios. The section looks at application contexts of the past, the changes that have since occurred, challenges that still need to be overcome in the domain of mobile applications, and what the future still has in store. Section 2.4.2 follows this up with a description of the mobile scenarios that form the basis of this dissertation, namely outdoor and indoor pedestrian navigation, and interactive shopping.

2.4.1 Progression from Stationary Computing to Mobile Computing

In the past, the majority of interface design and system design has centred on a stationary desktop computing paradigm. The desktop computer was designed to be used by a single person at a time and is suitable for general purpose tasks such as word processing, programming, sending messages or digital documents to other computers on the network (i.e. email), multimedia editing, game play, and Web browsing. The ‘desktop metaphor’ that is still used for interacting with personal computers today was created in the 1970s by a group of researchers from Xerox and was designed to allow for the technical details of the computer to be concealed by a friendly and familiar working environment (Müller-Prove, 2002). A closely related interaction concept is that of WIMP (Windows, Icon, Menu, Pointer) (Edwards, 1988), which is used to sum up interaction conducted using the contained elements. The target user group for desktop computing was originally taken to be the average office worker and the main human-input-devices used for interacting in such contexts were the keyboard and mouse.

Modern day application contexts are no longer limited to office scenarios, and this is indeed seen in that the term ‘personal computer’ is now no longer used to encompass only the desktop computer, but rather also devices such as laptops, tablet PCs, PDAs, and wearable computers. As Weiser (1991) points out, computing trends are such that many people used to share a single mainframe computer, which then led to the use of a single personal computer per user, and finally the use of many computers per user (ubiquitous computing). In comparison to the desktop computer, the newer more mobile devices all offer reduced size and increased mobility, albeit some more than others (e.g. see figure 6.1). The list of mobile computing devices around today further includes devices like mobile phones, music players, cameras, and Bluetooth GPS receivers, all of which have seen large market penetration in recent years.

2.4.1.1 Mobile Device Limitations

Mobile application contexts do however differ significantly from traditional desktop application contexts, particularly with respect to users, the environment, and mobile devices. Such applications also quite often have restrictions that would render WIMP interaction inadequate. ‘Users’ may for example be standing or moving (but not sitting at a desk in front of a keyboard) and thus limiting the type of I/O devices that may be used for interaction. Constantly changing parameters of the ‘environment’ in which interaction takes place might also contribute adversely to user interaction (e.g. noise, crowds, the time of day, the weather), thus limiting the type of communication that may be used in a given environment context. ‘Mobile devices’ also differ from their desktop counterparts in that they are generally more restricted (e.g. display size, processing power, and memory). As a result, applications designed for mobile contexts like navigation and sightseeing, shopping, restaurant finders, and museum guides are only now beginning to emerge in the market place, and much of the research and usability that was conducted previously for stationary desktop computers (e.g. interface design) is still lacking for mobile contexts.

The hardware and software limitations of mobile devices is a point that requires special attention as these limitations specifically restrict aspects in the design and implementation of the shopping and navigation applications that are described in this dissertation. The importance of mobile system design and indeed also of usability field studies was outlined in a talk by Anthony Jameson at the Kloster Irsee Conference in 2002 (Jameson, 2002), in which it was said that only realistic tests conducted early enough in the design process can reliably prevent mobile systems being designed without consideration for the actual conditions under which they are used. Indeed, in (Kumar et al., 2004), the study of a prototype mobile multimodal system used under exerted conditions identified technical limitations that exist with current wireless technologies. The study reported that the wireless 802.11b receivers used in the PDAs could not properly maintain a network connection while the subjects were running, and this led to substantial time delays and poor multimodal performance. Singh, Jain, and Singh (1999) further support the validity of concern with regards to wireless communication: “mobile commerce systems have so far been largely rejected by consumers, even by those who were initially eager to try them out”. The reason is cited to be technical limitations in current wireless technology that can lead to long waits and frequent interruptions of connections.

Bohnenberger (2005) states that one remedy for such device limitations is to “aim for designs that work well with the current limited technology, checking with users to see whether they really do work well enough”. Two requirements that were adhered to during the creation of the MSA and the BPN applications outlined in this dissertation were that the applications had to function in real-time and stand-alone. In (Asthana, Cravatts, & Krzyzanowski, 1994) the design of a personal shopping assistant system based on mobile end-devices was said to be dictated by the following device constraints: size, weight, power consumption and frequency bandwidth. The hardware and software factors that influenced the design of the MSA and the BPN applications are briefly outlined below:

- **Display size:** Typically 320x240 pixels or 640x480 pixels on current state-of-the-art PDAs. The display size had an affect on the design of presentation output and also input interaction. For example, due to the limited display space, it was not always possible for all of a textual output to be presented on the display at the one instance in time, instead requiring the use of scrolling text and supporting audio output. For user handwriting to be effective, the entire display space was allocated to the user, in effect meaning that the user would write over

what was currently being displayed on the screen (e.g. shopping products or navigation maps). As shown in usability studies conducted on the MSA, writing on top of images often took users some time to get acquainted with (see section 6.2.5).

- **Processing power:** Typically between 205MHz (e.g. Compaq iPAQ 3600) and 624MHz for current state-of-the art PDAs (e.g. Dell Axim x51v). Processing power affected the map size and the number and complexity of 3D buildings that could be presented in the BPN using VRML (Virtual Reality Markup Language), with the result that many buildings were only represented as 2D floor plan graphics and with a limited number of edges.
- **Size and speed of memory:** Typically between 64 and 196MB. The memory of a PDA is used not only by the installed applications, but also by the operating system, thus affecting the overall amount available to third party applications. In current devices, two separate types of memory are used: flash ROM memory and SDRAM. Removable SD and CF cards⁵ can also be used, but these are noticeably slower when used for applications and large software components such as speech recognizers and concatenative synthesizers. Due to memory restrictions, the MSA and BPN could not both be demonstrated on earlier PDA versions. The number and size of the XML files that could be read by the applications was also limited due to memory constraints.
- **Connectivity to networks (Always Best Connected):** PDAs are typically equipped with infrared, Bluetooth, and wireless LAN technology. However, these technologies require supporting infrastructure like network access points, which are not always available in contexts that are representative of mobile scenarios (e.g. outdoor pedestrian navigation and shopping). Connection to mobile phone networks (e.g. GSM/GPRS and UMTS/HSDPA, over a Bluetooth connection to a mobile phone) is possible, but not currently affordable for casual users, although flat rates are beginning to emerge. It is due to this lack of wide-spread connectivity and affordability that the BPN and MSA were designed as stand-alone applications. The architecture of the applications does however support an Always Best Connected (ABC) methodology, meaning that communication can take place over the most appropriate communication channels. This is seen in the BPN for example in that map data can be downloaded via either a USB, infrared, or Bluetooth connection with a computer, or directly via the GSM/UMTS phone network. User positioning was also possible via either GPS, infrared beacons, and/or active RFID tags. Interaction recognizers for the MSA and the BPN were configurable to function in either a distributed or an embedded configuration, meaning that if a network connection was available and the communication between the PDA and the server was fast enough to deliver recognition results in real-time, user input could be evaluated by a more powerful remote server rather than locally on the mobile device.
- **Suitable I/O devices:** Designing for mobile contexts removes the ability for an on-the-go user to communicate via traditional I/O devices like keyboard and mouse. These are replaced on the mobile device by interfaces like the touch-pad and microphone. A variety of sensors like accelerometers, magnetic compass, and 3-axis attitude sensor arrays (pitch, roll, yaw) further allow for input into the system, as do devices like cameras and laser scanners that are capable of interpreting Augmented Reality (AR) tags and bar codes respectively.

⁵SD: Secure Digital, CF: Compact Flash

Mobile applications also often make use of I/O devices situated in the environment, for example public microphone arrays and passive RFID tag readers.

- **Operating system, available software, and libraries:** Due to the above mentioned mobile device limitations - primarily speed and memory - software for PDAs is compiled differently to software for desktop computers, and this renders much of the existing software non-usable for mobile devices. Even platform-independent programming languages like Java are limited by the interpreters and the programming packages available for mobile devices, and when software packages designed for desktop computers are ported to PDA devices, they often come with a lesser degree of encompassed functionality.

Further restrictions that originate from the above mentioned mobile device limitations took the following form:

- **Multimodal knowledge representation and user input modelling:** Due to speed and memory restrictions on current PDA devices, the parsing of XML documents and their size was limited to only the most essential tasks like database queries and retrievals, and the communication of recognizer grammars (including speech and handwriting) and various system events. Data Type Definitions (DTDs) for XML documents were not supported and larger XML documents (for example coded in RDFS) took too long to process and reduced the robustness of the developed systems. This had the effect that some standards including the W3C EMMA Working Draft for representing multimodal input were replaced by simplified solutions.
- **Word hypothesis graphs:** No embedded speech recognizers are known by the author to date to return word hypothesis graphs for recognized spoken input. This is due to the embedded speech recognizers often being streamlined to be more efficient with respect to processing and memory requirements. This has the effect that properties such as timestamp and confidence value can not be retrieved on a per word basis, but rather only on an utterance basis from which the expected temporal order of words and an overall utterance confidence value can be derived. This limitation is common in other mobile demonstrators (e.g. (Kumar et al., 2004)), but could be overcome by incorporating server-side recognizers, although not even all server-sided speech recognizers provide such functionality in their returned results.
- **Timing information:** Functionality in the underlying operating system only provides access to timing information that is exact to the second. It is however often desirable to have timestamp information that is exact to the millisecond, for example in identifying the time that individual semantic constituents in an utterance (e.g. demonstrative pronouns like ‘this’ and ‘that’) are provided by a user in a particular modality. Timing in milliseconds for the MSA and BPN applications was achieved on the PDA through the use of system timers, but was found to be too resource intensive for the CPU when used for long periods of time.

Despite the difficulties arising from device restrictions and dynamically changing environment conditions, users are gradually breaking free from the traditional stationary desktop computing paradigm and entering the realms of mobile, ubiquitous, and pervasive computing. This is confirmed by a recent report by the market analysis company Gartner⁶, which shows that 816.6 million mobile terminals - such as mobile phones and smart phones - were sold globally in 2005,

⁶Gartner is an analysis company with a focus on the global information technology industry, <http://www.gartner.com>

in contrast to only an estimated 219 million PCs in the same period (Milanesi et al., 2006). The worldwide PDA market was also stated to have reached a record 14.9 million units shipped in 2005 (up 19% from 2004), from which 7.05 million PDAs were based on the Microsoft Windows CE operating system (up 33% from 2004) and 2.96 million were based on the Palm OS (down 34% from 2004) (Kort, Cozza, Maita, & Tay, 2006).

Supporting the take up of mobile devices is the technological improvements that are occurring in the field. For example, the hardware performance for PDAs is steadily increasing, with a three-fold increase in CPU speed and memory size over the last 4 years. Display sizes have doubled in resolution over this same period. The underlying operating systems are also becoming more robust (e.g. in terms of memory management), as too are commercially available third party software packages including speech and handwriting engines as used by the MSA and the BPN applications. Furthermore, with the emergence of mobile devices that combine Pocket PC+Phone functionality (e.g. MDA Pro⁷) and the emergence of UMTS phone networks, higher bandwidth 3G communications are becoming a reality.

2.4.1.2 Current Trends for Mobile Computing

While stationary desktop computing scenarios concentrate on software applications like document processing and are based on the WIMP metaphor for human-computer interaction, the market for mobile computing is directed not only at application services that support the user while in the office (e.g. time, calendar, contacts, and calculator aids), but also - and more importantly - for services that support a user in environments outside the office.

One project supporting a basic infrastructure for such mobile services is that of FLAME⁸, in which a multi-channel service platform is being designed to allow third party providers to post their Web services to a single services platform, from which end users can then subscribe and retrieve relevant mobile services on their device (Holtkamp, Gartmann, & Han, 2003). The services described in FLAME are defined in an ontology that provides semantic descriptions of each integrated Web service. The focus of the work is on the personalization of services for individual users, based on aspects like user preference, location, and time of day. A wide range of services in FLAME and its successor project COMPASS 2008⁹ have been defined to be relevant for mobile users and in particular for tourists destined for the upcoming Beijing Olympic Games in 2008. The depicted service descriptions concentrate on domains like sightseeing, dining, shopping, and transport, and include for example tour guides for historic and cultural sites and museums; point of interest locators for commercial buildings like restaurants (e.g. Chinese, German, Japanese) and shops (e.g. electronics, handcrafts); aids for selecting a meal from the menu within a restaurant or a product within a shop; services that provide information on train timetabling, train routing, route calculation, taxi fare estimation, traffic congestion, and weather forecasts; language translation services (e.g. German -> Chinese, and vice versa); how-to guides (e.g. for learning traditions and customs of a nationality), friend-location/buddy support; and emergency support (phone numbers, hospital locations).

The range of services described above brings together a broad spectrum of applications, but these need to be well integrated within a single end device for them to be beneficial to an end user. This places a great deal of importance on being able to sort and easily locate different services on

⁷T-Mobile, <http://www.t-mobile.de/shop/>

⁸FLAME: http://www.isst.fraunhofer.de/deutsch/inhalt/Projektarchiv/2005/FLAME_2008/

⁹COMPASS 2008: <http://compass.dfki.de>

the mobile device. Services can be sorted for example based on user preference (i.e. only those relating to a service category like sightseeing), but also on time (e.g. food services may be more relevant at times like 12:00 and 18:00) and on location (e.g. a user is often likely to be more interested in services that are in their vicinity and in the same town in comparison to those that are further away). The MSA and the BPN (as described in the next section) are well integrated in this respect as a user can be navigating a path within the BPN, while at the same time interacting with different objects on the map and in the real-world (such as landmarks and shops). Dependent on the type of object (e.g. a shop), the user may then load up the MSA application to provide assistance in selecting an appropriate shopping product.

The means in which one can access information is also a vital factor for mobile devices. For example, an interface with a deep menu structure will not be as convenient for a mobile user as that of an interface that makes intuitive use of natural languages like speech and handwriting. Natural language interaction and more specifically multimodal interaction is another area in which the MSA and the BPN excel. In these applications, a user is provided a high level of flexibility in choosing a modality or modality combination in which to communicate with the system. This is particularly important for mobile scenarios as environment contexts are destined to change and some modalities like speech are better suited when the user is in motion, compared for example to handwriting, which is a better modality to use if a high amount of background noise is present. Other factors that have an influence on modality include speed, convenience, privacy, and recognition accuracy (Wasinger & Krüger, 2005).

Another important trend that is expected to have wide market penetration is the area of tangible user interfaces. *Tangible User Interfaces* (TUIs) give physical form to digital information, employing physical artefacts both as representations and as controls for computational media (Ullmer & Ishii, 2001). In essence they couple physical representations (e.g. spatially manipulable physical objects) with digital representations (e.g. graphics and audio), yielding interactive systems that are computationally mediated. In the MSA for example, an intuitive one-to-one mapping between physical shopping items on the shelf and elements of digital information on the mobile device's display is employed (Wasinger & Wahlster, 2006). In this fashion, when coupled with user interactions, the situative context of a shopping item (e.g. in or out of a shelf) is used to compute the meaning of an interaction (i.e. product actively selected). TUIs find their origin from the term *direct manipulation*, which Shneiderman (1992) described in 1983 as encompassing "the visibility of the objects and actions of interest; rapid, reversible, incremental actions; and the replacement of complex command-language syntax by direct manipulation of the object of interest". The example the author uses to describe the use of the term in a real-world environment is that of driving an automobile: "To turn left, the driver simply rotates the steering wheel to the left. The response is immediate and the scene changes, providing feedback to refine the turn". Computing examples that relate to the era in which the book was written were more simplistic and incorporated the use of a track-ball or mouse to select entities on a display. As outlined in (Krüger et al., 2004), the design of modern tangible user interfaces also needs to consider that users may be interacting with computationally empowered artefacts that provide no obvious clue on their computational abilities and thus need to be simple, intuitive, and easy-to-learn.

In the following section, two mobile scenarios based on the BPN and the MSA applications are described. In these scenarios, a user is able to navigate a route consisting of both indoor and outdoor paths, and then interact with products in a shop. An important feature of the applications is that the user can interact with objects both on the mobile device's display and in their surrounding environment, while navigating and exploring a map as well as while shopping. This interaction

takes place with objects that are relevant to the individual contexts, such as buildings, rooms, shelves, and shopping products.

2.4.2 MSA/BPN System Descriptions and Scenarios

This section describes two interlinked scenarios upon which the dissertation builds. The first scenario is based on the BMW Personal Navigator (BPN, see figure 2.6A), which supports a user during indoor and outdoor map navigation and exploration (Krüger et al., 2004; Wasinger, Stahl, & Krüger, 2003a). The second scenario is based on the Mobile ShopAssist demonstrator (MSA, see figure 2.6B), which supports a user in retrieving product information through product attribute and product comparison queries within a shopping context (Wasinger, Krüger, & Jacobs, 2005; Wasinger & Wahlster, 2006).



Figure 2.6: The two mobile demonstrators created under the scope of this dissertation: A) The BPN and B) the MSA.

In mobile contexts like pedestrian navigation and shopping, the user is often on-the-go and thus interacting in a constantly changing environment. Users might at one moment be located in front of an electronics shop, while at the next moment find themselves inside a restaurant or inside the grounds of a park. The variability of the environment is also not helped by the fact that pedestrians (unlike cars) tend not to stick to streets, but rather will often take short-cuts from a clearly defined path (e.g. when walking through a park) (Stahl & Hauptert, 2006). Actions like crossing a busy road can also place the user in a very different and potentially adverse context within a short instance in time, thus requiring that communication with the user through different modalities take into consideration the surrounding context. As such, a user interface needs to adapt dynamically. One good medium for communication with the user is the environment itself, presented either as the physical real-world that exists in the world around the user (in its tangible form), or as digital entities represented on the mobile device's display. Within the BPN and the MSA, users can interact through a range of different modalities in order to access information on

entities represented on the mobile device's display and also in the physical world around them. In this way, the mode of communication can be adapted by a user to the surrounding environment. A user that is on-the-go might for example speak and point to objects in the environment rather than write the object names on the mobile device's display (which would cause the user to slow down or stop entirely to avoid tripping over potential obstacles). In a different context, a user desiring privacy might be more likely to write than speak, and when close to an object might (especially when shopping) prefer to touch or pickup the object rather than select it from the mobile device's display.

One aspect that ties the BPN and MSA applications together is the interconnection between navigation and interaction, as shown in figure 2.7. This relationship is best demonstrated when considering how a user combines 'pedestrian navigation' with 'interaction' in a larger scenario such as shopping in a foreign city (Wasinger & Krüger, 2004). In such a scenario, pedestrian navigation may take place both outdoors and indoors. In an outdoor context, navigation will be based on street and footpath networks that are used to guide a user through a 'city environment'. During such navigation, interaction may focus on the querying of surrounding building objects like a shopping mall. In an indoor context, navigation will be based on corridor networks that are used to guide a user through a 'building environment', during which interaction may focus on the surrounding rooms such as individual shops within the shopping mall. Within a particular 'room environment', shopping isles will be used to guide a user, and interaction may focus on surrounding containers such as tables, shelves, and shopping trolleys. A 'container' (e.g. a shelf) will be subsequently navigated based on its individual shelf levels, and interaction at this point takes place with 'items' such as digital cameras. At any point in time, the current objective (either navigation or interaction) may shift in importance, for example at this final level, interaction with the digital camera will take the primary objective, although this still consists of navigating the different product attributes such as price and megapixels. The combined use of the BPN and the MSA systems provide for indoor and outdoor pedestrian navigation as well as interaction when exploring a navigation map and when shopping for a product. This navigation and interaction takes place in everyday contexts that users can easily relate to, like that of a city, building, room, container, and individual product items.

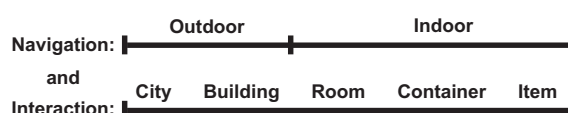


Figure 2.7: Environment contexts used during interaction with the BPN and the MSA.

2.4.2.1 BMW Personal Navigator Scenario

The BMW Personal Navigator (BPN) was developed under the projects COLLATE¹⁰, REAL, and READY¹¹, together with the BMW research division in Munich¹². It is an entirely implemented system that combines a desktop event and route planner, a car navigation system, and a multimodal

¹⁰COLLATE: <http://collate.dfki.de>

¹¹REAL and READY: http://w5.cs.uni-sb.de/website_old/bair/

¹²BMW Group Forschung und Technik: http://www.bmwgroup.com/bmwgroup_prod/e/0_0_www_bmwgroup_com/forschung_entwicklung/forschung_entwicklung.html

indoor and outdoor pedestrian navigation system for PDAs (Krüger et al., 2004). The communication modes used in the system and particularly for the pedestrian navigation component, include speech and selection-gesture (in the form of pointing and line drawing), which can be used during map navigation and map exploration. These modes can be used either unimodally or multimodally, and support interaction with the real-world in addition to on-device interaction with the display. In this section, the entirely implemented system is described, including discussion on the underlying motivation for creating such a navigation system, and the individual components of the system including the mobile pedestrian navigation component that is most relevant for this dissertation.

Traditional navigation systems are designed to work for a specific platform in a well defined environment, for example Deep Map (Kray, Elting, Laakso, & Coors, 2003), which is only targeted at pedestrians, and Telmaris (Malaka & Zipf, 2000), which does not cater for in-door navigation. Navigation systems can be categorized into several classes. One typical class is that of Web-based route finders. These services are optimized for the PC and usually provide only little support for other devices like PDAs, relying instead on the directions being printed on paper. Another prominent class are car navigation systems, which can be divided into two subclasses. Whereas the subclass of built-in navigation systems, often shipped with cars are restricted to supporting the user while driving, PDA based navigation systems (e.g. the TomTom navigator¹³) can also be used outside the car. The advantage of in-built systems is a higher positioning accuracy and very good usability under driving conditions due to a larger display and specialized input methods. The BMW IDrive Controller (see figure 2.10D) is one example of a specialized input method that is used for in-car menu navigation and has also recently been tested for multimodal use (together with speech) in accessing music (Becker et al., 2006). PDA based systems are more flexible, but rely on the same maps for in-car and on-foot conditions, causing suboptimal results when providing route descriptions for pedestrians.

The solution provided by this work is a situated personalized navigation service that transparently combines the desktop PC at home, a built-in car navigation system, and a PDA. Figure 2.8 shows the three different types of situations in which navigational services were considered to be of interest. At home, the desktop PC is used to make all travel arrangements provided by a personal navigation server that can be accessed over the Internet. The travel itinerary is then synchronized with the PDA, which allows access to the travel itinerary without the need for a direct Internet connection. In the car, the PDA connects locally to the car navigation system, which in turn permits the travel itinerary to be transferred from the PDA to the car navigation system. During the navigation task in the car, the PDA remains predominantly silent and the car navigation system takes control in guiding the traveller to the selected destination. Before leaving the car, the PDA receives the actual parking spot coordinates, which are added to the travel itinerary and which may help to find the way back to the car later on. On foot, the PDA plays a much more vital role. It displays the 3D map information included in the travel itinerary and guides the traveller with verbal and graphical route directions. It can also be used to store geo-referenced user data (e.g. voice memos) and to respond to multimodal requests regarding landmarks in the environment (e.g. “What is this building?”). If required, the PDA can also connect to a navigation server and receive updated information (e.g. on path directions), which is important for users that have lost their way.

To better understand the functionality of this system, and in particular the mobile pedestrian navigation component, consider the following scenario:

¹³TomTom, <http://www.tomtom.com>

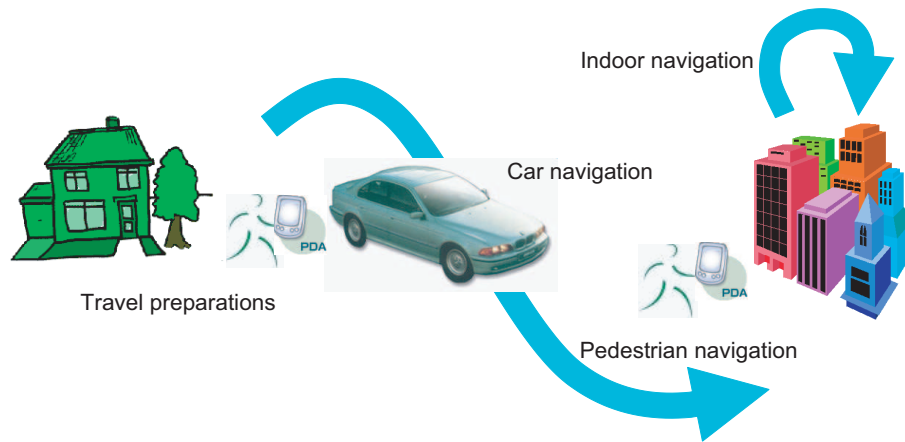


Figure 2.8: Different situations during a navigational task.



Figure 2.9: Downloading a travel itinerary onto the PDA: A) the desktop interface and B), C) the navigation routes and route information as they appear on the PDA. Note also the ‘walking’ and ‘car’ icons which denote whether a particular route is for pedestrians or for automobiles.

At home: Mr. S from Saarbrücken plans to drive to Munich for the weekend for a spot of shopping and some sightseeing. He has a particular shopping mall in mind in which one can buy digital cameras at discounted prices. He starts to prepare his trip while sitting in front of his desktop computer at home. First, he logs onto the personal navigation Web server, which provides him with information on earlier trips and travel itineraries. After entering the destination coordinates, the server provides Mr. S with a travel itinerary (figure 2.9A) that is then downloaded to his PDA and contains weather information, route navigation instructions, as well as sights of interest and suggestions on where to park in Munich (figures 2.9B, and 2.9C). Assuming that Mr. S will use the suggested parking station, the system also returns outdoor path directions from the parking station to his final destination, and even indoor directions that lead Mr. S directly to the specific shop that he intends to visit.

In car: When entering the car, the PDA is used to program the in-car navigation system at the push of a button (via a Bluetooth connection with the car server), making the cumbersome input of address details unnecessary. The car navigation now takes control and guides Mr. S safely to the parking station in Munich (figure 2.10).

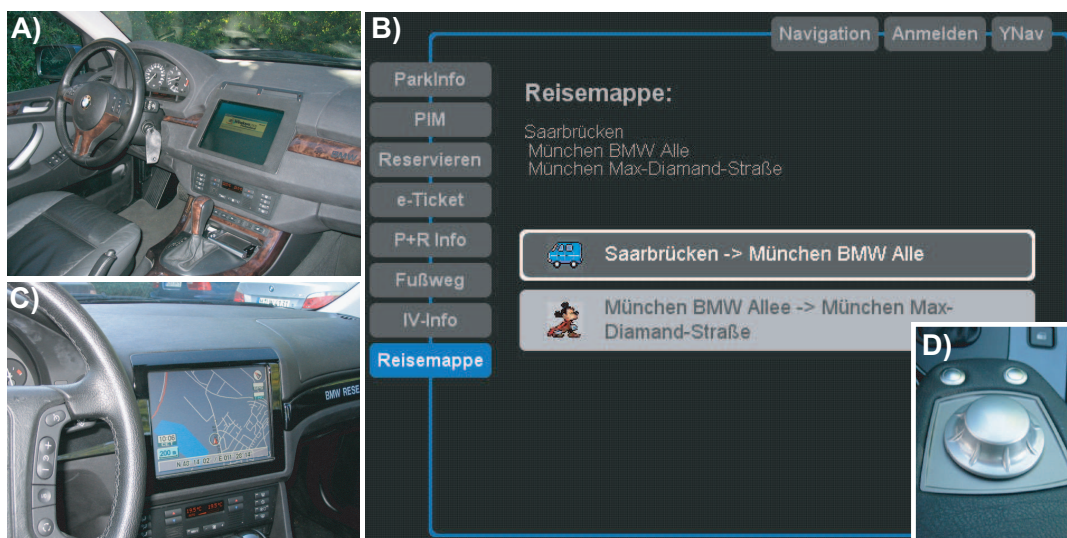


Figure 2.10: BMW Connected Drive showing A) the integration of the car display, B) the prototype navigation system with the uploaded PDA contents, C) car navigation in progress, and D) the BMW iDrive controller used for in-car interaction with the BMW navigation system.

On foot: After having parked at the designated location, Mr. S uses the PDA to navigate the remaining path to the shopping mall on foot. The PDA uses speech and 3D graphics to guide Mr. S and to provide information on surrounding landmarks and street names. Figure 2.11A shows three speech utterances that are provided by the system at different stages along a street, in which a longer start utterance is provided followed by a medium lengthed utterance and a final shorter utterance just before he needs to make a turn. Figure 2.11 also illustrates the different 2D and 3D perspectives that can be achieved in the demonstrator by zooming and by switching from a birds-eye view (figure 2.11A) to an egocentric view (figure 2.11D).

Along the way to the shopping mall, the BPN informs Mr. S of points of interest that are within a given radius of his current location, including an ingenious looking art sculpture that is located to his right: “Located on your right is the Richard-Serra sculpture”. Captivated by the sight of this sculpture Mr. S decides to find out a little bit more about its history before continuing on his way to the shops.

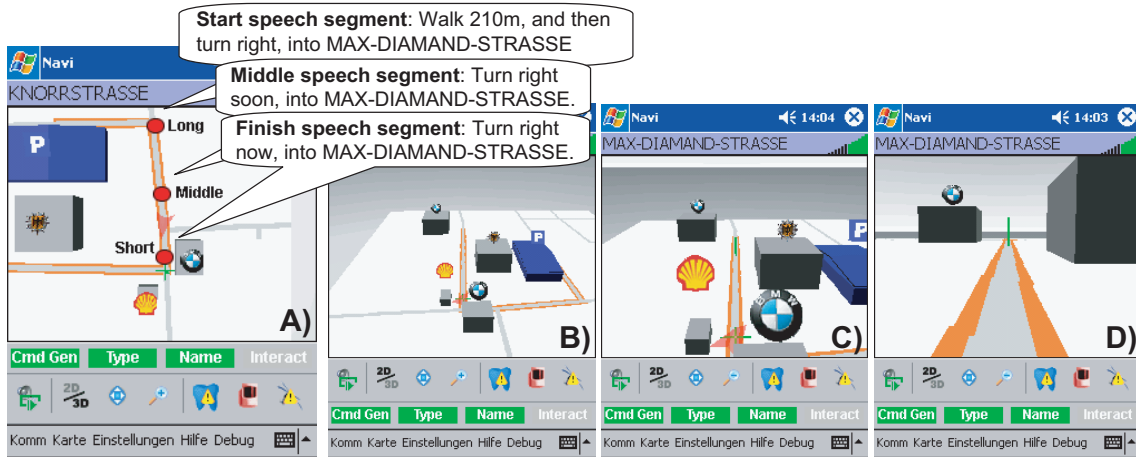


Figure 2.11: BPN screen-shots demonstrating speech output (A) and different 2D/3D visual perspectives ranging from birds-eye (A) to egocentric (D). Figures B, C, and D also demonstrate the BPN’s zoom feature.

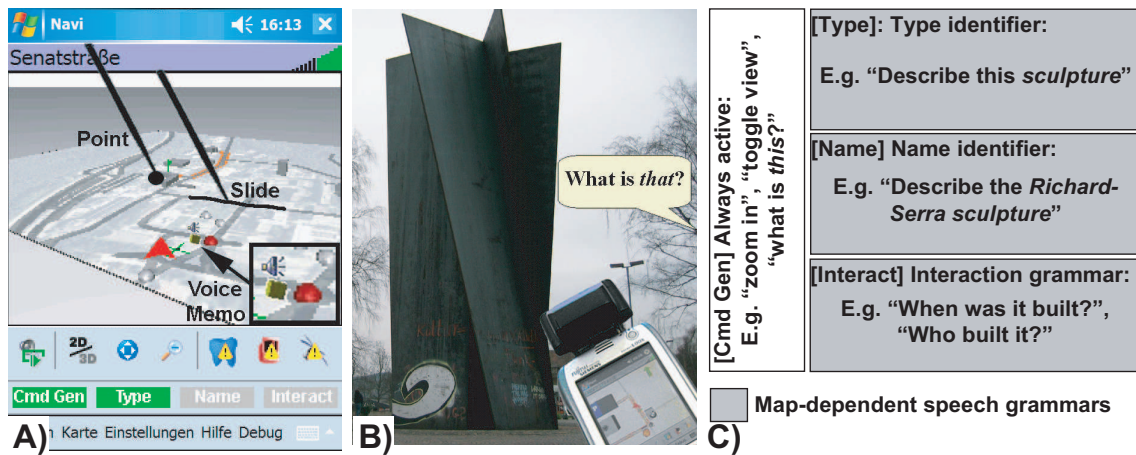


Figure 2.12: Interaction with the BPN system, illustrating A) point and slide intra-gestures, B) a combined speech-gesture(extra) interaction, and C) the speech grammars: ‘Cmd Gen’, ‘Type’, ‘Name’, and ‘Interact’. Figure A) also illustrates the use of voice memos (lower right) and the activation of the speech grammars ‘Cmd Gen’ and ‘Type’ (just below the map region).

It can be noted at this point that the BPN system caters for speech and speech-gesture combined interaction, whereby gesture can be further classified as either intra-gesture or extra-gesture. Intra-gesture refers to on-device interaction in the form of ‘point’ and ‘slide’ actions used to select map entities on the mobile device’s display (see figure 2.12A), while extra-gesture refers to off-device

interaction in the form of ‘point’ actions used to select objects in the real-world (see figure 2.12B). Furthermore, the speech grammars used by the system allow for a user to address map entities by referring to the object either directly via pronouns (e.g. demonstrative pronouns like ‘this’ and ‘that’), or via the object’s type (e.g. ‘sculpture’) or name (i.e. ‘Richard-Serra sculpture’), as shown in figure 2.12C. Based on the map’s size, the associated map grammars may contain anywhere between a few dozen words to many hundreds of words, but this number of active words can be minimized to improve recognition accuracy by selecting only a subset of the ‘name’, ‘type’, and/or ‘pronoun’ grammars to be active.

Switching off the ‘name’ speech grammar (just below the map region in figure 2.12A), Mr. S uses the communication modes of speech and real-world pointing gesture to enquire about the sculpture, “What is that?”, and accompanies this by a pointing gesture in which the PDA is pointed in the direction of the sculpture (figure 2.12B). With the object now visible on the display, Mr. S continues to interact with the object by asking “When was it built?” and “Who was the sculpture built by?”. These queries are made possible by landmark interaction grammars, which are loaded once a particular object has been selected. Impressed by the sculpture, Mr. S records a short geo-referenced voice memo onto the PDA to remind him of his experience later on (bottom right insert in figure 2.12A).

Keen to get to the shops, he continues on his walk until he arrives at the shopping mall, where he then enters the building and continues to use his PDA indoors in the same way that he had outdoors (location positioning indoors is based on an active RFID and infrared positioning system that is installed in the building, see (Brandherm & Schwartz, 2005)). First he goes up a flight of stairs and then he navigates the mall to reach the electronics shop where he hopes to find a new digital camera. Figure 2.13 shows the indoor maps that Mr. S sees along his path. Note also the ShopAssist icon located at his destination, which is a direct link to the MSA application.



Figure 2.13: Two typical indoor navigation images showing the user’s path to the electronics shop.

2.4.2.2 Mobile ShopAssist Scenario

The Mobile ShopAssist (MSA) application was developed under the BMBF funded project COL-LATE¹⁴ for the purpose of presenting state-of-the-art Language Technology (LT) products to the public and for demonstrating a wide range of different mobile and multimodal interaction possibilities. A subsidiary goal of the development was to deliver a multimodal testbed that can be used to test the effectiveness of multimodal interaction under different contexts. The primary context of the MSA is that of shopping, during which a user is able to interact multimodally with a set of items like ‘NLT’ and ‘digital camera’ products, to compare them and query their features (Wasinger et al., 2005). The MSA supports the user throughout the buying process, including finding a product, querying and comparing products, and finally purchasing the product. Like the BPN, it is designed for mobile PDA devices and all of the processing (e.g. speech and handwriting recognition) is performed locally on the device itself. This means that a user need only connect once to a supporting shop server (and its product database) in order to download a relevant dataset of products. The MSA is multilingual (English, German), and interaction with the system is derived from the base modalities speech, handwriting, and selection-gesture, whereby selection-gesture can be categorized as intra-gesture (i.e. on-device interaction) or extra-gesture (i.e. off-device interaction). In the MSA, intra-gesture refers to the selection of referents on the mobile device’s display via a ‘pointing’ action, and extra-gesture refers to the selection of objects in the real-world via ‘point’, ‘pickup’, and ‘putdown’ actions. Tangible real-world interaction is another commonality between the MSA and the BPN systems. As described in detail in section 4.2, among the types of multimodal interaction that the MSA demonstrator supports, are those that are temporally overlapped (i.e. different modalities overlapped in time) and those that are semantically overlapped (i.e. different modalities overlapped with respect to semantic content). The mobile and multimodal demonstrator also supports ‘anthropomorphized objects’ and ‘direct’ and ‘indirect’ product interaction (see section 4.3).

One motivation for shops to provide rich interaction types such as extra-gesture is that this supports the concept of ‘interaction shopping’. In contrast to ‘window shopping’, where a user is generally limited to viewing products locked behind a glass window, interaction shopping permits a user to physically handle and query the objects. Based on observations that were carried out during the usability studies outlined in chapter 6, tangible product queries are thought to provide users with a greater sense of certainty that a product will in fact live up to their expectations. Being able to physically handle a product is also thought to positively affect the quality of a shopping experience (e.g. shopping becomes fun), when compared to not being able to handle a product, such as is the case during window shopping and Internet shopping. Even when a user is unable to touch an object, multimodal interaction can still provide benefits, for example consider shopping on a Sunday when most shops are closed or even during the week outside of business hours. Rich interaction through modalities like speech, handwriting, and intra-gesture will in this case still permit a user the flexibility to enquire about a product and to purchase the product electronically.

Although research on interfaces for ubiquitous computing (Dey, Ljungstrand, & Schmidt, 2001) and for the domain of shopping (see section 3.2.1) is now beginning to develop, the combination of these areas with that of multimodal interaction ((Oviatt, 2003), and section 3.1) is still novel. This combination of a shopping context scenario with that of mobile and ubiquitous computing and multimodal interaction is however the precise focus of the MSA. Two different scenarios are catered for in the MSA application. The first scenario supports the use of a shop-

¹⁴COLLATE: <http://collate.dfki.de>

ping trolley in a grocery shop and has a central focus on plan recognition strategies (Schneider, 2003). The second scenario supports the use of a PDA within a shop (e.g. an electronics shop) and focuses on mobile multimodal interaction. It is the second scenario that this dissertation is based on. To better understand the functionality of the interactive shopping component, consider the following scenario which follows from the previously described mobile pedestrian navigation scenario.

In shop: Attracted by all of the sales he can see through the shop's display window, Mr. S enters the shop and navigates his way to a real-world shelf of products containing, among other products, digital cameras. He clicks the MSA icon that is displayed on the PDA display and waits for the application to load. He then connects to the shelf (each shelf's unique ID is provided by an infrared ID beacon¹⁵, see figure 2.18B on top of the shelf) and selects the product type that he is interested in, in this case 'digital cameras' (see figure 2.14). This downloads all relevant product information for each of the digital camera products on the shelf, including pictures and interaction grammars. The data set also includes products of a particular type that are not currently available on the shelf such as products that are out-of-stock or only available online. Even products that are not sold by the shop may be downloaded onto the PDA for comparison purposes, assuming that these products are represented in a compatible XML format.



Figure 2.14: Using the MSA to select from a number of different product types situated on a shelf.

Figures 2.15A, 2.15B, and 2.15C show three different product views available to Mr. S in addition to the real-world products that he can see in front of him. Each view demonstrates a different trade-off between graphical and textual information, with the 9x view showing 9 products but no text, while the 4x view showing fewer products but each accompanied with a name, and the 1x view showing only a single product but accompanied with several of the most relevant attributes. Also visible in the lower part of the figures is the menu that contains access to specialized application functionality, and the graphical toolbar buttons, from left to right: connect (to shelf),

¹⁵Eyed infrared ID beacon, <http://www.eyeled.de>

synchronize (with shelf), browse products, compare products, previous page, and next page. Being truly multimodal, the functionality in this toolbar is also available in the other modalities.

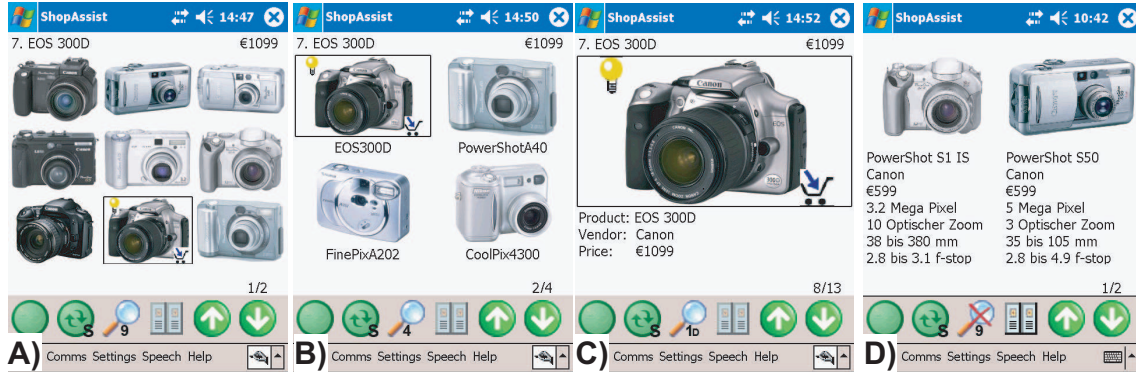


Figure 2.15: Different product views in the MSA showing A) a 9x view, B) a 4x view, C) a 1x view including description, and D) a product comparison view. The graphical toolbar (above the application menu) illustrates the following buttons from left to right: connect to shelf, synchronize with shelf, browse products, compare products, previous page, and next page.

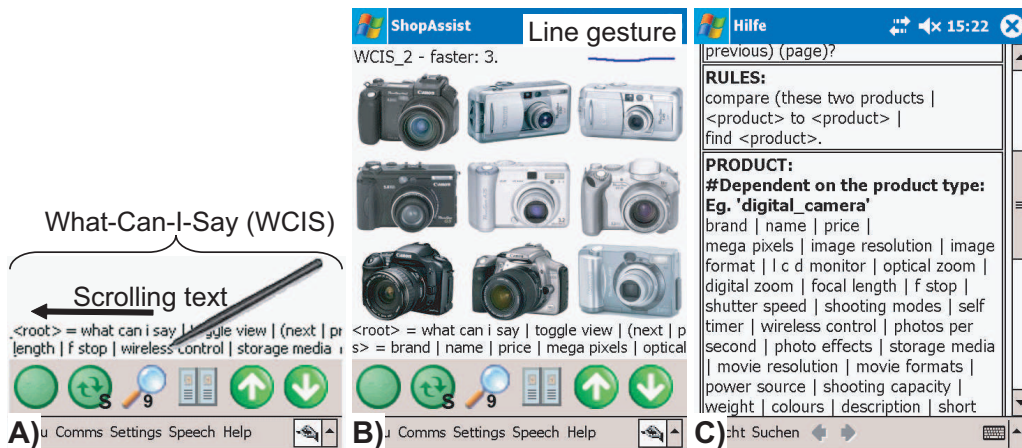


Figure 2.16: MSA functionality as defined in the interaction grammars for digital cameras. The figures show: A) the visual What-Can-I-Say scroll bar, B) how to increase or decrease the speed of the WCIS text through a line gesture (top right), and C) the WCIS information represented as a HTML page.

After synchronizing with the shelf, Mr. S begins to interact with the products. Not knowing what he can say, he first speaks the utterance “What Can I Say?”, which instantly loads the visual What-Can-I-Say (WCIS) scrolling text bar at the bottom of the display as shown in figure 2.16A. He then increases the speed at which the text scrolls, by sliding the stylus across the display as shown in figure 2.16B. Alternatively, Mr. S could have configured the system to provide him with the WCIS grammar in the modality of speech, or as a HTML page (see figure 2.16C), but decided not to do this because speech can be a slow modality and the HTML page hides the camera objects that would otherwise be visible on the PDA’s display.

Mr. S is keen to interact with the physical products that he can see on the shelf. An intuitive one-to-one mapping exists between physical shopping items and their digitally-represented counterparts. As a result, Mr. S is able to interact with the ‘digital’ set of products visible on the PDA’s display, but also with the ‘physical’ set of products available on the shelf and with a combination of ‘physical and digital’ products making up the data space. Figure 2.17 demonstrates the latter case in which Mr. S compares two products using the modalities speech, intra-gesture (GI), and extra-gesture (GE): “Compare this camera <GE> to this one <GI>”. In reply to the request, the system provides the visual output shown in figure 2.15D, where two cameras are displayed side-by-side with their most relevant attributes listed below. This visual output is further accompanied by a spoken summary of the cameras, the speed of which can be increased, decreased, or stopped. At any point in time during Mr. S’s visit to the store, he may decide to disconnect from the shop server and continue browsing entirely offline. By doing this, the range of interactions available to Mr. S is limited (i.e. the recognition of real-world pickup and putdown actions will not be possible), but the level of user privacy is increased because the shop will no longer be able to identify Mr. S interacting with the products via the PDA. Much of the current work on RFID enabled technologies focus on the benefits for retailers through improved inventory management and tracking, but the MSA scenario attempts to balance this out for the shopper by letting them not just choose whether or not to have their interactions logged, but by also providing incentives for this, such as real-world tangible interaction to retrieve product information, comparison shopping (as demonstrated in figures 2.17 and 2.15D), and the generation of cross-selling recommendations like camera accessories.



Figure 2.17: MSA interaction illustrating the use of speech, intra-gesture, and extra-gesture, during a product comparison query.

Speech being Mr. S’s favourite modality, he interacts with the system as follows: “How many megapixels does the EOS 300D have?”. The system replies: “EOS 300D. Megapixels. 6.3 megapixels” (see figure 2.18B). Under a different environment context, Mr. S may have preferred to use a different modality or modality combination for interacting with the system, for example handwriting combined with intra-gesture as shown in figure 2.18A, which would be more appropriate in contexts where a high level of background noise is present, or where user privacy is desired. A point to note is that Mr. S is just a single user and his modality preferences can not be

realistically taken to represent all user groups such as children and the elderly, men and women, and people with disabilities. Fortunately, the multimodal interactions described in this scenario are just the tip of the iceberg with respect to the multimodal combinations that are possible when interacting with the MSA system, the depth of which will be discussed in detail in chapter 4.

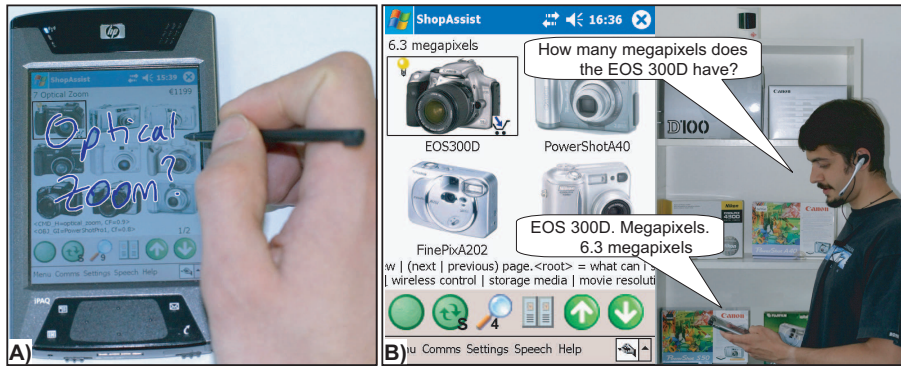


Figure 2.18: MSA interaction illustrating A) handwriting+intra-gesture and B) speech (for both input and output).

After interacting some more with the MSA and narrowing down the cameras that suit him best, Mr. S notices a large public display situated next to the shelf and uses the PDA to load onto this public display a URL of the manufacturer's complete set of specifications for the camera he is most interested in, as shown in figure 2.19A. Deciding that the EOS 300D is indeed the best camera, Mr. S adds this to his electronic shopping basket (figure 2.19B) and continues to explore the store, during which time a cross-selling service (outside the scope of this dissertation) is activated to present Mr. S with information on accessories for the selected camera, such as lenses and a camera case. The ability to combine the MSA application with other services and devices in the instrumented environment demonstrate the flexibility and future potential of such a mobile system. Indeed, instead of adding the camera product to his electronic shopping basket on the PDA, Mr. S might instead have taken one of the store's RFID-instrumented shopping trolleys and thus simply placed the product inside the trolley (Schneider, 2004), as shown in figure 2.19C.



Figure 2.19: MSA ties to the instrumented environment, showing A) a public display, B) the on-device shopping basket to which Mr. S added the product (see inset), and C) an RFID-instrumented shopping trolley that can be combined with the MSA (Schneider, 2004).

2.4.3 MSA/BPN System Architecture

The goal of this section is to provide a brief overview of the architecture of the MSA/BPN system. This architecture and the accompanying figure can be taken as a visual index into the many components that are later discussed in the dissertation.

As outlined in section 2.4.2, the MSA and the BPN are two interlinked demonstrators designed for mobile handheld devices, namely PDAs. The demonstrators are compiled for the Microsoft Windows Mobile platform¹⁶. This platform is the most common platform for PDAs and encompasses the following types of device: Pocket PC, Pocket PC+Phone, and Smartphone. Some of the devices on which the MSA/BPN was tested include the Siemens Pocket LOOX¹⁷, the HP iPAQ (as well as the earlier Compaq iPAQ)¹⁸, and the Dell Axim¹⁹. With the exception of the HP range of iPAQs, all of these devices represent Pocket PC devices rather than Pocket PC+Phone or Smartphones.

Figure 2.20 shows the architecture of the MSA/BPN as well as the data and program flow between the various components. The main components that are shown include the PDA and server components, the devices used for input interaction and output presentation, and several external applications that can be interfaced from the MSA/BPN.

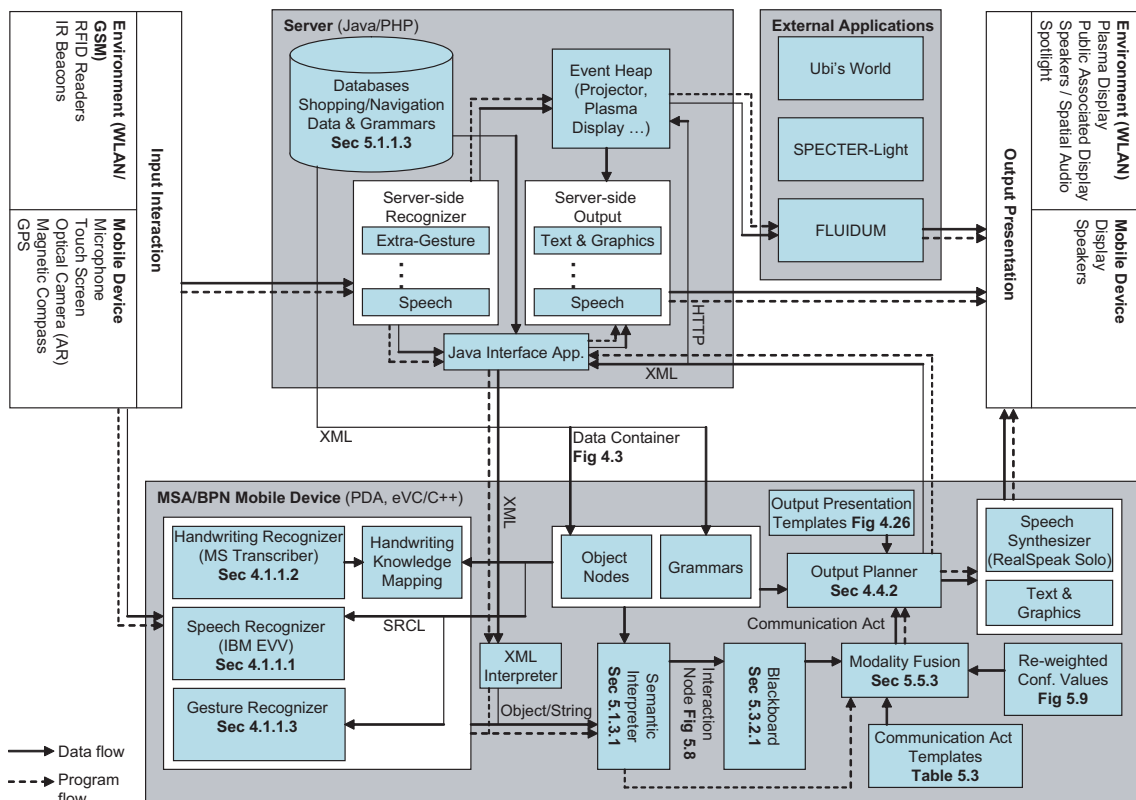


Figure 2.20: The MSA/BPN architecture showing the data flow between components.

¹⁶Windows Mobile, <http://www.microsoft.com/windowsmobile>

¹⁷Siemens Pocket LOOX, <http://www.fujitsu-siemens.com/products/mobile/handhelds>

¹⁸HP iPAQ, <http://www.hp.com.au/ipaq/>

¹⁹Dell Axim, http://www.dell.com/content/products/productdetails.aspx/axim_x51v

The mobile device component is the primary focus in this dissertation, as this is where the MSA and BPN applications reside. These applications are programmed in embedded Visual C/C++. The figure illustrates the relevant modules that are located on the PDA, including the embedded recognizers for speech, handwriting, and gesture. IBM Embedded ViaVoice was used for speech recognition, and Microsoft Transcriber was used for handprint recognition. The handprint recognizer can also be seen to be augmented by a self-implemented module that maps a string of recognized characters to known values in the handwriting grammars. The gesture recognizer is also part of the MSA/BPN implementation and maps (x,y) coordinates from the PDA's display onto possible referents. See section 4.1.1 for information on the recognizers used in the MSA/BPN.

The recognizers take as input a user's speech, handwriting, and/or gesture utterance, and provide as output a string or object interpretation of this utterance. The recognition process is also reliant on the grammars that are loaded at runtime and stored as linked-lists during program use. The recognition results are then sent to a semantic interpreter that is responsible for searching these results for semantic constituents belonging to a modality-free language (as defined in the communication act templates, see table 5.3). These constituents are then written as individual interaction nodes onto the blackboard (see figure 5.8). The semantic interpreter is also responsible for triggering the modality fusion module, which occurs each time a query+feature or command is identified to exist. The modality fusion module (see section 5.3.2) has the responsibility to fuse the events on the blackboard to form a complete communication act. This process entails several important steps, including determining an appropriate user interaction timeframe in which to consider collected input, re-weighting recognizer confidence values to avoid individual recognizer biases, the selection of an appropriate communication act, blackboard event filtering, and the fusion of semantic elements based on the selected communication act (see section 5.2 and section 5.3).

Interpreted communication acts are then passed to the output planner (see section 4.4), which has access to the object node and grammar data that is stored on the PDA. This data is used to determine answers to user queries and also provides formatting instructions (e.g. feature/object/value output in comparison to just feature/object output, see section 4.4.2). The output planner also has access to a variety of output planning templates that define which communication modes the system should use when interacting with the user, for example speech synthesis (ScanSoft RealSpeak Solo is used for speech synthesis) and visual graphics, as well as whether such information should be presented privately to the user (i.e. on the PDA device) or through the use of public infrastructure (e.g. a large plasma display or public associated displays).

A second component shown in the MSA/BPN architecture is the server. The server contains the MySQL²⁰ databases that contain navigation (street and landmark) and shopping (product) information, as well as information required to generate grammars for the embedded recognizers located on the PDA device. This information is downloaded onto the PDA in XML format when, for example, a navigation route is selected in the navigation scenario or when the user synchronizes with a store shelf in the shopping scenario. Another component running on the server is that of the server-sided recognizers. One server-sided recognizer used in the MSA/BPN scenario is that of extra-gesture, which recognizes the RFID tag of a product and whether the action is a pickup or putdown event. A Java interface application then maps the recognized RFID tag to a particular product and encodes the information into XML, which is in turn transmitted to the PDA.

²⁰MySQL, <http://www.mysql.com>

This application is similarly used to interface the server-sided speech synthesizer (ScanSoft Real-Speak Solo) and could in the future be used to interface additional input and output modules used for input recognition and output presentation (e.g. a server-sided speech recognizer). In addition to the Java interface application, an event heap (developed at Stanford University (Johanson & Fox, 2002)) is also located on the server. In comparison to the Java interface application, where communication is based on a client-server design, the event heap provides an indirect interaction mechanism that offers a high degree of fault tolerance and is based on a commonly accessible tuplespace design (Stahl, Baus, Brandherm, Schmitz, & Schwartz, 2005). This event heap is used to interface devices like the RFID-instrumented shelves, the spotlight, and the public associated displays (see section 4.5.3). Communication with the event heap is conducted via HTTP Web requests. Also located on the server is the Apache Web server, which serves PHP scripts to the public plasma display each time user input is detected by the server-sided extra-gesture recognizer. FLUIDUM²¹ (which contains the spotlight and public associated display functionality, see section 4.5.3), SPECTER-Light (Schneider, Kröner, & Wasinger, 2006), and UbisWorld (Heckmann, 2005) are several additional applications that can be connected to the MSA/BPN application.

²¹FLUIDUM, <http://www.fluidum.org>

Computing environments in which the user is mobile have seen significant advancements in recent years as applications begin to span multiple and changing contexts. Multimodal interaction has also emerged as an integral area of development for these contexts, especially as users break free from the stationary desktop computing paradigm and enter the realms of mobile computing, in which more flexible, more natural, and more robust interaction is required. This chapter outlines relevant work in the area of multimodal interaction, by summarizing significant and novel projects that have contributed to state-of-the-art research in the field. The chapter also outlines relevant work in the mobile computing domains of shopping and navigation, two mobile contexts that stand to gain from the incorporation of multimodal interfaces and that apply to the majority of people on a daily or near-daily basis.

3.1 Projects with a Focus on Multimodal Interaction

This section outlines state-of-the-art research that has been and still is being conducted in the area of multimodal interaction over recent years. The goals of this section are to outline the leading research organizations behind the drive and the projects that have resulted from their research. The section analyses several of the larger projects in some depth. In particular, the goals and highlights of each project are summarized and comparisons are drawn between these projects and the work outlined in this dissertation. This section ends with a table outlining the primary differences between each of the systems.

The systems described below all form part of an elite class of state-of-the-art multimodal dialogue systems. Many of the application domains that are focused on in the systems can be categorized into a select few groups. The most prominent domain to be seen in almost all of these multimodal projects is that of ‘map-based interaction’. This is a focus point in the projects QUICKSET, MATCH, MUST, EMBASSI, SMARTKOM, COMPASS, and the MSA/BPN (Mobile ShopAssist / BMW Personal Navigator). Typical implementations of map-based interaction concentrate on navigation and city exploration (e.g. BPN, MATCH), tourist guides (e.g. MUST), and smaller add-on services for the domain of tourism, such as ticket reservation, weather, subway information, and restaurant guides (e.g. COMPASS, MATCH). Other prominent application domains include home entertainment, public information kiosks, and car infotainment (all of which were prominent in the SMARTKOM and EMBASSI projects). Some of the more recent projects are also focussed on new and more difficult domains with regards to multimodal interaction. COMIC for example focused on bathroom design, MIAMM on interaction with music databases, and the MSA

on real-world shopping. Only two of the projects (both of which are among the newest) have a focus on real-world interaction, MSA/BPN and RASA, and this might give some indication into the future direction that mobile and multimodal systems are likely to take. The combination of different application domains is another area that shows future potential, the likes of which can be seen in projects such as COMPASS and MATCH (which combine navigation with restaurant finders) and the MSA/BPN (which combines navigation and exploration with a shopping assistant).

A notable feature of these projects is that they are all research projects rather than final commercializations, and this shows the very young nature of this field of study. With the exception of the pioneering PUT-THAT-THERE system published in 1980, all projects have been active within the relatively short timeframe of the last 5 to 10 years (and indeed the vast majority of the projects are less than 5 years old). Furthermore, many of the projects are still being developed in one form or another, including the larger systems like QUICKSET (which has during the course of its life produced a number of extensions such as RASA), and SMARTKOM (where multimodal components are being ported to an open-domain question-answering system in the project SMARTWEB).

Another notable feature of these projects is that the majority of the work has been conducted in either Europe or the USA. Some of the European research organizations behind the outlined projects include for example, the DFKI (German Research Center for Artificial Intelligence), Loria (Lorraine Laboratory of IT Research and its Applications), Eurescom (European Institute for Research and Strategic Studies in Telecommunications), the EML (European Media Laboratory), and the ITC-irst (Center for Scientific and Technological Research). From the USA, some research organisations behind the outlined projects include the CHCC/OGI (Center for Human Computer Communication, Oregon Graduate Institute of Science and Technology) and the MIT (Massachusetts Institute of Technology). This list of research centres is far from exhaustive and also does not include all of the university and industrial partners that have contributed to research on multimodal systems in the projects outlined in this section.

3.1.1 Put-That-There

Richard Bolt and Chris Schmandt (Bolt, 1980) from MIT describe a system which leverages the joint use of voice-input and gesture-recognition to command events on a large graphics display. The PUT-THAT-THERE system falls under their research into a Spatial Data Management System (SDMS) (Bolt, 1979), in which the goal was to spatially index data that exists within everyday experiences, like sitting at an office desk. The application scenario is situated in a so-called ‘media room’ and consists of a centrally situated armchair, two displays (one on either side of the chair), and a third large projected display some distance in front of the chair. One of the side displays shows the entirety of information in the SDMS including a ‘you-are-here’ rectangle. The sub-portion of this rectangle is then portrayed with increased detail on the second side display, which effectively represents a magnifying window. Two joysticks (also on either side of the armchair) are used to navigate the two displays. One joystick is used to move around the coordinate axis of the first display, while the other is used on the second display to zoom in on information represented by multimedia items such as maps, electronic books, and videos as found in the virtual room. Mini-computers are used to drive the displays and other devices resident in the media room, while the walls are fitted with two sets of loudspeaker banks, one on either side of the projected display and the other on either side of the user’s armchair.

Interaction within the PUT-THAT-THERE system is limited to commanding simple shapes about the central graphics display. Shape types include circles, squares, and diamonds, all of

which can be created, deleted, moved about, replicated, and their attributes (colour and size) altered. The idea is that these shapes act as placeholders that represent physical real-world items like an inbox or a calendar. The multimodal capabilities of the system are such that a user can interact via speech-only, or via speech-gesture (pointing) combined interaction. In the case of speech-gesture interaction, the pointing gesture always accompanies speech input and is used for the resolution of references occurring in the speech input. The actual fusion techniques are not outlined in the paper. An example of the system's capabilities is demonstrated by the speech utterance "Move the blue triangle to the right of the green square". This utterance can also be expressed multimodally via speech and gesture: "Put that there", where the pronoun 'that' refers to "the blue triangle" and 'there' refers to "to the right of the green square". A highlight of the system is its ability to resolve relational expressions, for example when modifying the attributes of an object (e.g. 'smaller', 'larger') or when moving an object (e.g. 'to the right of').

The underlying hardware required for the described multimodal speech and gesture interaction includes a speech recognizer capable of recognizing a maximum of 120 words and an orientation sensor providing pitch, heading, and roll data that is used for pointing at coordinates on the central display. This sensor is worn by the user on the wrist and is accompanied by a second sensor required to calculate the physical position of the sensor while the user is sitting in the armchair.

PUT-THAT-THERE depicts an instrumented environment infrastructure in which the user, although not sitting in front of a desktop computer, still needs to sit in a fixed armchair for the system to function correctly. A final distinction is that processing in this system is carried out on stationary PCs rather than embedded on a mobile device. In contrast to the PUT-THAT-THERE system in which pointing-gesture input is used solely to augment speech input, the work outlined in this dissertation focuses on providing the same degree of input flexibility in each available modality (made possible through the use of a blackboard architecture and multimodal fusion techniques). The MSA and the BPN also differ from the system in that they do not limit a user's physical mobility.

3.1.2 XTRA

Another pioneering system that incorporates multiple communication modes is XTRA (eXpert TRAnslator) (Kobsa et al., 1986; Wahlster, 1991), which combines natural language and deixis input. XTRA provides written natural language access to expert systems, and had the aim of rendering interaction with expert systems easier for inexperienced users. XTRA allows a user to combine natural language input (German) together with pointing gestures on a computer display, in order to refer to objects on the display. Input can be in the form of written linguistic descriptions (typed on a keyboard), pointing gestures with a pointing device (i.e. a mouse), or both.

XTRA has been developed independently to any single expert system. In its first application, access to an expert system in the income tax domain was realized, in which the system assists a user in filling out his or her annual withholding tax adjustment form. During a dialogue with the user, the system displays the relevant page of the income tax form on the computer display. This form also contains a number of rectangular regions that may themselves contain additional embedded regions, all of which can be referred to by the user. A highlight of the system is its ability to cater for the resolution of pars-pro-toto deixis (see section 2.3.1). 'Pars-pro-toto' deixis occurs in the XTRA system when a user points at an embedded region when actually intending to refer to a superordinated region. Several sources of information are used for identifying relevant regions on the tax form, including (where available) the linguistic descriptor, the location and type

of deictic gesture, intra-sentential context, and the dialogue context. Combined, this information is said to almost always allow for a precise identification of a referent.

Although XTRA sets out to cater for a variety of different pointing gestures, these are simulated by simple mouse clicks on a computer display. Respective of the times, natural language is also represented by keyboard input rather than speech recognition and/or character recognition. Little focus is placed on aspects like the timing and synchronization of multimodal input, and the described tax form implementation is representative of a stationary domain rather than a mobile one.

3.1.3 QuickSet

QUICKSET (Cohen et al., 1997) is a collaborative, multimodal pen and voice system for interacting with a number of distributed applications. Its main role is to provide the multimodal interface to these applications based on components such as recognizer agents, a natural language agent, and a multimodal integration agent. At the heart of the system is its ability to integrate continuous spoken language and continuous pen input (written, symbols, and pointing gesture). The system also focuses on providing the same user input capabilities for different types of devices including handhelds, desktops, and wall-sized terminals. Communication is achieved via wireless LAN through a distributed Open Agent Architecture (Cohen, Cheyer, Wang, & Baeg, 1998). This type of architecture generally allows the processing of input to be carried out remotely on resource-rich devices, with only a minimal amount of processing required on end devices such as handhelds, but as a result does require additional hardware infrastructure to be present for the system to function. The motivation behind QUICKSET was to develop technologies that can aid in substantially reducing the time and effort needed to create large-scale interactive simulations such as those required by a country's military forces.

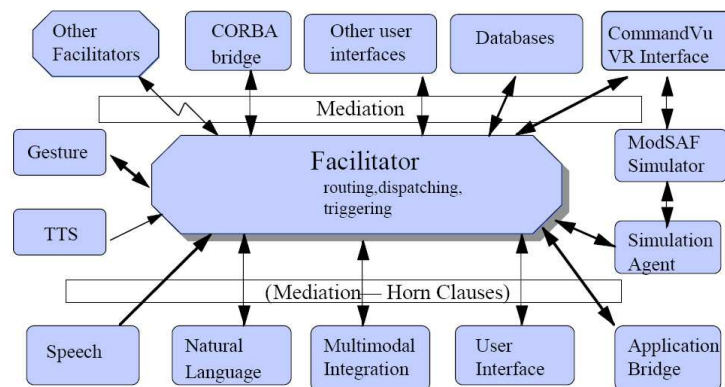


Figure 3.1: QUICKSET system architecture (Cohen et al., 1997).

The QUICKSET interface is specialized to map-based tasks, and this is a commonality in all the resulting applications that make use of QUICKSET. The interface provides a geo-referenced map of some region, and it also provides pan and zoom capabilities, multiple overlays, and icons. Employing pen, voice, or multimodal input, the user can annotate the map, creating points, lines, and areas of various types. The user can also create entities and define their behaviour. One difference between QUICKSET and the MSA/BPN is that QUICKSET uses a tap-to-speak interface

to activate the speech recognizer(s). This is often important in mobile applications as it provides substantially more intelligible results than open-microphone¹ interaction. The MSA/BPN employs a similar interaction metaphor for this reason (i.e. push-to-talk), but rather than requiring a touch-display (and more importantly knowing where on the display to touch), the MSA/BPN utilizes a physical button on the mobile device that the user can both see as well as feel.

Three applications where QUICKSET has been applied include EXINIT, LEATHERNET, and MIMI (Cohen et al., 1997). EXINIT (Exercise Initialization) is an application enabling users to create large-scale military exercise simulations. An example user interaction would be a user providing the following spoken utterance: “Multiple boundaries”, followed in succession by a series of multimodal utterances such as “Battalion <draw line>” and “Company <draw line>”. LEATHERNET is a second military application in which a user can create and position entities, give them missions, and control an associated virtual reality environment. For example, a user might hold the pen at the desired location and speak: “Red T72 platoon” resulting in a new platoon of the specified type being created. The third application is called MIMI (Multimodal Interaction with Medical Information). In contrast to the other two scenarios, this is a non-military application that allows users to find appropriate health care centres in a city and to then ask follow up questions about the centres, including transportation means to those sites. In this scenario, a user might for example say “Show me all psychiatrists in this neighbourhood” and combine this utterance with a circling gesture on the map.

The speech vocabulary of the system consists of noun phrases used to label entities and a variety of imperative constructs used to supply entity behaviour. A novel aspect of QUICKSET is that a user’s speech input is recognized by multiple speech recognizers at the same time - two from IBM (VoiceType Application Factory and VoiceType 3.0) and a third from Microsoft (Whisper). 68 pen gestures are also possible within QUICKSET, and these include various military map symbols (platoon, mortar, fortified line), editing gestures (deletion, grouping), route indications, area indications, and taps. Multimodal input recognized by the individual recognition engines is passed to a natural language agent, which employs a definite clause grammar and produces typed feature structures (Carpenter, 1992) as a representation of the utterance meaning. A multimodal integration agent then analyses the incoming typed feature structures representing individual interpretations of speech and gesture, and it also identifies the best unified interpretation, be that multimodal or unimodal. The authors outline that their modality fusion procedure is based on typed feature structure unification, and that this is advantageous because it supports partiality, mutual compensation, structure sharing, and multimodal discourse. Partiality refers to the ability for a recognizer to deliver part results that can later be unified with other part results. Structure sharing and multimodal discourse refer to the ability for the feature structure templates to be continually modified as more information becomes apparent. Mutual compensation refers to the ability to choose from multiple results in one communication mode based on a given result in a different communication mode (e.g. choosing a point gesture over a line gesture based on what is known from an accompanying speech input²).

QUICKSET is one of the larger projects that focus on multimodal input. Its design is however based entirely around a distributed agent architecture. In contrast, the applications developed under this dissertation focus on embedded multimodal interaction, which has a number of different

¹The term ‘open-microphone’ refers to a microphone that is always actively listening. For the purposes of speech recognition, such an approach requires that speech (and not background noise) be detected automatically

²A trade-off for mutual compensation is an amplified error in the case that the assumed correct input was actually incorrect

advantages, including the ability to function even when the device is disconnected from a network. Another difference is that while QUICKSET concentrates on providing the same functionality to a range of different device platforms, the MSA/BPN concentrates on providing the same functionality to a range of different communication modes, be that speech, writing, gesture, or multimodal. The result of this is that users are able to interact via a wide range of multimodal input combinations, rather than being forced to use a limited set of multimodal combinations specified by the system. A final difference is that the MSA/BPN represents a lower-cost multimodal solution to QUICKSET in that both the application and interaction strategies are embedded on a single end device rather than on a set of distributed servers plus an end device.

3.1.3.1 QuickSet-Rasa

RASA (McGee & Cohen, 2001; Cohen & McGee, 2004) is a tangible, augmented reality environment that digitally enhances the existing paper-based command and control capability in a military command post. It is based on the distributed multi-agent QUICKSET project, and it links physical objects (i.e. Post-itTM notes) on a command post map to their digitally represented military unit counterparts. Input communication channels in this project include speech, pen, and touch. The authors note that in various high stress environments, people choose tools that are robust, malleable, physical, and high in resolution (i.e. pencil and paper), and this has resulted in large paper maps and post-it notes to be preferred over high-value command and control software systems. The motivation of RASA was to support the tools of paper maps, post-it notes, and pencil, while at the same time capturing a digital representation of the map and its entities.



Figure 3.2: Users collaborating with RASA (McGee & Cohen, 2001).

RASA's underlying hardware includes a touch-sensitive Smartboard upon which the user can affix a map, and a digital Anoto^T M pen and tablet PC for writing on the post-it notes. The Anoto pen³ is used for drawing on the post-it notes. Like any other pen, the Anoto pen produces ink, but it also has a Strong ARM CPU, memory, Bluetooth capabilities, and a camera that lets it see a special Anoto dot pattern that can be printed onto any ordinary piece of paper. The underlying

³Anoto pen, <http://www.anoto.com>

tablet PC thus captures the digital ink while the pen simultaneously produces real ink on the post-it notes. The location of the post-it note is then captured by the touch-sensitive display each time a post-it note is affixed to the map.

Gestures recognized by the RASA system include symbolic and editing gestures such as points, lines, arrows, deletion, and grouping, as well as military symbology including unit symbols and various control measures like barbed wire (around 200 symbols in total). Speech is recognized via a microphone array attached to the top of the SmartBoard or alternatively via a wireless close-talking microphone. An off-the-shelf speech recognizer is used for this purpose (Dragon Systems Naturally Speaking), together with context-free grammars containing a vocabulary of 675 words.

3.1.3.2 QuickSet-3D Hand: 3D Hand Gesture Extension to QuickSet

One extension to the QUICKSET system describes how QUICKSET's digital ink capabilities are extended by a 3D hand gesture recognizer (Corradini, Wesson, & Cohen, 2002). The motivation behind this extension was to create a body-centred multimodal architecture employing both speech and 3D hand gestures. A separate goal of the authors was to create a more mobile interaction alternative to the electronic pen used in QUICKSET, which is connected by wire to a specific interface and thus limits user mobility.

In the scenario, a map is projected onto a virtual plane in space, and the user combines 3D hand movements with speech commands to create and move entities on the map. Two types of gesture are recognized by this system: pointing (used to select points on the map) and hand twisting about the index finger (used to signal a user's wish to pan over the map). These hand gestures are recognized through the use of a magnetic field tracker called Flock of Birds (FOB)⁴, which is used to monitor the position and orientation of the hand, and a PinchGlove⁵, which is used to simulate pen-down and pen-up actions. Before using the system, users must first calibrate the regions that they wish to paint in, which is done by pointing at three of the vertices of a chosen rectangle. A limitation of the system as outlined in their work is that although drawings using free hand movements allow for non-proximity and transparency to the interface, the creation of detailed drawings is not easy and human pointing is not very accurate (Corradini & Cohen, 2002). A second limitation of the system is that the hand tracker does still have cables that connect it to a stationary computer.

3.1.3.3 QuickSet-ExertEnv: Mobile System for Exerted Conditions

Perhaps the most closely related work to the MSA/BPN is that from Kumar et al. (2004), in which a mobile multimodal system was built for a study that analysed the relationships of speech, pen-based gesture, and multimodal recognition as a function of a user's state of exertion. In the study, subjects completed multimodal tasks that required them to use speech and gesture to provide the system with two attributes for a first object ('shape' and 'composition') and then two attributes for a second object ('direction' and 'object'). In this system, the shape and direction attributes are only permissible via the modality of gesture, and the composition and object attributes are only permissible via the modality of speech. This meant that each object was in effect identified by information originating from two separate modality types. The hardware used in this system

⁴Flock of Birds, <http://www.ascension-tech.com/>

⁵PinchGlove, <http://www.fakespace.com/>

consists of a PDA communicating with a server via wireless LAN, and a close-talking noise-cancellation microphone. A digital voice recorder and a polar heart rate monitor were also used for study purposes, but did not provide any direct input into the system. ScanSoft's ASR3200 speaker-independent speech recognizer is used in combination with vocabularies catering for 85 different utterances to capture speech input and to generate a 3-best list of hypotheses. A sketch recognizer from Natural Interaction Systems is used for the recognition of five different gestures (dot, cross, arrow, line, and area).

In contrast to the QUICKSET project, speech recognition in this project is performed locally on the PDA device. However, in contrast to the MSA/BPN, the digital ink is still required to be transmitted over the wireless network to a remote server for gesture recognition, and multimodal integration also takes place on the server rather than locally on the PDA. This reliance on supporting infrastructure and in particular the reliance on a working wireless LAN connection while the subject is running, is outlined by the author as creating substantial time delays of 15-20 seconds, and this led to poor multimodal performance each time the system switched between base stations. The author was only able to overcome this problem by allowing subjects to run without the mobile device in their possession and then handing subjects a mobile device each time they arrived at a particular destination. A second difference between this system and the MSA/BPN is that the system only supports the use of a communication mode in combination with a specific semantic type (e.g. speech used only for composition and object information, and gesture used only for shape information and directional arrows), while the goal of the MSA/BPN is to allow a user any modality (speech, handwriting, gesture) in combination with any semantic term (e.g. feature, object).

3.1.4 MATCH

MATCH (Multimodal Access To City Help) is a multimodal city-guide and navigation system that enables mobile users to access restaurant and subway information for the city of New York (Johnston, Bangalore, Stent, Vasireddy, & Ehlen, 2002; Johnston et al., 2002). Input into the system can be expressed by speech, pen (writing, or symbolic gestures), or multimodally (speech-pen combined interaction). The architecture of MATCH is agent-based and similar in style to the hub-and-spoke architecture used in GALAXY (Goddeau et al., 1994). In contrast to almost all other distributed architectures, MATCH runs stand-alone on a Fujitsu pen computer (a mobile computer closely resembling a tablet PC). Support for multimodal interaction is achieved by a single declarative multimodal context-free grammar that is compiled in the form of a Finite State Automaton (FSA). This FSA defines the paths that make up valid interaction with the system and is tightly bound to a speech-act based multimodal dialogue manager that allows for mixed-initiative multimodal dialogue that can span multiple user-turns.

Speech input is provided to the system via a 'click-to-speak' action, in which a user is required to press a graphical button on the visual display to activate the microphone (similar to that described in QUICKSET). In line with the experience gained while designing the MSA/BPN, this is said to be preferred to an open-microphone, as an always listening recognizer is more susceptible to an increased degree of spurious speech, especially in noisy environments. AT&T's Watson speech engine is used for recognition, the output of which is a lattice of possible word string hypotheses and associated costs. Pen input includes area, point, line, and arrow gestures (10 in total), and a total of 285 handwritten words are accommodated by the system. Gesture and handwriting recognition is based on a variant of Rubine's (Rubine, 1991) template-based gesture

recognition algorithm trained on a corpus of sample gestures.

An example multimodal interaction in the MATCH system is as follows: “Show cheap Italian restaurants in this neighbourhood <gesture>”, which could also be represented entirely via handwriting and gesture (e.g. cheap Italian <circle gesture>), or entirely via speech (e.g. “Show cheap Italian restaurants in Chelsea”). Aside from restaurant information, the user can also ask for subway directions (e.g. “How do I get to this place?”).

MATCH and the MSA/BPN have much in common. However, in comparison to MATCH where the grammars are compiled based on a predefined grammar, the grammars in the MSA are generated dynamically based on the type of objects currently in a particular shelf. The MSA also caters for speech-only, handwriting-only, and multimodal interaction, but can in addition support pointing-gesture-only interaction, and unlike MATCH, the MSA/BPN supports physical interaction in a real-world environment.

3.1.5 MUST

The EURESCOM MUST project (MULTimodal, multilingual information Services for small mobile Terminals) is a multilingual (Norwegian, Portuguese, French, and English) and multimodal tourist guide demonstrator for the city of Paris. It was developed by the research departments of three telecommunication operators and two academic institutes over a period of two years from 2001 (Almeida et al., 2002; Boves & Os, 2002). The objectives of the project were to develop a realistic multimodal and multilingual service to gather an understanding of how such services might integrate with future UMTS networks and to conduct human factor experiments with users to evaluate the worthiness of the multimodal interaction.

Interaction within MUST is user-initiated, but in contrast to the MSA/BPN, MUST focuses almost entirely on multimodal interaction consisting of only combined speech and pen (pointing) interaction. Similar to the MSA/BPN, multimodal interaction within MUST may be provided either sequentially (i.e. first speech then gesture) or simultaneously (i.e. gesture provided during a speech utterance, titled in MUST as ‘tap-while-talk’ mode). Output in MUST takes the form of synthesized speech, text, and graphics. A multilingual Question/Answering system is also incorporated to handle out of domain requests.

The MUST tourist guide is based on a fairly rigid interaction template. First a user is presented with an overview map showing all POIs, such as the ‘Eiffel tower’ and the ‘Arc de Triumph’. The user must then select a POI (via speech, pointing-gesture, or both). Having selected a POI object, the user is then able to interact with the POI via speech, for example by speaking “What is this building?” or “What are the opening hours?”. A second example of multimodal input is the spoken utterance “What restaurants are in this neighbourhood?” accompanied by a pointing-gesture. One novel aspect of the project is that the speech recognizer is always open and listening, which is rare for mobile outdoor systems as they often struggle with decreased recognition performance due to the high levels of noise found in the surrounding environment.

As shown in figure 3.3, MUST is based on a distributed client-server Hub architecture that was originally developed under the GALAXY project (Goddeau et al., 1994). The server is responsible for almost all of the processing, while the client device - a Compaq iPAQ PDA - is used merely as an access portal to the service. This heavy reliance on supporting infrastructure (including WLAN and GSM/UMTS networks) is perhaps the most notable difference between the embedded MSA/BPN applications and this distributed system. Philips SpeechPearl2000 is used for the server-sided speech recognition and ScanSoft RealSpeak (among others) is used for server-sided

speech synthesis. Spoken utterances are transferred to/from the server via a GSM-based phone, while text and pen inputs are transferred from/to the client GUI via a TCP/IP connection with the GUI server. Speech recognizer and GUI inputs are fused by a multimodal server and then passed to the dialogue manager where a system response is created and sent back to the client device. Communication within MUST is represented in an XML based markup language named MxML (MUST XML), which is used to represent most of the multimodal content. The individual modules are written in Java and C/C++.

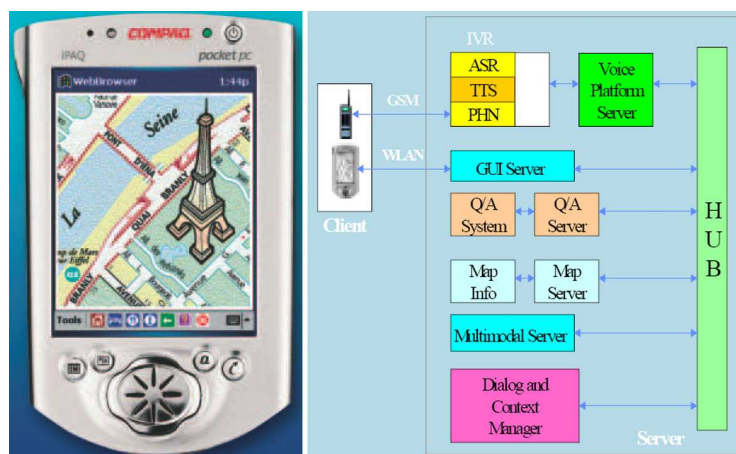


Figure 3.3: MUST mobile tourist guide (left) and its distributed architecture (right) (Almeida et al., 2002).

3.1.6 EMBASSI

EMBASSI (Kirste, Herfet, & Schnaider, 2001; Elting, Rapp, Möhler, & Strube, 2003) was a joint research project funded by the BMBF, with a total of 19 partners from industry and academia. The project stands for “Multimodal Assistance for Infotainment and Service Infrastructures”, whereby the EMBASSI acronym itself is derived from the German translation of the project: “Elektronische Multimediale Bedien- und service ASSIstenz”. The focus of EMBASSI was the development of new paradigms and architectures for man-machine interaction with technical infrastructures for the non-professional everyday life. Central themes of the project include: intelligent assistance, multimodal interaction, and anthropomorphic user interfaces. Around 200 scientific papers and conference talks were generated over the life of the project.

Some application contexts in EMBASSI include home entertainment, public terminal, and car infotainment systems. In the home environment and in particular in a living room scenario, the goals are to allow a user to control devices (e.g. TV) and the environment (e.g. light or sound intensity, and temperature). For example, the user might say: “I want to see the news” or “Please record the thriller tonight”, in a manner similar to the SMARTKOM project. In the automotive environment, the user is able to interact with infotainment devices such as radios, navigation systems, and mobile phones, for which speech input and output were classified as essential modalities (e.g. “I want my favourite station”). The public terminals environment focussed on special user groups like disabled persons because such public environments are most often engineered for the average user and do not allow modification by individuals. Specific interfaces that were considered

included an automatic teller machine and a food vending terminal.

One of the main scenarios is that of the home environment. In this scenario, input is provided via speech, pointing gesture, and a graphical user interface that is controlled by a remote. It is stated that a directed laser beam can also be used in place of video traced pointing gestures, which are often ambiguous. EMBASSI incorporates conventional output modalities like displays, lights, and acoustical signals, but also speech, and non-verbal visual information like facial expressions and gestures through the use of virtual characters. Multimodality is managed in EMBASSI by a 'polymodal input module' that is responsible for merging the modalities. A user can choose to use any of the three supported modalities alone (i.e. unimodal input), or choose to interact with the system multimodally. It is for this reason that the term 'polymodal' is preferred in this project, which is stated to encompass "one or more modalities" rather than only multiple modalities (Kirste et al., 2001). Input modalities available in the other scenarios (all based on a stereoscopic video recording approach) include the recognition of gesture, facial expression, emotion, lip-reading, eye-tracking, and stick-pointing. One novel aspect of the system is that it also caters for indirect interaction. For example in the driving scenario, driver attentiveness is measured based on the direction in which the driver is looking, and driver drowsiness is measured based on the frequency of eyelid movements.

EMBASSI is a multi-agent system based on a layered and distributed architecture. Some of the more prominent modules within EMBASSI include the user input components (also known in EMBASSI as I-components), user output components (or O-components), the dialogue manager, and the context manager. Individual interaction events are first handled by the I-components, and are then transformed into semantic representations that reflect the intention of a user's action. These semantic representations are then fused together; in particular, any references made between the modalities and/or past entities of the dialogue are resolved. Similar to the MSA/BPN, typical user interactions in EMBASSI consist of the combination of a command and object reference (Eltling et al., 2003). In EMBASSI, the semantic definitions and the relationships between semantic types like commands and objects are outlined in an ontology that is defined via an XML DTD (containing the basic hierarchy and syntax) and a corresponding description logic representation (containing the semantics and interdependencies). The modality fusion module for EMBASSI is implemented using Java and Prolog.

One interesting aspect within EMBASSI is with regards to the placement of interaction events and the synchronization of user input. In contrast to the MSA/BPN, which uses a central blackboard to store active events, in EMBASSI, an interaction event is sent directly to the modality fusion component, which then queries all other modality analysers independently to determine if they have information that might contribute to the current interaction based on a set timeframe (t_a, t_b). Additionally, EMBASSI is capable of analysing a single recorded signal through the use of multiple different-type recognizers (e.g. the same eye recording may be used for multiple and different purposes). This aspect shows similarities to SMARTKOM, but differs to the MSA/BPN where a single signal (e.g. speech) is analysed by multiple same-type recognizers with the goal to increase recognition accuracy in the one captured modality.

3.1.7 SmartKom

SMARTKOM (Wahlster, Blocher, & Reithinger, 2001; Wahlster, 2006b) is a mixed-initiative multilingual (German, English) and multimodal dialogue system that combines speech, gesture, and facial expressions for input and output. It was the follow up project to Verbmobil (1993-2000)

(Wahlster, 2000) and was funded by the German Federal Ministry of Education and Research (BMBF). It consisted of over 10 consortium partners led by the DFKI⁶, and over the course of the project 51 patents, 29 spin-off products, and 246 publications were generated.

Three scenario platforms defined under the SMARTKOM project were: SMARTKOM-PUBLIC, SMARTKOM-HOME/OFFICE, and SMARTKOM-MOBILE.

- SMARTKOM-PUBLIC is a public multimodal communication kiosk for domains like airports and train stations. Input modalities consist of speech, facial expression, and gesture. Speech input is captured with a directional microphone and is based on a speaker-independent recognizer. Facial expressions of emotion are captured through real-time video analysis, and gestures are tracked through the use of an infrared camera and an extended version of the Siemens Virtual Touch screen (SIVIT). Output is provided in the form of graphics, audio, and a life-like character called Smartakus that is capable of multimodal interaction in the form of combined graphical output, speech and gesture.
- SMARTKOM-HOME/OFFICE is an infotainment companion that is implemented using a Fujitsu Stylistic 3500X tablet PC and caters for home and office domains. Multimodal services include an electronic programme guide for the TV and an interface for controlling consumer electronic devices like TVs, VCRs, and DVD players. Input interaction takes the form of either just speech (referred to in SMARTKOM as 'lean-back' mode), or coordinated speech and gesture ('lean-forward' mode).
- SMARTKOM-MOBILE is a mobile travel companion that uses a PDA as a front-end device and caters for the domains of car/pedestrian navigation and point-of-interest information retrieval. Multimodal interaction consists of speech input that can be combined with pen-based pointing. A simplified version of the Smartakus interface agent provides output in the form speech, gesture, and facial expression.

SMARTKOM is based on a distributed component architecture called MULTIPLATFORM (Multiple Language Target Integration Platform for Modules, (Herzog, Kirchmann, Merten, Ndiaye, & Poller, 2003)). The SMARTKOM system consists of more than 40 asynchronously running modules coded in four different programming languages: C, C++, Java, and Prolog. An important feature of the architecture is its multi-blackboard design (see figure 3.4). The communication protocol is based on the Multimodal Markup Language (M3L) (Herzog et al., 2004) and is used for exchanging information via the various blackboards and for expressing word hypothesis graphs, gesture hypothesis graphs, hypotheses about facial expressions, media fusion results, and the presentation goals.

The integration and mutual disambiguation of multimodal input and output in SMARTKOM is based on symbolic and statistical methods for the fusion of modalities, which is conducted on both a semantic and a pragmatic level. Unification, overlay, constraint solving, and planning are all used in SMARTKOM's modality fusion and modality fission components. The modality fusion component unifies all scored hypothesis graphs stemming from a user interaction. These graphs are generated by the speech recognizer (word hypothesis graphs), prosody component (clause and sentence boundary hypothesis graphs), gesture recognizer (hypotheses for possible reference objects in the visual context), and facial expression interpreter (hypotheses about the emotional state

⁶German Research Center for Artificial Intelligence (DFKI GmbH), <http://www.dfki.de>

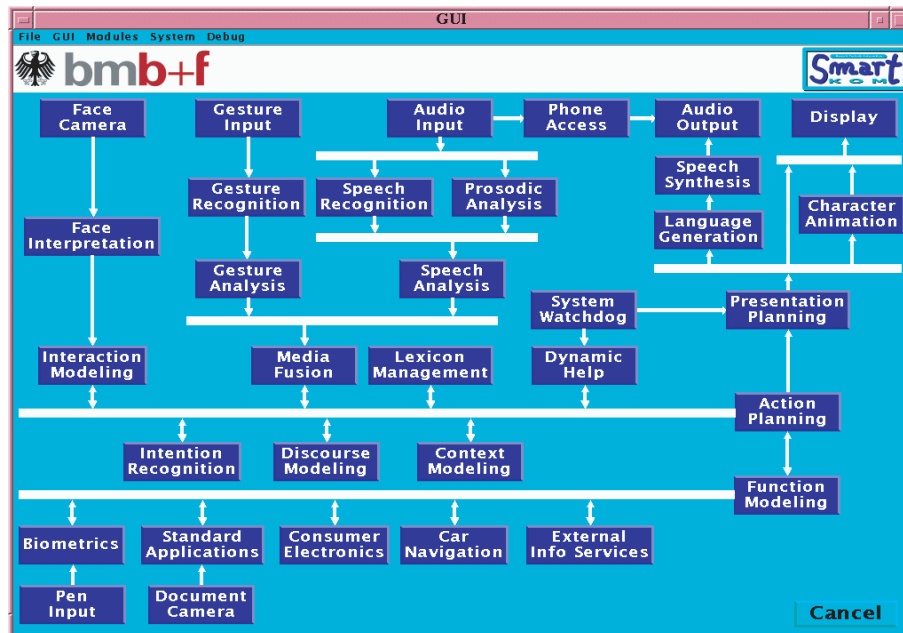


Figure 3.4: SMARTKOM architecture showing the multi-blackboard design (Wahlster et al., 2001).

of the user). An intention recognizer then ranks the resulting interaction hypotheses. The modality fusion component is also augmented by a multimodal discourse model that can rank the resulting hypotheses within the scope of a particular context. One novel aspect of SMARTKOM is that the mutual disambiguation component is also capable of understanding non-standard interpretations such as irony and sarcasm.

SMARTKOM has a number of significant contributions that distinguish it from other projects. One such aspect is its plug-and-pay architecture, which supports multiple recognizers for a single modality (e.g. a user's speech signal is processed by three unimodal recognizers in parallel for speech recognition, to determine emotional prosody and to determine clause boundaries). Modality analysers may also be added and removed dynamically while the system is running, thus supporting the changing demands of users and the situative context that they are in. Another interesting aspect of SMARTKOM is its equal focus on output presentation as on input interaction. Multimodal output in SMARTKOM centres around an interface agent called Smartakus that can provide an anthropomorphic and affective user interface and contains a large repertoire of gestures including pointing-gestures, body postures, and facial expressions (Wahlster et al., 2001). Also of interest is that SMARTKOM is capable of not only understanding and representing a user's multimodal input, but also its own multimodal output as shown in the following example taken from Wahlster (2002b):

U: "I would like to go to the movies tonight"
 S: "This <gesture> is a list of films showing in Heidelberg"
 U: "Please reserve a ticket for the first one"

In this example, 'the first one' refers to a visual antecedent provided by the system rather than a

linguistic antecedent provided by the user. Another important result of the SMARTKOM project was the collection of a large amount of multi-channel audio and video data from experiments that in total consisted of 448 multimodal Wizard of Oz sessions (or 1.6 terabytes of data), useful for the functional and ergonomic design of systems.

SMARTKOM is one of the largest projects world-wide that tackle the issues concerning multimodal interaction. Several aspects of the project relate closely to the work outlined in this dissertation, while other aspects differ in their overall objectives. One area in which the MSA/BPN exhibits commonality with SMARTKOM is with regards to discourse processing and in particular SMARTKOM's three-tiered multimodal dialogue discourse model (modality, discourse, and domain layers), which for the purposes of crossmodal reference resolution stores information not only on what is verbalized but also on what is visualized (Wahlster, 2003). For example, in the MSA, objects written to the blackboard are stored as 'modality objects' (i.e. speech, handwriting, intra-gesture, or extra-gesture). Each modality object also contains a link to both the proposed semantic function of the object (e.g. provider of 'feature' or 'object' information) and to the individual modality's proposed N-best list of hypotheses for the semantic function (e.g. in the case of an object referent: N1="PowerShot S50", N2="PowerShot S60", N3="PowerShot S45"). This modelling of 'semantic information providers' is analogous to SMARTKOM's discourse object layer. Finally, the domain objects within SMARTKOM can be likened to the multimodal 'communication acts' defined in the MSA/BPN. For example, communication acts in the MSA define that a <Feature><Object> entry is valid in contrast to a <Object><Object> entry. These acts are statically defined within the MSA/BPN's program code in comparison to SMARTKOM, which defines its domain model/ontology more flexibly using the ontology language OIL.

Symmetric multimodality is another area in which the MSA/BPN has commonalities with SMARTKOM. In SMARTKOM this is based on the use of a virtual character that exhibits many human characteristics, thus allowing both the user and the virtual character to communicate via speech, pointing-gesture, and facial expression. Similar to a user, the virtual character can speak, use his body, arms, and hand movements to define gesture, and facial expression to portray emotions. The MSA also caters for symmetric multimodality (for the modes speech, handwriting, and gesture), but does not focus on the use of virtual characters to do this. The validity of the implementation can be seen by considering the MSA's system output; speech is presented via a speech synthesizer, handwriting is displayed as text, and pointing-gestures used for referent selection by the system are modelled as boxes drawn around a referent (for intra-gesture) and as a spotlight directed at a referent (for extra-gesture). The primary difference between these two systems is that whereas SMARTKOM models the actual physical actions via a virtual character, the MSA models just the result of the action (i.e. the resulting audio output, written text, or referent selection).

In contrast to SMARTKOM, the MSA takes interaction one step further by incorporating symmetric multimodal interaction with real-world interaction, for example a user may select an object by picking it up, and the system may select an object by casting a spotlight onto it. Real-world interaction is one theme that is not addressed within the SMARTKOM project.

Anthropomorphization is another concept covered by both systems (i.e. assigning human characteristics to non-human things, see Wasinger and Wahlster (2006) and section 4.3). In SMARTKOM, a user can interact with Smartakus as though Smartakus were a real person. This differs in the MSA where a set of shopping products are instead used. In the MSA, an additional communication layer (both for user input and system output) is modelled so that users can interact either 'directly' or 'indirectly' with products that exhibit anthropomorphic characteristics when spoken to directly, for example each object is assigned different voice characteristics.

One other difference between SMARTKOM and the MSA/BPN is that the MSA additionally caters for the modality of handwriting, which was only used in the SMARTKOM system during biometric signature identification. Finally, SMARTKOM was designed to be distributed across a number of servers. As a result, even the mobile versions of SMARTKOM (both the tablet PC and PDA versions) are linked to servers via WLAN. In the case of the SMARTKOM PDA implementation, the mobile device works via a VNC connection to a distributed server, such that all interaction and processing is conducted remotely as discussed in section 3.1.7.1.

3.1.7.1 SmartKom-Mobile

Bühler, Minker, Häußler, and Krüger (2002) describe the SMARTKOM-MOBILE system in more detail. SMARTKOM-MOBILE is defined as a prototype system for multimodal human-computer interaction in mobile environments, and it supports location-aware in-car and on-foot navigation. SMARTKOM-MOBILE is based on a client-server architecture in which the PDA client is only responsible for managing the content of the iPAQ screen and gathering pen events. This data is then transmitted to a server for interpretation via a VNC connection (see figure 3.5).

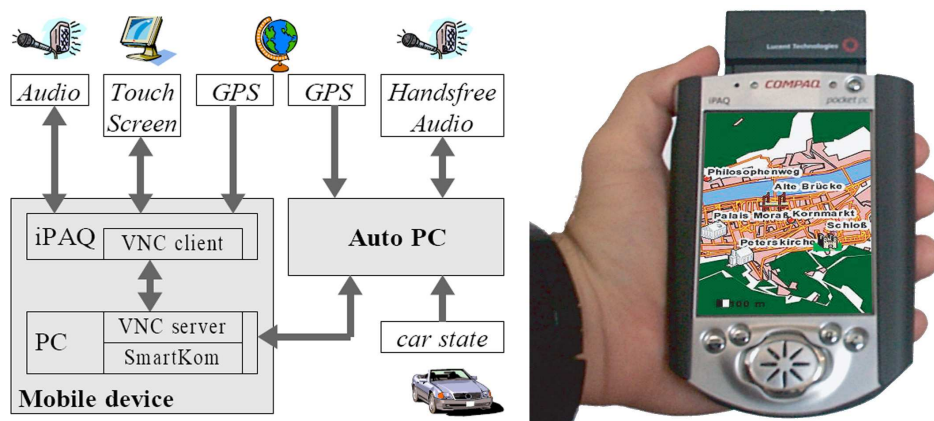


Figure 3.5: SMARTKOM-MOBILE: Hardware and software components used in the scenario (Bühler et al., 2002) and a picture of the interface used for navigation (Wahlster, 2002b).

SMARTKOM-MOBILE is novel and similar to the work in this dissertation in that it defines a range of different modality combinations for interaction with the system. The combinations defined in SMARTKOM-MOBILE (based on speech and graphics/pointing-gesture) are: default, listener, silent, speech-only, and suspend. In the default mode, all modalities are enabled for input and output (useful when privacy or disturbing others is not an issue). In the listener mode, speech and graphics/gestures are available for output, but only gesture is available for input. In the silent mode, only graphics/gestures are available for input and output (similar to a traditional GUI), while in the speech-only mode, only speech is available for input and output (useful when the user is driving a car). Finally, the suspend mode disables all input and output modalities (useful when the user has a high cognitive load or is in a dangerous situation). Another interesting aspect of the system is that it tries to encourage a user to switch modes if recognition accuracy is too low (e.g. by asking the user to ‘show’ rather than to ‘tell’ the system something). In comparison to SMARTKOM-MOBILE, the MSA defines and evaluates a much larger number of multimodal combinations (23 in total) for use with mobile devices.

3.1.7.2 SmartWeb

SMARTWEB (2004-2008) (Wahlster, 2006a; Reithinger et al., 2005) is the current follow up project to SMARTKOM (1999-2003). Although SMARTKOM addresses multiple domains (e.g. TV programme guide, telecommunications assistant, travel guide), it only supports restricted-domain dialogue understanding. SMARTWEB goes beyond SMARTKOM in that it will combine the previous focus of multimodal interaction with that of open-domain question-answering, using the entire Web as its knowledge base. The main goal of the SMARTWEB project is to lay the foundations for multimodal interfaces to wireless Internet terminals (e.g. smart phones, Internet phones, PDAs) that offer flexible access to Web services.

The SMARTWEB project brings together research in the fields of mobile Web services, intelligent user interfaces, multimodal dialogue systems, language and speech technology, information extraction, and Semantic Web technologies. The project is based on three efforts that have the potential to form the basis for the next generation of the Web. The first effort is the Semantic Web (Fensel, Hendler, Lieberman, & Wahlster, 2003), which provides the tools for the explicit markup of the content of webpages. The second effort is the development of semantic Web services, which results in a Web where programs act as autonomous agents to become the producers and consumers of information. The third important effort is information extraction from large volumes of rich text corpora available on the Web.

By exploiting machine-understandable content in semantic webpages, SMARTWEB will extend today's search engines by allowing not only for intelligent information-seeking dialogues but also for task-oriented dialogues (e.g. programming a navigation system to find an appropriate soccer stadium). Since semantically annotated webpages are still rare due to the time-consuming and costly manual markup, SMARTWEB uses advanced language technology, information extraction methods, and machine learning, for the automatic annotation of traditional webpages encoded in HTML and XML. SMARTWEB generates such semantic webpages offline and stores the results in an ontology-based database of facts that can be accessed via a knowledge server. In addition, SMARTWEB uses online question-answering methods based on real-time extraction of relevant information from retrieved webpages. SMARTWEB is furthermore expected to provide context-aware user interfaces to support a user in different roles such as driving a car, driving a motor bike, as a pedestrian, and as a sports spectator.

3.1.8 COMIC

COMIC (Conversational Multimodal Interaction with Computers) (Boves et al., 2004) is a project that was funded by the EU in the area of long term and high risk research. The project combined software and system development with experiments in human-human and human-computer interaction in language-centric multimodal environments.

The focus of COMIC was on cognitive-science research relating to multimodality. Its objectives were to build models of the cognitive processes involved in multimodal interaction in both fixed and mobile working environments, and to show the usability of these cognitive models for novel eWork and eCommerce services. One application that is outlined under COMIC is a design tool for bathrooms, which was extended so that interaction is not just possible via a mouse, but rather also via multimodal input. System components developed in the COMIC project include: dialogue and action management, multimodal fusion and fission, speech recognition, pen-based gesture recognition (handwriting, sketches, and deictic gestures), and output presentation. Speech

recognition in COMIC is implemented using the Hidden Markov Model Toolkit⁷. Supporting the recognizer are context-dependent phone models that are trained using the German Speech-Dat database and a language model that is inferred from recordings taken during experiments. Pen input recognition is implemented with algorithms developed by the NICI⁸. The modality fusion components of COMIC are based on the procedures and software developed at the DFKI in the framework of the SMARTKOM project (Boves et al., 2004). The distributed component architecture upon which COMIC is based, called MULTIPLATFORM (Herzog et al., 2003), was also originally developed in the Verbmobil and SMARTKOM projects.

Peculiar to the COMIC system is that user interaction is system-driven (rather than user-driven or mixed-initiative) (Boves et al., 2004). This means that a user's speech and pen input are confined to a fixed time window following the end of a system prompt. The authors state that if the user interacts outside this time period - which is denoted by a green or red coloured square on the tablet PC's display - their interaction is disregarded. During usability studies, this interaction paradigm was shown to be difficult for subjects to abide by, and recognition accuracy declined as a substantial proportion of input utterances were truncated because they exceeded the maximum allotted time window. Interaction in COMIC was defined to be drawing or writing input combined with speech input. This too is an aspect which differs to the MSA/BPN where a user can choose to interact either unimodally (e.g. just speech or just gesture) or multimodally (e.g. speech and gesture combined) when providing input to the system.

3.1.9 MIAMM

MIAMM (Multidimensional Information Access using Multiple Modalities) (Reithinger, Lauer, & Romary, 2002; Pecourt & Reithinger, 2004; Reithinger et al., 2005) is an EU funded project that was carried out by a large consortium including INRIA-LORIA⁹, DFKI, TNO¹⁰, Sony International, and Canon Research. The aim of MIAMM was to develop new concepts and techniques in the field of multimodal interaction to allow for faster and more natural access to multimedia databases. In addition to interaction via a graphical user interface, multilingual speech (German, French, English) and haptic interaction (senso-motoric, and tactile) are available for data selection and data visualization of a large-scale MP3 music database. On the output side, an MP3 player (to play music) and pre-recorded speech prompts (to provide acoustic feedback) were incorporated into the system.

The specific type of haptic device used in MIAMM is a PHANToM force-feedback unit¹¹ with three degrees of freedom. This is required to simulate the buttons of the music player and can actively exert forces to resist or aid the movement of a user's fingers and to provide vibrotactile signals to the fingertips. Envisaged uses of haptic feedback in MIAMM were for example strong/weak feedback to imply the existence of large or small amounts of data while navigating a music database and haptic feedback in the form of rhythms (for describing the type of music currently selected). Four different visualizations were created for haptic-visual interaction over the life of the project: conveyor belt/wheel, timeline, lexicon, and map/terrain (Reithinger et al., 2005).

⁷Hidden Markov Model Toolkit, <http://htk.eng.cam.ac.uk>

⁸NICI: Nijmegen Institute for Cognition and Information

⁹INRIA: Institut National de Recherche en Informatique et Automatique

¹⁰TNO: Netherlands Organisation for Applied Scientific Research

¹¹PHANToM, http://www.sensable.com/products/phantom_ghost/

From a functional point of view, multimodal fusion in MIAMM is divided into three mechanisms: interpretation, context processing, and reference resolution and fusion. During interpretation, semantically significant entities such as the relational predicates like ‘subject’ and ‘object’ are identified and added to the live discourse context. During context processing, the significant entities are mapped to task-oriented dialogues that MIAMM calls context frames. A novel feature of MIAMM is that context frames can be populated either during a single utterance or over several consecutive utterances. During the final stage of reference resolution and fusion, under-specified reference domains are integrated or merged (Kumar, Pecourt, & Romary, 2002) with the live context frame. This method for interpreting multimodal interaction shares similar aspects to the previously described processing layers found in SMARTKOM (i.e. modality object, discourse object, domain object) and the MSA/BPN (i.e. modality objects, semantic objects, communication acts).

MIAMM is based on a distributed hub-and-spoke infrastructure similar to that used in the GALAXY project (Goddeau et al., 1994), through which multimodal interaction on the portable music device is made possible.

3.1.10 COMPASS

COMPASS (COMprehensive Public informAtion Services System) (Aslan, Xu, Uszkoreit, Krüger, & Steffen, 2005) is an ongoing Sino-German cooperation aimed at creating an information system that will help mobile visitors of the Beijing 2008 Olympic Games to access information services through the use of multimodal (speech, handwriting, pointing-gesture) and multilingual (English, German, and Chinese) technologies. Multimodality in the system provides for natural interaction with mobile devices, while multilingual and crosslingual technologies allow information sources represented in foreign languages to be exploited by a user. The system’s architecture supports a wide range of differing services and is based on the FLAME2008 service platform from Fraunhofer (Holtkamp et al., 2003). Three service categories exist in COMPASS: information services (e.g. eating and drinking, and weather), transaction services (e.g. e-commerce), and composed services (i.e. services that integrate multiple services to deal with more complex tasks, e.g. a taxi dining service). The taxonomy for these services and in particular that for the domain of tourism is derived from the project MIETTA (Xu, 2003).

Typical user interaction in the system consists of a user first finding an applicable service based on category names (e.g. city info) and/or keyword searches (e.g. ‘Forbidden City’). Once a particular service has been selected, the user can then interact and retrieve service-specific information. This is based on query templates that are designed for each individual service category. For example, one service described in COMPASS is that of a smart dining service, in which a visitor to Beijing interacts with the system to find an appropriate eatery/restaurant. Typical multimodal interactions with this service include for example “Show me the ingredients of this <gesture> dish” and “Translate this English text <handwriting or gesture> to Chinese”. In the latter example, the system is also capable of speaking out the resulting text in Chinese.

The multimodal components of COMPASS originate from the work that was conducted on the MSA/BPN, and the interaction modalities supported in COMPASS are being chosen based on the results of user studies conducted on the MSA (Wasinger et al., 2005; Wasinger & Krüger, 2005, 2006). Once finished, COMPASS will support unimodal and multimodal interaction based on the communication modes speech, handwriting, pointing-gesture, and relevant combinations thereof. The COMPASS project also foresees the integration of OCR handwriting recognition based on

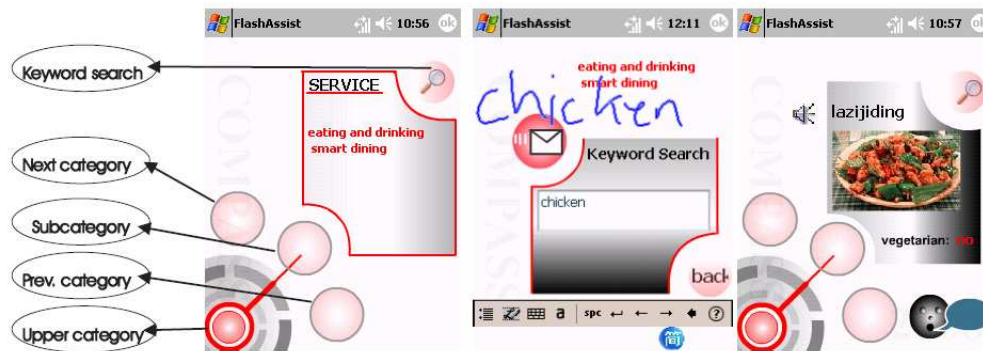


Figure 3.6: COMPASS smart dining service (Aslan et al., 2005).

capturing characters (e.g. Chinese characters) via a digital camera. Similar to the MSA/BPN, IBM Embedded ViaVoice is used for speech recognition and ScanSoft RealSpeak Solo is used for speech output. Both these packages are designed for PDAs and support the multilingual requirements of the project (i.e. English, German, Chinese). Microsoft Transcriber is required for English/German character recognition on the PDA, while CE-Star Suite v2.5 is being used for the Chinese character recognition on the PDA. The PDA components are programmed in a combination of C/C++ and Macromedia's Flash (vector animation) for the graphical user interface.

COMPASS and the MSA/BPN have many commonalities (e.g. multimodal, multilingual) and this is due to some components within the COMPASS project being modelled on the MSA/BPN project. One novel aspect of COMPASS is its multilinguality, which is defined in a component called the COMPASS Translation Centre. This component provides an interface for open-domain machine translation (via services such as Google Translate¹² and AltaVista's Babelfish¹³) and tourism-specific translation (via a handcrafted digital tourism phrase book). When solving translation requests, open-domain translation is only used if the closed-domain phrase book can not find a result first. Another novel feature of the COMPASS project is that it combines multimodal interaction not just with multilingual interaction, but also with crosslingual (e.g. when a Chinese speaking taxi driver and an English speaking tourist are in dialogue with one another) and mixed-lingual interaction (e.g. when a bilingual tourist formulates utterances that are partly in English and partly in Chinese). This flexibility is particularly useful when a user is unable to pronounce a foreign-language referent (e.g. "What does this mean?" <point-gesture=Chinese character>).

3.1.11 Project Short-form Comparisons

In this section, the similarities and differences of the described projects are consolidated and presented in a concise table. The aspects on which the projects are compared include their range of supported communication modes, their support for real-world interaction, multi-lingual interaction, symmetric multimodality, and anthropomorphic interfaces. Comparisons are also drawn between the degree of user mobility when using the system and the type of computing platform upon which the system is built. Many of the projects use different terminology when defining their

¹²Google Translate, http://www.google.com/language_tools

¹³AltaVista Babelfish, <http://babelfish.altavista.com>

work. For this reason, several of the categories are explained below:

- **Multimodal communication:** The ‘communication modes’ that arise in this literature study include the set: speech (S), writing (W), and gesture (G). The term ‘writing’ encompasses both characters (C) (e.g. the Latin alphabet) and symbols (S) (e.g. Braille, graphical languages like Chinese, military symbology, primitive symbols like shapes, and drawings). The term ‘gesture’ encompasses communication modes like pointing (P) (e.g. with a finger, pen, or mouse), hand movement (H) (e.g. hand-twists, pickup & putdown gestures), facial expressions (F), and eye-gaze (E). The complete number of communication modes considered in these project comparisons can thus be outlined by the set: {S, WC, WS, GP, GH, GF, GE}. Some systems categorize their communication modes via different terms such as ‘voice’ and ‘pen’, as in the case of QUICKSET. These terms are considered as part of the instrumentation required for a communication mode rather than the actual communication mode itself and are thus not included in the definition.
- **Real-world interaction:** ‘Real-world interaction’ refers to the user’s ability to physically interact with objects in their surrounding environment. Only two systems fulfilled this requirement (MSA/BPN and RASA), although there was also a visible trend for virtual interaction (e.g. PUT-THAT-THERE and QUICKSET-3D HAND).
- **Multilinguality:** Most of the surveyed systems cater for just one language, e.g. English or German. Those that are ‘multilingual’ have their relevant languages outlined in the table.
- **Symmetric multimodality and anthropomorphic interfaces:** Most (but not all) of the systems that contained a virtual character¹⁴ support both ‘symmetric multimodality’ and an ‘anthropomorphic interface’. The MSA is one exception where symmetric multimodality and an anthropomorphic interface are provided without using a virtual character for presenting system output.
- **User mobility:** Most of the analysed systems are built for scenarios that differ from the traditional (and old fashioned) desktop computing paradigm. Some systems require the user to sit down in an armchair (e.g. PUT-THAT-THERE, EMBASSI, and SMARTKOM-HOME). Other systems depict the user standing up in their scenarios (e.g. in front of a public information kiosk like in parts of the SMARTKOM-PUBLIC scenario, RASA, and QUICKSET-3D HAND). Yet other systems allow the user to move around, either by walking or running (e.g. QUICKSET, MATCH, MUST, SMARTKOM-MOBILE, COMIC, and COMPASS) or even driving (e.g. EMBASSI, SMARTKOM-MOBILE). The degree of ‘user mobility’ is defined as being either stationary (S, e.g. sitting, standing) or mobile (M, e.g. walking, running).
- **Computing platform:** With regards to choice of ‘computing platform’, most are distributed. MATCH is the only system capable of working entirely distributed and entirely embedded (although the computing platform used for MATCH is a tablet PC rather than for example a smaller and more resource restricted PDA). The MSA/BPN and COMPASS are two systems that are capable of working entirely embedded for most modalities (and providing full system functionality for these modalities), but when distributed cater for a larger range of modality combinations (e.g. extra-gesture interaction with physical real-world objects). Extending this, very few systems are capable of working offline under mobile conditions, and

¹⁴Virtual character, see also <http://www.virtual-human.org>

even fewer place emphasis on specifically designing for the limitations of handheld devices (e.g. lack of screen space, and restricted memory and computing power). For the purposes of this study, the systems are classified as being either: distributed (D) or embedded (E).

System	Multimodal Communication Modes	Real-world Interaction	Multi-lingual	Symmetric Multi-modality	Anthropomorphic Interface	User Mobility	Computing Platform
Put-That-There MIT, USA	S, GP	Partial	–	–	–	S	D
XTRA DFKI, Germany	WC, GP	–	–	–	–	S	–
QuickSet CHCC/OGI, USA	S, WC, WS, GP	–	–	–	–	M	D
QS-Rasa CHCC/OGI, USA	S, WS, GP	Yes	–	–	–	S	D
QS-3D Hand CHCC/OGI, USA	S, GP, GH	Partial	–	–	–	S	D
QS-ExertEnv CHCC/OGI, USA	S, WS, GP	–	–	–	–	M	D
MATCH AT&T Labs, USA	S, GP, WC, WS	–	–	–	–	M	E, D
MUST Eurescom ¹ , Germany	S, GP	–	En, Fr, No, Pt	–	–	M	D
EMBASSI Grundig ¹ , Germany	S, GP, GH, GF, GE	–	–	Yes	Yes	M	D
SmartKom-Public & SK-Home DFKI ¹ , Germany	S, GP, GF	–	En, De	Yes	Yes	S	D
SK-Mobile DFKI ¹ , Germany	S, GP	–	En, De	Yes	Yes	M	D
COMIC MPI ¹ , Netherlands	S, WC, WS, GP	–	–	Yes	No	M	D
MIAMM Loria ¹ , France	S, GP ²	–	En, De, Fr	–	–	S	D
COMPASS DFKI ¹ , Germany	S, WC, WS, GP	–	En, Ch, De	–	–	M	E, D
MSA/BPN DFKI, Germany	S, WC, WS ³ , GP, GH	Yes	En, De	Yes	Yes	M	E, D

Table 3.1: Multimodal project comparisons. ¹Lead organizer but not the only partner in the project. ²MIAMM also incorporates haptic, a mode not covered in this table. ³WS in the MSA is represented by a directional line to increase/decrease the speech of the visual-WCIS scroll bar.

One area of interest that is outside the scope of the table is that not all of the systems provide the same degree of flexibility with regards to usable modality combinations (i.e. combinations formed from the supported base modality groups). For example, the PUT-THAT-THERE system which caters for the modes S and GP (see table 3.1) only caters for the modality combinations S-only and S-GP combined. Other systems that support only a fairly limited number of modality combinations include XTRA (WC-only, GP-only), QUICKSET-3D HAND (S-GP and S-GH(twisting)), MUST (S-GP only), SMARTKOM-MOBILE (S-Only, GP-Only, S-GP combined), and COMIC (S-WC, S-WS, S-GP). On the other hand, systems like QUICKSET, SMARTKOM-PUBLIC (and -Home), EMBASSI, and the MSA/BPN support a more flexible range of modality combinations. Reasons for this difference between systems are varied. For example, some modality combinations are more suited to specific application domains, some systems focus on providing only one set of

easy-to-learn multimodal interactions, and yet other systems simply do not have an architecture flexible enough to capture a wide range of modality combinations.

Another aspect not covered in the table is that only a select few architectures actually cater for semantically overlapped information in their error resolution strategies (e.g. QUICKSET, SMARTKOM, EMBASSI, COMPASS and the MSA/BPN). A reason for this might be that users dislike providing redundant information to a system, especially when system accuracy is already good and/or when the lack of errors is non-critical (Wasinger et al., 2005; Wasinger & Krüger, 2006).

Two final aspects worthy of comparison relate to the multimodal integration patterns and modality fusion processes found in the systems. Regarding multimodal integration patterns, to the best of the author's knowledge all systems cater for both sequential integration and simultaneous integration of modalities¹⁵. With regards to modality fusion, all of the surveyed systems (with the exception of a select few like SMARTKOM) focus solely on late fusion rather than early fusion. A common reason for this is that more knowledge sources become available later on in the processing stage and thus allow for the more robust interpretation of possibly incomplete and inconsistent multimodal input (Wahlster, 2003).

3.2 Mobile Users and Instrumented Environments

The previous section surveyed only projects that had a prominent focus on multimodal interaction without much regard for the scenario. A focus of this dissertation is however to design 'mobile' multimodal systems that support users in everyday tasks. It is for this reason that the focus is now moved from projects specializing in multimodality, to projects that have been designed for environments in which the user is mobile and in need of assistance while performing everyday tasks. Two scenarios are considered, in particular that of shopping and navigation. In section 3.2.1 a range of commercial and research shopping assistants covering a variety of themes from decision-theoretic planning to mobile interface design are described. The context of shopping is still almost completely untouched by the advancements of multimodal interaction over the last few years, and the MSA builds upon these scenarios by providing a shopping system that does cater for multimodal interaction for mobile users. Section 3.2.2 then describes a variety of mobile map-based guides covering themes like outdoor and indoor navigation and interaction within confined environments, and the section also summarizes different aspects relating to navigation and map-based interaction.

Aside from shopping and navigation being tasks that apply to the majority of users everyday, these two scenarios were also selected for their ability to combine into one larger scenario, and for their ability to complement each other in aspects like mobile navigation and multimodal interaction. Such a scenario is described in (Wasinger & Krüger, 2004), where navigation at different scales (e.g. outdoor and indoor) is combined with interaction with different types of objects (e.g. buildings and shopping products). For example, when outdoors, a user might navigate streets and footpaths in a city environment and interact with surrounding building objects like a shopping complex. When indoors, a user might navigate corridors within a building environment and interact with surrounding rooms like an electronics store. Within a room environment, the user might then continue to navigate isles and interact with containers like tables or shelves, and finally, a user

¹⁵Sequential interaction occurs when a user provides input in multiple modalities one after the other in time, while simultaneous interaction occurs when a user provides input in multiple modalities together in time

might navigate the levels within a shelf and interact with the physical objects contained within it. This short scenario outline is representative of the MSA/BPN, and it also shows how mobile multimodal interaction can be seamlessly applied to a range of everyday tasks.

3.2.1 Shopping Assistance

Shopping has been identified by Falk and Campbell (1997) as a “realm of social action, interaction and experience which increasingly structures the everyday practices of urban people”. Sociologists have described the shopping experience as complex and ambiguous, and full of contradictions and tensions. Lehtonen and Mäenpää (1997) for example state that shopping is ambiguous in nature because it is essentially a private experience that occurs in a public setting. They argue that shopping is contradictory in that it is an experience that yields both pleasure and anxiety which can easily morph into a nightmare. The act of shopping can also be seen to entail tension in the form of rationality versus impulse, and between a pleasurable social form and a necessary maintenance activity. From these perspectives and the associated intricacies, it is evident that shopping is a subject consisting of considerable depth. Shopping has a central role within society, and it is thus a prime field of study for mobile applications that provide benefits for retailers and/or customers.

This section summarizes a range of shopping assistants that focus on mobile and ubiquitous computing. Matching the diversity that entails the act of shopping itself, the described assistants cover a wide range of product domains such as everyday grocery items like bread and milk, electronic items like digital cameras, and car sales. Some of the described implementations are location and context aware and delve into the realms of mobile, ubiquitous, and pervasive computing, and ambient intelligence. Their architectures are often based on instrumented environments encompassing shopping trolleys and handheld devices that accompany a user around a store. Extending upon the roles of the traditional real-world sales assistant, the main practical goals that these shopping assistants focus on include guiding a user around a store and providing users with supporting information in the form of personal shopping lists and product specifications. The research goals of these systems focus on topics like conversational dialogues, augmented reality, and plan recognition. Many of the systems are location and context aware, which is also a commonality of other types of mobile guides including navigational guides like REAL (Baus, Krüger, & Wahlster, 2002) (see also section 3.2.2), and museum guides like ALFRESCO (Stock, 1991) and PEACH (Rocchi, Stock, Zancanaro, Kruppa, & Krüger, 2004). Some systems are now also beginning to merge different application domains together, for example the MSA/BPN described in this dissertation, in which a multimodal shopping assistant is tightly linked to a mobile pedestrian navigation and exploration system (Wasinger & Krüger, 2004).

A second class of shopping assistant are those based on Web-agents. These assistants collate data from many different product vendors and then allow customers to access the results via the Web in the form of comparison charts. In contrast to mobile shopping assistants that have the goal of improving a customer’s ‘in-store’ shopping experience, Web-agents focus on optimizing a customer’s ‘online’ shopping experience and are generally related to the paradigm of home Internet shopping. This class of shopping assistant will not be discussed as it has little to do with either mobile and ubiquitous computing or multimodal interaction. For more information on this type of shopping assistant see (Menczer, Street, Vishwakarma, Monge, & Jakobsson, 2002), where a number of such Web-based shopping agents are outlined.

3.2.1.1 Commercial Shopping Assistance Systems

Commercial shopping assistants have the primary goal of improving a customer's in-store shopping experience, while at the same time increasing the store's level of efficiency and thus profits. Two commercial shopping assistants include the METRO Group's FUTURE STORE¹⁶ and IBM's SHOPPING BUDDY¹⁷.

The goal of the FUTURE STORE was to integrate multiple emerging technologies into an existing store and to evaluate the technologies as a preliminary step towards broad integration of the technologies throughout the retail chain. Key technology components used in the store include servers, RFID readers, kiosks, desktop and mobile PCs, handheld devices, and network components. The FUTURE STORE installation can be seen to benefit both retailers and customers. From a retailer's point of view, RFID tags can be placed on pallets and individual products to allow inventory throughout the store's supply chain to be tracked. This is achieved through the use of RFID readers, which for example if attached to shelves can notify staff when products need to be replenished. The system also allows staff to access business intelligence through mobile PDA interfaces that allow stock levels to be checked, item information to be requested, and product prices to be automatically changed on electronic advertising displays. From a customer's point of view, benefits revolve around a more convenient, engaging, and customized shopping experience. A loyalty card allows customers to begin shopping before they enter the store by selecting goods that they plan to purchase from a website and saving these to the card for later use in conjunction with an instrumented shopping trolley. Touch screen tablet PCs mounted on top of trolleys provide shopping lists, product descriptions and pictures, pricing information and store maps, along with running totals for products placed inside the trolley. Promotional offers are also displayed on the trolley's display based on the customer's location in the store, and 19" displays mounted above product areas offer further promotional information using video and animation.



Figure 3.7: METRO's FUTURE STORE instrumented trolley (left) and IBM's SHOPPING BUDDY (right).

On a similar front, IBM's SHOPPING BUDDY has been deployed in several test stores and has many of the same goals as that of the METRO FUTURE STORE. The SHOPPING BUDDY for example displays running totals of how much customers have spent and saved during their visit. It

¹⁶FUTURE STORE: Creating the Future at METRO Group, <http://www.future-store.org>

¹⁷SHOPPING BUDDY: Stop & Shop grocery drives sales and boosts customer loyalty with IBM personal shopping assistant, <http://www.pc.ibm.com/store/products/psa/>

reminds them of past purchases, and allows them to place orders with the supermarket's deli from their trolley and to pickup their requests once the system indicates they are ready. Complementing the trolley's functionality, a location tracking system permits the delivery of targeted promotions and is also capable of helping customers navigate through the store and locate products. Similar to the FUTURE STORE, this system reduces checkout lines by allowing customers to scan and bag items as they shop and then complete their transactions using a self-checkout system.

3.2.1.2 Research Shopping Assistance Systems

Research is being conducted on a number of fronts with the aim of extending the capabilities of the commercial shopping assistant implementations. Current research covers expansions to the general in-store scenario to cater for: additional surroundings like that of the family home (Kourouthanassis, Koukara, Lazaris, & Thiveos, 2001); the use of plan recognition (Schneider, 2004) and decision theoretic planning (Bohnenberger, Jameson, Krüger, & Butz, 2002) to better predict and guide a customer throughout a shop; the incorporation of augmented reality (Zhu, Owen, Li, & Lee, 2004); conversational interfaces (Chai, Horvath, Kambhatla, Nicolov, & Stys-Budzikowska, 2001; Rist et al., 2002); and the design of shopping assistant interfaces for mobile devices (Newcomb, Pashley, & Stasko, 2003) and the visually impaired (Ebaugh & Chatterjee, 2004).

The MYGROCER project (Kourouthanassis et al., 2001) extends the general in-store scenario to cater for in-house and on-the-move interaction. Whereas the functionality of the in-store scenario is based on an instrumented shopping trolley and includes displaying a user's shopping list as well as in-store promotions based on previous customer buying behaviour, the in-house scenario allows products that are removed from a particular location to be added to the user's shopping list and accessed via a mobile phone connection. The on-the-move scenario incorporates notifications about products that have run out-of-stock and allows for the home delivery of such products.

The SMART SHOPPING ASSISTANT (Schneider, 2004) is an adaptive shopping assistant that utilizes plan recognition techniques to aid the user while shopping. It provides a proactive user interface driven by implicit interaction in a real-world shopping scenario. If a user picks up a product that has not been previously handled, the system may for example provide detailed product information, or may alternatively display a list of similar products or product chart-comparisons (e.g. if the user has two products, one in each hand). If the system infers that the user intends to cook a particular dish, a list of related products may also be displayed. Presentation output takes the form of dynamic HTML pages that are displayed on the shopping trolley's display.

In (Bohnenberger et al., 2002), a PDA-based system is developed to give customers directions through a shopping mall based on the type of products that the customer has expressed interest in, the customer's current location, and the purchases that the customer has made so far. The approach uses decision-theoretic planning to compute a policy that optimizes the expected utility of a customer's walk through the shopping mall, taking into account uncertainty about whether the customer will actually find a suitable product in a given location and the time required for each purchase.

Another interesting design is the PROMOPAD (Zhu et al., 2004). This is an in-store e-commerce system that provides context-sensitive shopping assistance and personalized advertising through augmented reality techniques. Individual objects that are encountered in the real-world are augmented with virtual complements so as to make the real objects more meaningful and appealing. The system is novel in that aside from adding new imagery relative to a focal product,

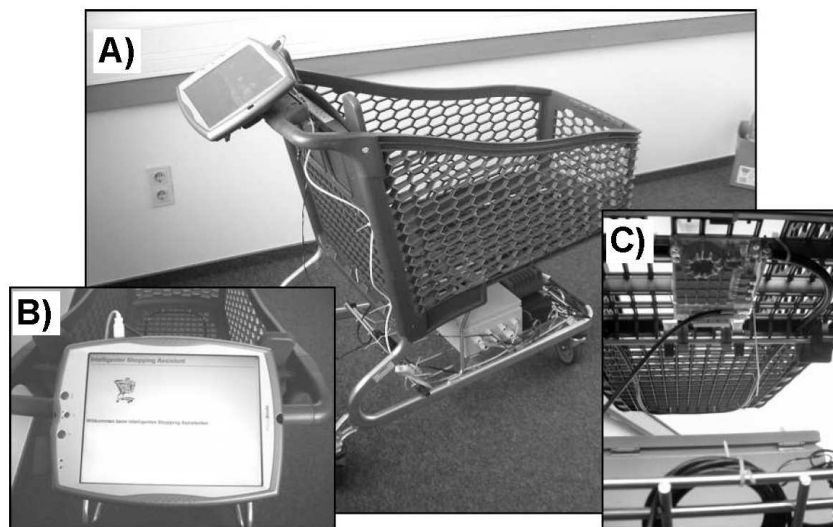


Figure 3.8: *The SMART SHOPPING ASSISTANT with plan-recognition technology (Schneider, 2004).*

the system can also remove elements of the image that may distract from the focal product. This system is based on a tablet PC with a camera mounted on the back. The display on the tablet provides a modified version of the camera image, which the customer can look at as though it were a ‘magic frame’ (i.e. see-through).

Two shopping assistants supporting conversational interfaces include the NLA (Natural Language Assistant) (Chai et al., 2001) and CROSSTALK (Rist et al., 2002). The NLA is an online conversational dialogue system that assists users in finding laptops by engaging them in dialogue. Based on a market survey, an appropriate set of natural language user vocabulary consisting of 195 keywords and phrases was acquired, and statistical n-gram models and a shallow noun phrase grammar for extracting keywords and phrases from user input were also generated. Subsequent user studies found that when compared to a menu driven system, the use of a conversational interface reduced the average number of clicks by 63% and the average interaction time by 33%. In comparison to this system, which encourages direct human-computer conversation, CROSSTALK (Rist et al., 2002) is an interactive installation in which agents engage in conversational car sale dialogues. It builds on the IMP (Inhabited Market Place) (André & Rist, 2001) by adding a virtual hostess called Cyberella to act as mediator between human visitors and the IMP application. The IMP is a virtual place (i.e. a showroom) where seller agents provide product information to potential buyer agents in the form of typical multi-party sales dialogues. This allows for human users observing the dialogue to learn about the features of a car.

SAVI (Ebaugh & Chatterjee, 2004) is a shopping assistant that specifically caters for the visually impaired. It is designed to aid blind and sight-impaired customers in identifying and selecting products from store shelves, by verbalizing the name, brand, and price of an item. In contrast to other systems in which product IDs are detected by readers that are built into instrumented shelves, in this implementation the product IDs are detected via an iGlove¹⁸ that contains the RFID reader. The proposed solution is said to have the benefit that it can also be used when putting items away

¹⁸Intel iGlove, http://www.intel.com/research/network/seattle_human_activity_recognition.htm

at the customer's home, provided the right infrastructure exists.

Newcomb et al. (2003) take a different focus with their research into mobile shopping assistants and discuss, on the basis of studies into people's grocery shopping habits, what an interface for mobile devices should actually look like. With this goal in mind, they design and evaluate prototypes and also perform usability tests within a true shopping environment. Based on user preferences on nine different spatial and contextual interface designs, they develop a user interface that is divided into three segments, the middle more prominent region consisting of a shopping list, the top consisting of a spatial map, and a promotional area at the bottom displaying revolving store specials.



Figure 3.9: CROSSTALK, left (Rist et al., 2002) and an example of a mobile shopping assistant user interface designed on the basis of user studies, right (Newcomb et al., 2003).

From this brief overview of shopping assistants, it can be seen that the design of shopping assistants for mobile users and instrumented environments is growing in number and slowly forming a mature market. Indeed, some of the benefits of such research are already becoming visible in the commercial marketplace, particularly in the form of instrumented shops consisting of smart trolleys and mobile handheld devices. Although the focus of these systems is highly varied, none as yet concentrate on providing the user with a multimodal interface similar to that presented by the MSA/BPN.

3.2.2 Map-based Guides

In this section, map-based mobile guides that have been developed to assist users with spatial tasks are compared. These tasks mainly consist of navigation tasks, but also include map-based tasks that support user interaction in confined environments, e.g. interacting with objects inside a museum. The section concludes with a table that contrasts the functionality of these map-based systems, both in terms of the scale of navigation and the type of modalities provided. Due to the

large number of mobile guides that already exist, analysis is limited to those systems described in (Wasinger & Krüger, 2004) and (Baus, Cheverst, & Kray, 2005).

Early research on mobile spatial information systems focussed on the technical problems of mobile computational platforms, e.g. how to localize mobile devices. The CYBERGUIDE system (Abowd et al., 1997) is able to localize users indoors (via infrared) and outdoors (via GPS) and provided simple black and white maps to support orientation in an unknown environment. One of the most prominent mobile spatial information systems is the GUIDE system (Cheverst, Davies, Mitchell, Friday, & Efstratiou, 2000), which provides tourists with information on places of interest in the city of Lancaster. Both the CYBERGUIDE and the GUIDE system allow only for simple pointing gestures and were not explicitly designed to explore multimodal research issues. The HIPS (Oppermann & Specht, 2000) project aimed at designing a personalized electronic museum guide to provide information on objects in an exhibit. The presentations were tailored to the specific interests of a user with the help of a user model and the user's location within the rooms of a museum. The implementation allowed for simple point gestures and speech commands, but did not allow for the fusion or parallel processing of these two modes.

The REAL system (Baus et al., 2002; Stahl et al., 2004) is a navigation system that provides resource adapted information on the environment (see figure 3.10). The user can use pointing gestures to interact with landmarks in the physical real-world to obtain more information, but natural language technologies like speech interaction are not catered for. In contrast, DEEP MAP (Kray, 2003), an electronic tourist guide for the city of Heidelberg, combines both speech and pointing gesture to allow users to interact more freely with map-based presentations. SMARTKOM and SMARTKOM-MOBILE (see section 3.1.7) are among the first systems to follow the paradigm of symmetric multimodality. Input to SMARTKOM-MOBILE can be provided through the combination of speech and gesture. SMARTKOM-MOBILE provides travel assistance for the city of Heidelberg through synthesized speech and through gestures performed by a virtual character. The QUICKSET system (see section 3.1.3) is one of the earlier sophisticated multimodal systems that allows for pointing gesture accompanied by speech utterances. QUICKSET was designed to facilitate military operations and the coordination of civil protection forces (i.e. fire fighters). In addition to speech and pointing gesture, QUICKSET also understands a range of written symbols (e.g. military based symbology) and handwritten commands.

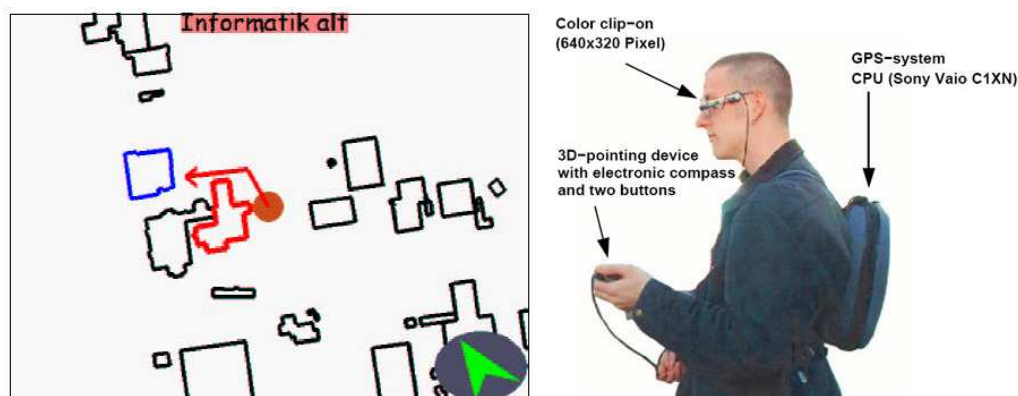


Figure 3.10: Example of the map interface used in REAL (left), showing also the components of the REAL system (right) (Baus et al., 2002).

The map-based guides described in this section share some commonalities with the MSA/BPN. For example the BPN provides navigational assistance on several scales (city, building, and room level) and is capable of interpreting speech input, pointing-gesture input, and multimodal input consisting of speech utterances fused with gesture. Similar to the REAL system, it also allows users to interact with and point to real-world objects in the surrounding environment to obtain more information. The BPN follows the basic principles of symmetric multimodality, as users can speak to the BPN and can receive spoken information from the system (i.e. information on landmarks and street names). The MSA takes this even further by providing information on products in a shop. When integrated into an intelligent environment, the MSA is able to detect which physical objects a user is currently holding and can perform ‘point-like’ system gestures by using a steerable projector as a light source (Butz, Schneider, & Spassova, 2004). The MSA also allows for handwriting, in addition to speech and gesture interaction (especially useful in noisy environments and environments where privacy is required). The summary of these systems is shown in table 3.2 below. The modalities intra-gesture and extra-gesture, as used in the table below, may be respectively defined as interaction with the virtual-world by selecting objects on a display and interaction with the real-world by selecting real physical objects from their surrounding environment. These concepts are described in detail in chapter 4.

System	Scale	Speech	Writing	Intra-Gesture (Type)	Extra-Gesture (Type)	Symmetric Modalities (Type)
Cyberguide	Building/Room	No	No	Point	No	No
GUIDE	City	No	No	Point	No	No
HIPS/Hippie	Room/Container	Yes	No	Point	No	No
REAL	City/Building	No	No	Point	Point with device	No
Deep Map	City	Yes	No	Point	No	No
SmartKom	City	Yes	No	Point	No	Speech, Gesture
QuickSet	City	Yes	Yes	Point and simple shapes	No	No
MSA/BPN	City/Building/Room/Container	Yes	Yes	Point and simple shapes	Point with device, pickup, putdown	Speech, Gesture

Table 3.2: Overview of map-based mobile guides (Wasinger & Krüger, 2004).

An important consideration for mobile and ubiquitous computing systems is how a user can interact with these systems. In section 4.1, the different communication modes used in the MSA/BPN are discussed. Particular focus is placed on tangible user interactions and gesture, the calculation of confidence values, and field study results on the accuracy, speed, and scalability of different communication modes that can be used in mobile settings. In section 4.2, multimodal interaction is categorized in terms of its temporal and semantic synchrony, and in terms of the degree of (semantic) overlap between different inputs. An outline of different recognizer configurations for capturing semantically overlapped multimodal input is also provided. Section 4.3, outlines the concepts of direct and indirect interaction, and anthropomorphization. This is followed in section 4.4 with an outline of symmetric multimodality in the MSA/BPN and a description of the encompassed presentation planning capabilities. The chapter closes with the analysis of an extended application context for the MSA/BPN in section 4.5, in which multiple users can interact with multiple devices and with a common set of applications, simultaneously.

4.1 Modal Interaction

In chapter 2, communication modes were identified to have unique strengths and weaknesses for given environments, tasks, and users. This aspect is particularly relevant for mobile users where the choice of a particular application task (e.g. navigation, shopping) and the features of a surrounding environment (e.g. noisy, crowded, rainy) may change with little warning. Users themselves also differ greatly, with typical demographic groups including age (children, middle-aged, elderly) and familiarity with a system (beginner, advanced) to name just a few.

This section forms a prelude to the following section on multimodal interaction. A range of different communication modes and their supporting software/hardware are described, including most importantly for this dissertation those that are implemented in the MSA/BPN applications. Particular focus is placed on tangible user interactions and gesture, which are very relevant to instrumented environments. This is followed with a discussion on the calculation of confidence values for the communication modes speech, handwriting, and gesture. In the final section the accuracy, speed, and scalability of the communication modes used in the MSA/BPN are discussed in relation to the results obtained from a field study conducted on the MSA at the CeBIT 2006 fair in Hannover¹.

¹CeBIT, <http://www.cebit.de>

4.1.1 MSA/BPN Communication Modes

Communication modes that are typical to the desktop computing interaction paradigm like mouse and keyboard are not discussed in this dissertation as they do not reflect state-of-the-art human-computer interaction in the fields of instrumented environments (Hagras & Callaghan, 2005) and ambient intelligence (Aarts & Encarnação, 2006), or ubiquitous (Beigl, Intille, Rekimoto, & Tokuda, 2005), pervasive (Gellersen, Want, & Schmidt, 2005), and mobile (Chittaro, 2003) computing. However, the evolution of such traditional devices does make for an interesting starting point. In particular, the mouse that was previously used to point at coordinates on a display can now be seen to be replaced by the use of a stylus and touch-sensitive screen in mobile device scenarios. In the MSA/BPN, such pointing action is extended even further to encompass not just stylus interaction, but also interaction with objects in the real-world. A similar evolution is occurring with the traditional keyboard device, where current research is focusing on the design of keyboards that are portable and can be used efficiently with only a single hand, for example the Twiddler² (Lyons, 2003).

In terms of human-human interaction, speech and handwriting are two forms of communication that are quite expressive and natural. For human-computer interaction, these forms of communication are being adopted slowly by the masses; a process which has so far taken over 30 years and has still only resulted in minimal impact. These two modes of communication do however remain very promising and the mobile device market is expected to also greatly influence their up-take. This is supported in that the primary communication mode for mobile devices like PDAs is now based on pen interaction in the form of pointing (the substitute for mouse interaction) and handwriting (the substitute for keyboard input).

Another evolving type of interaction is that of tangible interaction. Shneiderman (1992) defines the term ‘direct manipulation’ as referring to “the visual display of actions (the sliders or buttons) and objects (the query results in the task-domain display)” that a user can manipulate directly. In its original sense, direct manipulation represented actions like dragging visual objects across a display through the use of a mouse. Tangible User Interfaces (TUIs) extend upon this principle by employing physical artefacts both as representations and as controls for computational media (Ullmer & Ishii, 2001). In (Masui, Tsukada, & Sii, 2004) for example (see figure 4.1), a simple and versatile input device for ubiquitous computing is described. It is based on two optical mice (used to obtain 2D directional information on an object) and an RFID reader (to obtain ID information on an object). When an ordinary RFID-instrumented music CD is placed over the input device, music is played. The user can then rotate the CD to the right to forward to the next track, or rotate the CD to the left to rewind to the beginning of the current or previous audio track. This example of a tangible interface demonstrates how an ordinary object like a music CD case can be used as a control to the system. In the MSA/BPN, RFID-instrumented shopping products permit a user to directly manipulate the selection of items by either picking products up (to select them) or putting them down (to deselect them).

Oviatt (2000a) makes a distinction between ‘passive input’, which “requires no explicit user command to the computer”, and ‘active input’ like speech and handwriting where a user does intend for the interaction to be issued to the system. Vision and sensing technologies provide the basis for many types of passive input like gaze, head position, body posture, facial expressions, hand gestures, and user location and orientation. In the BAIR project, biosensors are attached to the user’s body to measure electrocardiogram and electrodermal activity, to infer a user’s state of

²Twiddler, Handykey Corporation, <http://www.handykey.com/>

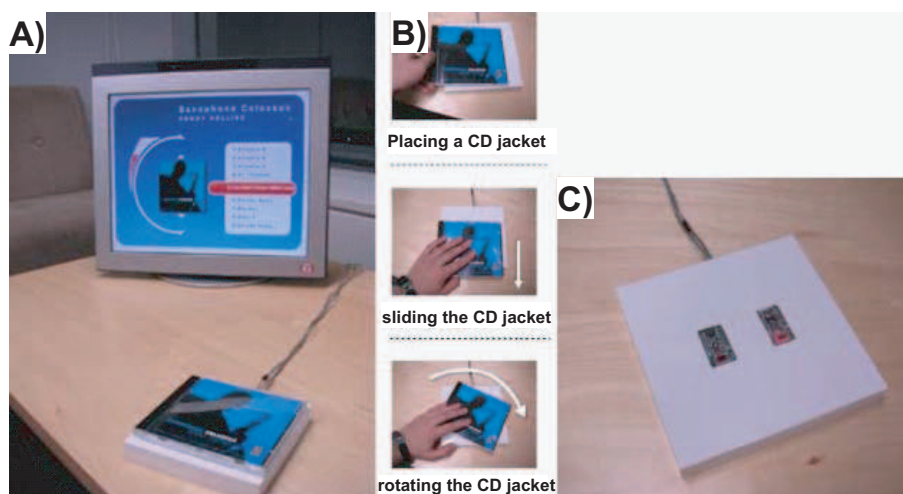


Figure 4.1: MouseField, showing A) and B) tangible interaction and C) the hardware consisting of two optical mice and an RFID reader (Masui et al., 2004).

being (e.g. under time pressure) (Wahlster, Krüger, & Baus, 2004b; Wilamowitz-Moellendorff, Müller, Jameson, Brandherm, & Schwartz, 2005). In SmartKom (Wahlster, 2002b), facial expressions of emotion are captured via a digital video camera to detect signs of annoyance and sarcasm based on the categorization of facial expressions into the groups negative, neutral, and positive. In the MSA/BPN, a variety of sensing technologies like GPS, magnetic compass, and 3-axis attitude sensor arrays (pitch, roll, and yaw) are used to determine a user's location and facing direction outdoors. When indoors, a user's location is detected via active RFID tags situated in an instrumented environment, and a user's facing direction may be determined based on signal strengths of the active RFID tags, and based on line-of-sight to infrared beacons that are used for indoor navigation and the identification of shelves in a room.

This section discusses the communication modes used in the MSA/BPN, namely speech, handwriting, and gesture. The MSA accepts input in all of the above mentioned modes and their combinations, while the BPN accepts only speech and speech-gesture combined input. Although the interaction examples given throughout this dissertation are in English, the MSA is capable of multilingual input and output in the languages of German and English, and the BPN is capable of input and output in German.

The communication modes used in the MSA/BPN are often abbreviated in this dissertation to: speech (S), handwriting (H), intra-gesture (GI), and extra-gesture (GE). The individual modality combinations that can be created from these modes (see section 4.2.2.2) are also often abbreviated based on the semantic constituents that they represent. For example <Feature modality="speech"><Object modality="intra-gesture"> is analogous to the modality combination 'SGI', where the feature always refers to the first abbreviation and object to the second (see section 5.1.3.1 for a definition of semantic constituents).

4.1.1.1 Speech Recognition

The speech recognizer that is employed locally on the mobile PDA device in the MSA and BPN applications is that of IBM Embedded ViaVoice (EVV)³. This recognizer is capable of recognizing continuous speech provided at speeds representative of normal human-human conversation, making the recognizer more flexible than discrete speech or isolated-word speech recognizers in which small pauses need to be left between each word that is dictated to the system (Cole et al., 1998). The recognizer is also speaker-independent, which means that no prior training of users is required for the system to be able to recognize their speech. IBM EVV uses an 11 kHz sampling rate, i.e. the analogue audio is digitally sampled at 11,000 samples per second. This represents a higher resolution than that currently used in mobile phones (8 kHz) but a lower resolution to that used by typical desktop recognizers (22 kHz). The recognizer requires only slightly more than 4MB of memory, but is therefore limited in certain aspects when compared to typical desktop recognizers. One limitation common to recognizers designed for embedded devices is that recognition results rarely provide timestamp information for individual words, thus making it difficult to resolve certain types of reference consisting of multiple referents.

Three important components of a speech recognizer are the acoustic models, vocabulary, and language models. Acoustic models are a mathematical representation of the sound (or phoneme) patterns in a language, and are often designed based on specific users and/or specific environment contexts. Because the BPN and MSA applications are required to function in public noisy environments like that of shopping and navigation, an acoustic model designed specifically for ‘automotive general-use’ is employed. A vocabulary is the list of words and their associated phoneme pronunciations that a speech recognizer can interpret. IBM EVV provides vocabularies consisting of over 30,000 predefined words. In the MSA and BPN applications, certain words (and their phonemes) such as camera product names or street and landmark names were additionally added to the application’s vocabulary list. Although 30,000+ words are recognizable by the speech engine, not all of these words are required by an application’s language model, and are thus not all relevant for the MSA/BPN applications. In a traditional sense, ‘language models’ represent statistical information associated with a vocabulary and describe the likelihood of words and sequences of words occurring in a user’s utterance.

There are two common methods for modelling spoken language: formal language and stochastic (or N-gram) language (Gorrell, 2004). In (W3C-StochasticLanguageModels, 2001), several of the main differences are defined. ‘Formal language models’ represent language via strict grammars such as context-free grammars (CFGs), or in the case of the MSA/BPN, finite-state grammars (FSGs) (Chomsky, 1956). When formal grammars are employed, users are only allowed to utter those sentences explicitly covered by the (often hand-written) grammar. ‘N-gram language models’ provide a recognizer with an a-priori likelihood of a given word sequence and are derived from large training texts that share the same language characteristics as the expected user input. Formal grammars are accepted to be more restricting than N-gram language models, but are simpler to design as they do not require the collection of data for large corpora. Formal grammars are particularly important for devices that have limited computing resources like PDAs.

In the MSA/BPN, the term ‘language model’ is used to refer to the set of MSA/BPN rule-grammars. These grammars are written in a format closely related to the Backus-Naur Form (BNF) called the Speech Recognition Command Language (SRCL), and they represent the set of allowable phrases that the recognizer will accept from a user. An ontology and a thesaurus often

³IBM Embedded ViaVoice, http://www.ibm.com/software/pervasive/products/voice/vv_enterprise.shtml

accompany the language model and these are used during semantic interpretation of the recognized utterances, as described in section 5.1. In the case of the MSA, a small ontology defines the various shopping product types and their relationships (e.g. a 'PowerShot S50' is a 'camera' is a 'shopping product' is an 'object'), while a thesaurus is used to represent a list of synonyms for object attributes (e.g. 'price' is a 'cost' is a 'worth'). In addition to the embedded speech recognizer, the MSA architecture also supports the incorporation of server-sided speech recognition, and a module was created for this purpose to transmit a user's speech signal over a TCP/IP connection as an 11 kHz 16-bit mono wav file (Feld, 2006). CMU Sphinx 4⁴ was chosen for server-sided speech recognition. This recognizer is written entirely in Java and supports both formal language models written in a format similar to BNF called the Java Speech Grammar Format (JSGF) and stochastic language models based on unigram, bigram, or trigram word-sequence predictions.

MSA and BPN Rule-Grammars: Rule-grammars in the MSA and BPN can be grouped into those that are static (or precompiled) and those that are dynamic (or generated during runtime). In the BPN, precompiled FSG grammars cover program control that is independent of a particular map, such as map control (e.g. "Zoom in"), trip functionality (e.g. "What is my destination?"), and generic multimodal interaction with objects (e.g. "What is that?") (Krüger et al., 2004). Interaction that is dependent on a particular map, for example specific 'street' and 'landmark' names, and the rule-grammars providing for individualized interaction with landmarks, are generated each time a new map is loaded. This has the advantage of keeping the number of active words and utterances to a minimum and thus increasing the likelihood of speech being correctly recognized. To further improve speech recognition in the BPN application, three separate dynamic grammars are employed. The first grammar allows for interaction with type identifiers (e.g. landmark type identifiers: "What is the name of this *sculpture* <gesture>?"), while the second allows for interaction with name identifiers (e.g. street and landmark names: "Describe the *Richard-Serra sculpture*", or "Take me to *Stuhlsatzenhausweg*"). The third grammar type defines interactions that may take place with a specific landmark, in effect allowing a user to find out information similar to what one might expect in a tourist pamphlet, like detailed descriptions, opening hours, and the cost of entrance (e.g. "What are the opening hours?"). Figure 2.12C in chapter 2 shows these three types of dynamic grammar for the BPN, along with the graphical ability to activate and deactivate the different grammar types (see bottom toolbar in figure 2.12A).

Figure 4.2 illustrates how the functionality in the BPN is entirely accessible via the modality of speech. This includes functionality on selecting trips and encompassed routes, and interaction with map referents. The grammars '1_main', '2_mfType', '3_mfName', and '4_mfInteraction' can all be activated at the same time. The effectiveness of speech-only interaction in the BPN has also been studied in field trials conducted on a 97% sight-impaired student. These studies showed that the speech recognition and synthesis provided by the BPN system was adequate for navigation and exploration, but that the system as a whole failed to provide distance and orientation information that was precise enough for the user. In particular, location information based on GPS can be up to 30m inaccurate, a distance which in the worst-case could place a pedestrian on the incorrect side of a building. This inaccuracy fell short of the student's desire of being told directional and meter information (accurate to within centimetres) on individual building entrances. In comparison to car navigation where map-matching techniques are used to map cars to a street, pedestrians quite frequently leave pedestrian paths, for example when cutting across fields.

⁴CMU Sphinx, <http://www.speech.cs.cmu.edu/sphinx/>

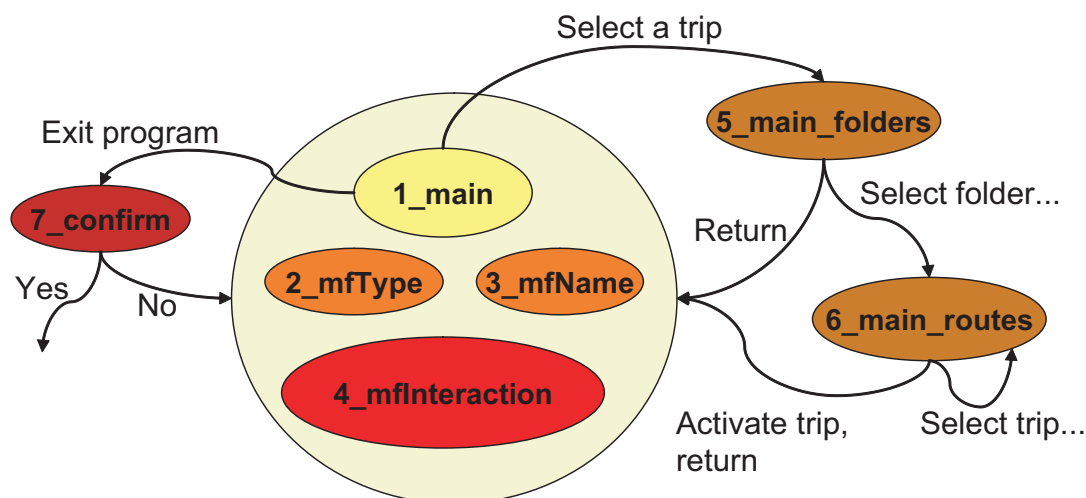


Figure 4.2: States and transitions of the BPN finite-state rule-grammars. Transitions like ‘Select trip’ represent utterances that the user would speak to move between particular states (translated into English).

Similar to the BPN, the MSA also comprises both static and dynamic rule-grammars. The static grammars cover program control (e.g. “What can I say?”, “Next/previous page”, “Connect with shelf”), while the dynamic grammars are derived from the product types located on a particular shelf that the user has synchronized with (e.g. ‘camera’ or ‘language technology’). The SQL⁵ database of products is stored on a remote server, and upon request, products of a particular type and their associated grammars are transmitted to the mobile device via a TCP/IP connection. These grammars are created by querying the SQL database and are sent to the mobile device in XML format. These XML files also contain grammar information for the other modalities like handwriting. Figure 4.3 illustrates the type of information transmitted to the mobile device when the user synchronizes with a shelf containing objects of type ‘digital_camera’. Figure 4.3A shows the XML data container containing product attributes and values, while figure 4.3B shows the associated speech grammars for products of type ‘digital_camera’, and figure 4.3C shows an additional speech grammar that is used when interacting with anthropomorphized objects (see section 4.3). Another interesting feature in the MSA is that a user is able to select whether he/she wishes to speak out just the keywords when querying a product (e.g. “price”, “megapixels”, “optical zoom”), complete sentences (e.g. “What is the price?”), or a combination of both (e.g. sometimes providing just keywords while at other times providing complete sentences). The highest levels of speech accuracy are however achieved when only the option to speak out complete sentences is selected. This is partly due to the increased uniqueness that exists for longer phrases and also to the fact that most recognition errors occur due to the beginning of a spoken utterance being accidentally truncated (e.g. when a user speaks before the recognizer has had time to activate itself).

In the BPN application, the rule-grammars contain around 100-150 unique words depending on the number of landmarks present in the currently loaded map and the complexity of any associated landmark interaction grammars. The MSA application also contains around 100-150

⁵SQL: Structured Query Language

```

<?xml version="1.0" encoding="ISO-8859-1" ?>
- <GETACTIVECONTAINER>
- <CONTAINER type="shelf" name="product shelf" id="00">
+ <OBJECT>
+ <OBJECT>
+ <OBJECT>
- <OBJECT>
  <ATTR id="id" value="54" />
  <ATTR id="type" value="digital_camera" />
  <ATTR id="name" value="PowerShotG5" />
  <ATTR id="price" value="799" />
  <ATTR id="image" value="PowerShotG5.gif" />
  <ATTR id="fullname" value="PowerShot G5" />
  <ATTR id="mega_pixels" value="5 megapixels" />
  <ATTR id="optical_zoom" value="4 Optical Zoom" />
  <ATTR id="digital_zoom" value="3.2 Digital Zoom" />
  <ATTR id="focal_length" value="28 to 200mm" />
  <ATTR id="f_stop" value="2.4 to 3.5 f-stop" />
  <ATTR id="self_timer" value="2 or 10 seconds" />
  <ATTR id="wireless_control" value="Yes" />
  ... ..
+ <OBJECT>
+ <OBJECT>
+ <OBJECT>
+ <OBJECT>
</CONTAINER>
</GETACTIVECONTAINER>

```

A) Data container

```

<?xml version="1.0" encoding="UTF-8" ?>
- <PRODUCTGRAMMAR grammar="digital_camera">
+ <RULES>
+ <COMPARE>
- <PRODUCT>
  <ATTR id="fullname"
  handwritingvalue="name"
  speechvalue_short="name"
  speechvalue_long="What is the name of
  <product>" />
  <ATTR id="price" handwritingvalue="price"
  speechvalue_short="price"
  speechvalue_long="What is the price of
  <product>" />
  <ATTR id="mega_pixels"
  handwritingvalue="mega pixels"
  speechvalue_short="mega pixels"
  speechvalue_long="How many mega
  pixels does <product> have" />
  <ATTR id="lcd_monitor"
  handwritingvalue="lcd monitor"
  speechvalue_short="l c d monitor"
  speechvalue_long="Does <product> have
  an l c d monitor" />
  <ATTR id="description_long"
  handwritingvalue="tell me about"
  speechvalue_short="Tell me about"
  speechvalue_long="Tell me about
  <product>" />
  ... ..

```

Product grammar
(digital cameras)

```

<?xml version="1.0" encoding="UTF-8" ?>
- <PRODUCTGRAMMAR grammar="digital_camera">
+ <RULES>
+ <COMPARE>
- <PRODUCT>
  <ATTR id="fullname"
  handwritingvalue="name"
  speechvalue_short="name"
  speechvalue_long="what is your name" />
  <ATTR id="price" handwritingvalue="price"
  speechvalue_short="price"
  speechvalue_long="what is your price" />
  ... ..

```

Product grammar,
talking objects
(digital cameras)

Figure 4.3: Data sources represented in XML in the MSA showing A) the data container representing a shelf of products, B) the accompanying digital camera product grammar, and C) the additional talking objects grammar for digital camera products.

unique words depending on the number of products and product attributes that are covered by the grammar. For the product type ‘digital_camera’, the database contains information on 25 attributes for 13 different camera products. The attributes for digital cameras include: brand, name, price, megapixels, image resolution, image formats, LCD monitor, optical zoom, digital zoom, focal length, f-stop, shutter speed, shooting modes, self timer, wireless control, photos per second, photo effects, storage media, movie resolution, movie formats, shooting capacity, weight, colours, description, and accessories. In (Cole et al., 1998), vocabularies of under 20 words are considered small while vocabularies of over 20,000 words are considered large. A vocabulary covering street navigation data for the whole of a country would undoubtedly strain the capability of a current state-of-the-art PDA device if it were embedded locally on the device. In figure 2.16, two different methods in which a user can access the available grammars (not just covering the modality of speech) are illustrated. In particular, figure 2.16A and figure 2.16B show two lines of scrolling visual ‘What-Can-I-Say’ text, and figure 2.16C shows a HTML page outlining important keywords in the grammars. Due to limited display space on mobile devices, only keywords (rather than complete phrases) are available to the user to guide them in their queries. A third means of determining the system’s capabilities is via the modality of speech. In the BPN, the application grammars are available in the form of speech output and as HTML.

Figure 4.4A illustrates the use of speech-only input in the MSA for selecting i) a particular object, ii) a particular feature, and iii) both a feature and an object. Figure 4.4B shows the user pressing the button on the mobile device that is used to start (and stop) speech recognition in the MSA and BPN applications. There are three different configurations available to the user in the MSA/BPN to activate and deactivate the speech recognizer. These are known as ‘always listening’ (AL), ‘push to activate’ (P2A), and ‘push to talk’ (P2T). In the always listening mode, a user need neither press a button to start or stop speech. This mode is however not suitable to mobile scenarios like shopping and outdoor navigation because of the high levels of background

noise which can cause frequent engine misfire (i.e. when background noise and especially words spoken by passersby are mistaken for the user's voice). P2A and P2T are the two modes commonly used in the MSA and BPN applications. While P2A only requires the user to manually start the recognizer (the recognizer stops when it detects a period of silence), P2T requires the user to both manually start and stop the recognizer. The use of a physical button that provides tactile feedback to the user is considered more adequate in mobile scenarios than visual touch screen buttons that one can not feel (often denoted as tap-to-speak interaction), as used in systems like (Oviatt, 2002). This is because users may not always have all modalities available to them, especially if they are also multitasking, thus for example a user may not be able to first look at the display in order to determine where to press the button in order to then speak to the recognizer.

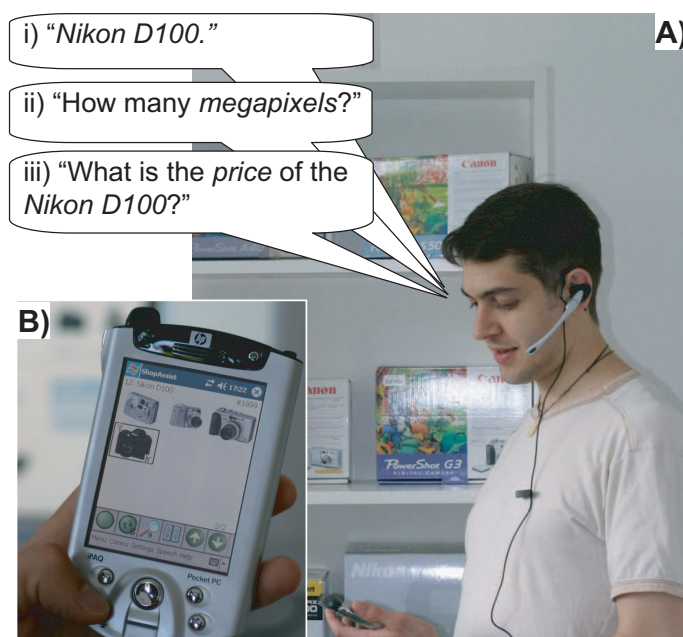


Figure 4.4: The use of speech-only input in the MSA for selecting A) an object (i), a feature (ii), and both a feature and an object (iii). B) shows the user pressing the button that is used to start (and stop) speech recognition in the MSA and BPN applications.

4.1.1.2 Handwriting Recognition

The handwriting character recognizer employed on the mobile PDA device in the MSA application is that of Microsoft's Transcriber⁶. Transcriber is a digital character recognizer that is capable of recognizing handprinted text on the fly. It is one of the primary input communication modes found on modern PDAs. Handwriting input is provided by a user through stylus interaction on the mobile device's display, and recognition of the handprint characters takes into account the order, speed, and direction of individual line segments as they are provided by the user. Similar to large-vocabulary speaker-independent speech recognition systems operating in a clean environment, which have a reported recognition accuracy of around 80-90% (Wikipedia, 2006f), the

⁶Microsoft Transcriber, <http://www.microsoft.com/windowsmobile/downloads/transcriber.msp>

recognition rates for neat, clean handprinted characters have also been reported to be around 80-90% (Wikipedia, 2006c). These recognition rates for handwriting do not however entail the use of cursive print which has much lower rates of recognition accuracy than individual handprinted characters. From the usability studies conducted as part of this dissertation (see chapter 6), it was found that users often combine both forms of handwriting, starting a word or phrase in handprint and ending it in cursive print. An important difference between the commercially available embedded speech and handwriting recognizers used in the MSA/BPN is that the speech recognizer is constrained by its rule-grammars to understand only a limited number of phrases and words. In comparison, the embedded handprint recognizer recognizes individual characters without regard for the actual words and phrases that they may form. It is for this reason that a separate module was written in the MSA to coexist with the implemented handprint recognizer. This module performs character and word matching algorithms on the output provided by the Transcriber software, and it also takes advantage of information contained within given constrained handwriting grammars. Similar to speech recognition, these grammars are derived from the product database and are generated based on product types located in a shelf that a user is currently synchronized with. Knowledge-based character recognition as a suitable means to improving OCR accuracy is described in detail in (Dengel et al., 1997). As shown in figure 4.3B and figure 4.3C, the XML grammar files downloaded onto the mobile device contain grammar information not just for speech (short and long utterances, and talking object interaction), but also for handwriting. This XML file is reformatted on the mobile device to correspond with the required recognizer grammar formats like SRCL, and the actual product names are at this point also integrated into the grammars based on information contained in the data container (figure 4.3A).

Figure 4.5 illustrates the use of handwriting-only input in the MSA for selecting A) a particular object, B) a particular feature, and C) both a feature and an object. Figure 4.5D shows the graphical button on the mobile device's display used to activate the handwriting recognizer (inactive button on the top, active button on the bottom). When activated, all 'line' interaction on the display (but not single point interaction) is redirected to the handwriting recognizer for recognition.



Figure 4.5: The use of handwriting-only input in the MSA for selecting A) an object, B) a feature, and C) both a feature and an object. D) shows the graphical button used to activate the handwriting recognizer in its deactivated (top) and activated (bottom) state.

4.1.1.3 Tangible Interaction and Gesture Recognition

Ullmer and Ishii (2001) describe how the last decade has seen a wave of new research into ways to link the physical and digital worlds, and how this common goal has created a range of new research themes like: mixed and augmented reality (Azuma, Bimber, & Sato, 2005), ubiquitous computing (Beigl et al., 2005), and wearable computing (Rhodes & Mase, 2005). *Tangible User Interfaces* (TUIs) (Ishii & Ullmer, 1997) build on the ‘direct manipulation’ work defined by Shneiderman (1992) and the work on ‘graspable user interfaces’ defined by Fitzmaurice, Ishii, and Buxton (1995). TUIs are interfaces that give physical form to digital information, thus making digital bits directly manipulable and perceptible. One common approach to designing TUIs includes the use of ‘found objects’ already existing in an environment and embedding these with position sensors or ID tags. One of the more renowned examples of a TUI is the “Marble Answering Machine” by Durrell Bishop (Crampton-Smith, 1995), in which a marble represents a single message left on the answering machine, and picking a marble up plays back the associated message. Another example of a TUI is given in (Butz & Schmitz, 2005), where an everyday beer coaster is instrumented with both a gravity sensor that can detect the motion of lifting up a drink and a pressure sensor to detect the weight of the drink. One goal of the beer coaster is to support a range of entertainment activities in pubs, and the authors make a point of preserving the coasters original functionality like absorbing liquid and providing advertising space.

Weiser (1991) states that “the most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.” Furthering this, Norman (1998) states that one should “design the tool to fit the task so well that the tool becomes part of the task, feeling like a natural extension of the work, a natural extension of the person”. It is based on this guideline that extra-gestures in the MSA and BPN applications were developed. In the BPN for example, an intuitive one-to-one mapping exists between the physical building and landmark objects in the surrounding environment and elements of digital information on the map shown on the mobile device’s display. In the MSA, a similar one-to-one mapping exists between the shopping items resting on a shelf in the real-world and their digital product representations shown on the mobile device’s display.

Two distinct types of selection-gesture are used in the MSA/BPN, and these are given the terms intra-gesture and extra-gesture. In (Ullmer & Ishii, 2001) the terms digital representation and physical representation are used to represent similar functions. In particular, ‘digital representation’ refers to computationally mediated displays that are perceptually observed in the world, but are not physically embodied and thus intangible in form, e.g. the pixels on a screen or the audio from a speaker. Intra-gesture is used for interacting with digital representations, namely through a PDA’s display. In comparison, ‘physical representation’ refers to information that is embodied in tangible form, e.g. physical chess pieces and chess boards. Physical representations are therefore interacted with through extra-gesture.

Table 4.1 illustrates the range of gestures that are available in the BPN and MSA applications. *Gesture* is a broad term defined in common usage as “motions of the limbs or body, used as a means of expression” (Merriam-Webster, 1998). In (Kendon, Drew, Goodwin, Gumperz, & Schiffrin, 1990), gestures are described to range from pointing at a person to draw their attention, to conveying information about space and temporal characteristics. Current research on gesture can be divided into fields like human body motion recognition (including facial expressions and hand movements) (Wahlster, 2002b; Baudel & Beaudouin-Lafon, 1993), pen and mouse based recognition (Pastel & Skalsky, 2004) and sign language. For an extended summary on

state-of-the-art research into gesture, Kipp (2003) provides a categorization based on “research aim (analysis, recognition, generation), method (linguistic, psychological, engineering), examined conversational domain (storytelling, psychotherapy, talk-show), and observational conditions (laboratory, field data, TV recordings)”. In table 4.1, gestures are classified as ‘selection-gestures’ because the majority of the gestures in the MSA/BPN are used to select semantic referents like buildings and landmarks in the navigation scenario (e.g. ‘Richard-Serra Sculpture’) and shopping products and their feature attributes in the shopping scenario (e.g. ‘PowerShot S50’ and ‘price’).

Selection Gestures	MSA	BPN
Intra-gesture (on-device)	point, slide	point, slide
Extra-gesture (off-device)	point, pickup, putdown	point

Table 4.1: Selection gestures available in the MSA/BPN.

Intra-gesture: *Intra-gestures* occur when a user interacts with the mobile device’s graphical display by pointing at referents or by drawing symbols. Because this type of gesture is closely associated to the mobile device and in particular the device’s display, this type of gesture is also referred to in this dissertation as ‘on-device interaction’ (Wasinger & Krüger, 2005). Intra-gestures are provided in the form of stylus or finger input and can be of the type ‘point’ (i.e. intra-point) or ‘slide’ (i.e. intra-slide). In a navigation context, point gestures are used to interact with buildings and landmarks, e.g. “What is the name of this building <Gesture-point>?”. Slide gestures are, in contrast, used to interact with streets, which are often well suited to such gestures due to their narrow and long form, e.g. “What is the name of this street <Gesture-slide>?”.

Usability study findings on the BPN indicate that when presented with limited display space (pixel resolutions of 240x320 are not uncommon) and maps containing a dense number of referents, the use of specialized gestures to distinguish between map referent types (e.g. buildings, streets) can lead to improved rates of referent disambiguation. Point and slide gestures as used in the BPN application are illustrated in figure 4.6 (Wasinger et al., 2003a). In a shopping context, point gestures are used to select products as shown in figure 4.7A and product feature attributes like ‘movie resolution’ as shown in figures 4.7B and 4.7C (Wasinger & Krüger, 2004). Selecting product feature attributes is made possible through the use of a visual What-Can-I-Say (WCIS) scrolling text bar as shown in figure 4.7B. The ability to carry out complex interactions in each communication mode (i.e. speech, handwriting, and gesture) and to access the entire functionality of a system is one of the novel aspects presented in this dissertation. A slide-gesture is also available in the MSA application, and this is used specifically to increase and decrease the speed of the visual-WCIS text, from 0 chars/sec up to and including 50 chars/sec (default = 15 chars/sec). The space allocated to the visual-WCIS text is two lines, each able to present around 50 characters at any one time. The upper of these two lines displays keywords relating to the static grammars (i.e. program control like what can i say, toggle view, and next page), while the lower of the lines displays keywords for the product-specific grammars (e.g. price, megapixels, and optical zoom for digital cameras). Refreshing the visual-WCIS scroll bar occurs every 200ms, i.e. 5 times per second.

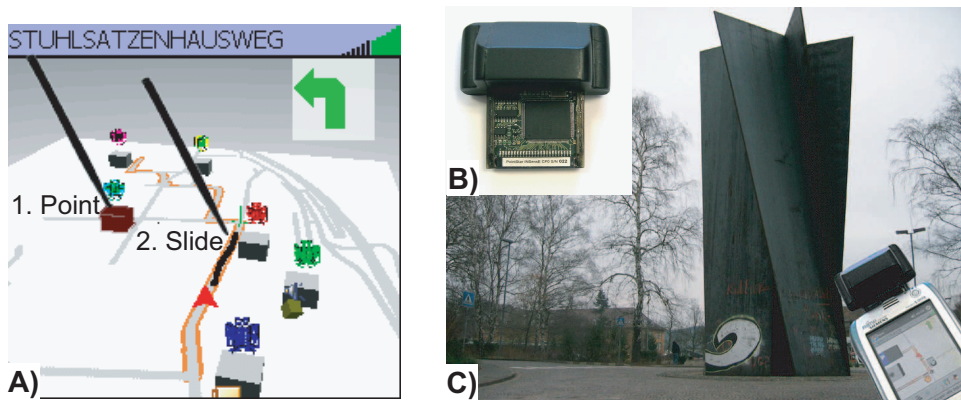


Figure 4.6: Selection gestures in the BPN application illustrating A) intra-point and intra-slide gestures, and C) an extra-point gesture. B) shows the magnetic compass/attitude sensor array device used for extra-gesture interaction.

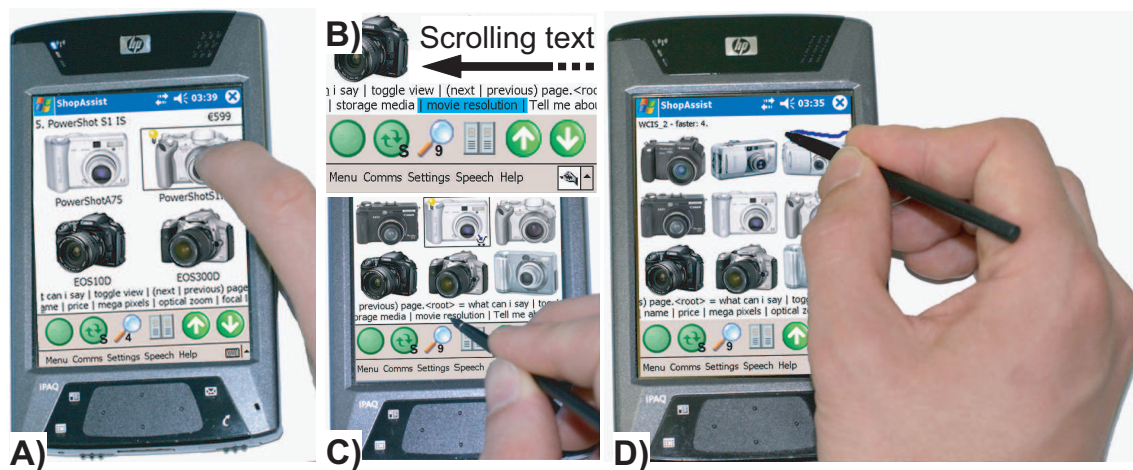


Figure 4.7: Intra-gestures in the MSA application showing A) intra-point object selection, C) intra-point feature selection, and D) an intra-slide gesture used to increase and decrease the speed of the visual What-Can-I-Say scrolling text bar shown in B).

Recognition of the point and slide gestures in the BPN takes place by mapping the 2D screen coordinates (x, y) onto the underlying data objects on the 3D map graphic. The visual component used in the BPN application utilizes a VRML browser (Virtual Reality Markup Language) to render map objects like buildings and landmarks. This graphics engine is called Cortona and was created specifically for mobile PDA devices by the company Parallel Graphics⁷. One challenging aspect in resolving intra-point gestures in the BPN application was that aside from requiring 2D screen coordinates to be mapped to the 3D scene space displayed on the PDA, the map graphic changes constantly to reflect the user's current position, and the user can also toggle between different map views like birds-eye and egocentric (see figure 2.11A and 2.11B). These program features require the map representation models to be constantly updated for point and slide gestures to be properly interpreted. In the MSA, recognition of point and slide gestures is done in a similar fashion, i.e. by mapping (x,y) screen coordinates onto the underlying product set, which can be displayed as a set of nine, four, or one product. Intra-gestures in the MSA do not specifically cater for a user's natural pointing behaviour, and this can be seen in that a user must point on (rather than near to) a product for it to be selected. Supporting natural pointing behaviour is not considered technically difficult, and such behaviour may in fact differ to the interaction that users have grown accustomed to through the use of point-and-click interfaces found in many applications.

Extra-gesture: *Extra-gestures* occur when a user interacts with physical world objects, by pointing at them or by physically handling them. Because this type of selection-gesture deals specifically with real-world tangible objects, it is also referred to as 'off-device interaction'. Extra-gestures can be of type 'point' (i.e. extra-point), 'pickup' (i.e. extra-pickup), or 'putdown' (i.e. extra-putdown). Similar to intra-point gestures, extra-point gestures are used in a navigational context to select buildings and landmarks, only rather than these objects being displayed on the mobile device's display, they are real physical entities found in the user's surrounding environment, as shown in figure 4.6C. In a shopping context, extra-point gestures refer to the user's ability to select shopping objects like cameras, by pointing at them in the real-world. Pickup and putdown gestures occur in a shopping context when a user picks a product up from the shelf or shopping trolley, and puts a product back down onto the shelf or into the shopping trolley. As shown in figure 4.8A and figure 4.8E, these three extra-gestures cover both 'proximal interaction', in which a user must touch the real-world object (i.e. extra-pickup, extra-putdown), and 'distal interaction', in which the user is able to point at the real-world object from a distance (i.e. extra-point) (Wasinger et al., 2005).

Recognition of the extra-gestures in the MSA/BPN is based on a variety of sensing and vision technologies. For both the BPN and MSA applications, extra-point gestures work on the assumption that the PDA device is used as a pointing stick; thus the user selects a referent in the real-world by pointing at it with the PDA device. In the BPN, extra-point gestures are detected through a magnetic compass combined with an attitude sensor array, together capable of providing real-time direction and orientation information about the PDA. The device responsible for this is a prototype CF-card from the company PointStar called INSense⁸. Directional information on the user (and velocity) is also attainable from a Bluetooth GPS device, but requires the user to be in motion for this to be accurate because it is based on the analysis of consecutive geo-coordinate locations of the user. On determining the PDA's orientation, the direction of the extra-point gesture can be

⁷Parallel Graphics, <http://www.parallelgraphics.com/products/cortona/>

⁸Pointstar, <http://www.pointstar.dk>

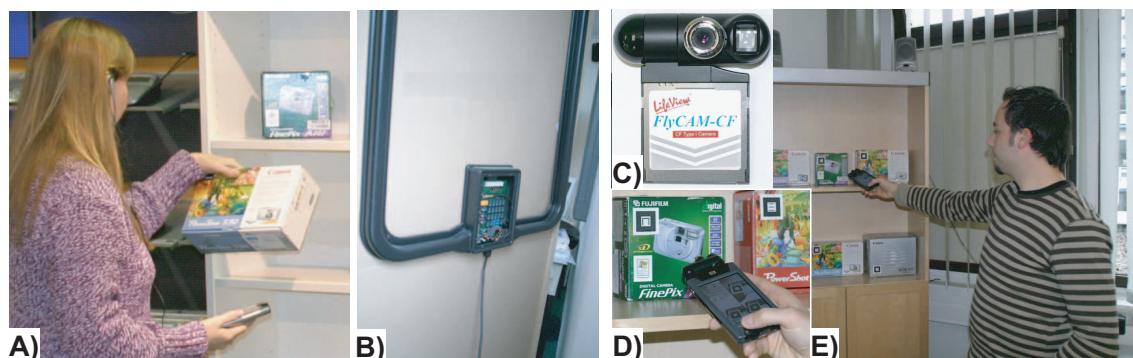


Figure 4.8: Extra-gestures in the MSA application showing A) an extra-pickup/extra-putdown gesture, and E) an extra-point gesture. B) shows the RFID antenna used to recognize objects taken out of and put back in the shelf, and D) and C) show a close-up picture of the extra-point accompanied with the actual CF-camera device underlying this type of interaction.

calculated and mapped to known locations of physical objects in the real-world, as described in (Krüger et al., 2004). In the MSA, extra-point gestures are detected through the recognition of Augmented Reality (AR) markers that are placed on the shopping products. AR marker recognition requires the sense of vision to function, and in the MSA application a 1.3 megapixel CF-card camera designed specifically for PDA devices and called FlyCAM-CF⁹ is used. Similar to the navigational context, a user points the PDA device in the direction of the shopping product and then presses a button on the PDA to capture and analyse the photo for encompassed AR tags.

The accuracy of extra-point gestures in the BPN application during outdoor navigation and exploration is influenced by the number of objects in the same direction relative to the user (i.e. one in front of the other, but different heights). The magnetic compass can also be influenced by metal objects (including walls of buildings) and is thus not reliable in an indoor environment. Similarly, GPS is also unreliable when inside buildings due to the obstructed line-of-sight to at least three satellites that is required for user-positioning. Extra-point gestures in the shopping context are affected by the focal length of the camera that is used. In particular, because the camera's focus is only manually adjustable and has no ability to auto-focus, the extra-point gestures are limited to a distance of around 30cm to 1m. Pointing gestures conducted over distances of more than one metre have a reduced accuracy due to the comparative resolution of the AR tag to the rest of the picture (which is only 640x480 pixels in resolution).

Extra-pickup and extra-putdown gestures in the MSA application are recognized through the use of RFID sensing technology. The critical elements in such a setup are the RFID readers, antennas, and ID tags. In the given shopping context, each shelf is instrumented with an RFID antenna connected to an RFID reader, and each shopping product is instrumented with a passive RFID tag¹⁰. In the MSA configuration, products that are placed in or out of a shelf, or in or out of a shopping trolley, have their RFID tags recognized by a server and sent to the mobile device via a wireless LAN connection. The employed RFID technology was created by the company FEIG Electronic¹¹.

⁹LifeView FlyCAM-CF, http://www.lifeview.com.tw/html/products/pc_camera/flycam_cf.htm

¹⁰In contrast to active RFID tags, passive RFID tags do not require a battery

¹¹FEIG Electronic, <http://www.feig.de>

One novel aspect of the MSA application is that RFID technology is implemented with the goal of benefiting the customer. RFID instrumented environments are very often designed to benefit the retailer rather than the customer through improved inventory management and inventory tracking (Wasinger & Wahlster, 2006) (see also section 3.2.1.1). In the MSA/BPN however, user benefits are provided in the form of new and novel interaction techniques that allow for ‘hands-on’ comparison shopping, cross-selling, and information retrieval. The interaction furthermore conforms to traditional shopping practices, which means that customers do not need training on how to use extra-gestures that they are already accustomed to. In (Newcomb et al., 2003), a series of design guidelines for a PDA based shopping assistant are described, and one of the points that is made is that shoppers often use their hands to touch the products; a feature that this dissertation has tried to incorporate into the design of the MSA/BPN mobile applications.

Another novel aspect in the MSA/BPN is that intra- and extra-gestures can be combined, as shown by the speech-gesture utterance: “Compare this camera <GI> to this one <GE>” (see figure 2.17). When such digital, real, and mixed interaction is possible, there is also a need for the layout of objects to be easily reordered. For example, camera objects shown on the mobile device’s display need not always align with the layout of the products on a particular real-world shelf. This misalignment can occur as a result of some products being out of stock, or perhaps due to a user sorting the digital object representations by feature attributes like price or optical zoom. To compensate for this type of disorientation, a spotlight service, as illustrated in figure 4.30, was integrated to find objects in the real-world that were referred to via their digital representation.

4.1.2 Confidence Values and Confidence Scoring

Confidence scoring refers to the process of attaching likelihoods to recognition results in an attempt to measure the certainty of finding a correct match to a user’s input. For each of the modalities within the MSA/BPN (speech, handwriting, and gesture), an N-best list of results is generated each time a user interacts with the system. These results are assigned confidence values (Cf) ranging between Cf=0.0 and Cf=1.0. ‘N’ in the case of the MSA/BPN is equal to three, meaning that the N-best list contains the three most likely results for a given modality. N-best lists play an essential role in the disambiguation of multimodal input, which might for example be semantically overlapped and conflicting (destructive) or semantically overlapped and non-conflicting (constructive). In fusing multiple N-best lists, the goal is to decrease the overall certainty of destructive combinations and to increase the overall certainty of constructive combinations. As described in section 2.2.4, the benefit of late semantic fusion, like that which occurs in the MSA/BPN, is that information otherwise discarded by a recognizer can be stored and kept till a later stage of processing, at which time information accumulated over a spread of different modalities may contribute to more reliable results. Confidence scoring within mobile multimodal systems is also advantageous because the methods used in calculating the confidence values are often specific to the individual modalities, and are thus affected differently by noise, be that noise in the form of sound (which would affect speech), motion (affecting handwriting), or the density of referents in a spatial plane (affecting gesture). This section is an extension to the work published in (Wasinger et al., 2005) and describes how the confidence values in the MSA/BPN application are calculated for each of the modalities speech, handwriting, and gesture. Particular focus is placed on the generation of confidence values in the MSA application because this application has a broader range of communication modes, including a variety of gestures and also handwriting. The generation of confidence values for modes like speech and gesture, which occur in both the MSA and the BPN, will only

be described in detail for the MSA system. The calculation of confidence values for these modes in the BPN system is similar, but based on preliminary versions of the approach used in the MSA (Wasinger, Stahl, & Krüger, 2003b).

Speech: Speech confidence values are generated by matching a user's spoken utterance to a sequence of word hypothesis defined in a given language model or vocabulary. In the MSA/BPN, the vocabulary consists of word-to-phoneme mappings and rule-grammars that define sequences of words to be recognized. The generation of N-best speech results and associated confidence values in the MSA/BPN is a functionality provided by the underlying commercial speech engine. State-of-the art speech engines that are designed to run embedded on mobile devices are limited in resources, and one of the consequences of this is that confidence values are only allocated per utterance. Speech engines with greater disposable resources are capable of returning confidence values for individual words in a recognized utterance's word lattice (Cole et al., 1998). This limitation means for example that if both a feature and an object are provided in a single spoken utterance (e.g. "What is the *price* of the *PowerShot S50*?"), the recognizer will assign the same confidence value to both semantic constituents.

Handwriting: Unlike the generation of confidence values for the modality of speech, the employed commercial character recognizer provides no API to gain access to the handwriting N-best lists and associated confidence values. The generation of N-best lists for handwriting input in the MSA is a two stage process. As shown in figure 4.9, written input (i), is first sent to the character recognizer which is capable of recognizing individual characters (ii), based on constraints like the order, speed, and direction of individual line segments provided during real-time user interaction. With typical character recognition rates of around 80% (Chellapilla, Larson, Simard, & Czerwinski, 2005), words of lengths greater than 5 characters stand to have at least one character incorrectly recognized (i.e. $0.2 \text{ error rate} \times 5 \text{ characters} = 1 \text{ in } 5 \text{ incorrect characters}$). To minimize the error rate, a module was designed to match the set of recognized characters with the entries defined in the MSA application's handwriting grammars (iii), based on calculated confidence values (iv). These grammars consist of feature keywords like 'optical zoom' and 'megapixels', object keywords like 'PowerShot S50' and 'PowerShot S70', and phrases consisting of both features and objects, e.g. 'price PowerShot S50' and 'PowerShot S50 movie resolution'.

The character-to-word mappings in the MSA are based on a character matching algorithm, in which the characters in a user's handwriting input (e.g. 'optinlzrein', see figure 4.9) are compared to the characters in each entry in the handwriting grammars (e.g. 'name' and 'optical zoom'). Each character that appears in both the user's input and the currently being compared grammar entry increases the confidence value for that particular grammar entry. To avoid multiple matches on the same character (e.g. 'optinlzrein' and 'image resolution' where two 'i's in the user's input exist compared to only one in the grammar entry), any characters already matched in the grammar entries are temporarily removed. Because long grammar entries like 'optical zoom' contain more characters and thus stand a greater chance of having more correctly matched characters, grammar entries that are either too long or too short when compared to the user's input, i.e. greater than or less than the total length plus or minus three characters respectively, are immediately filtered out. A positive bias of $C_f = C_f + 0.1$ is also given to grammar entries that have the same starting character as the user's input, and the reasoning behind this bias is that it is assumed user's will take more care writing the first character compared to subsequent and final characters. If the bias results in

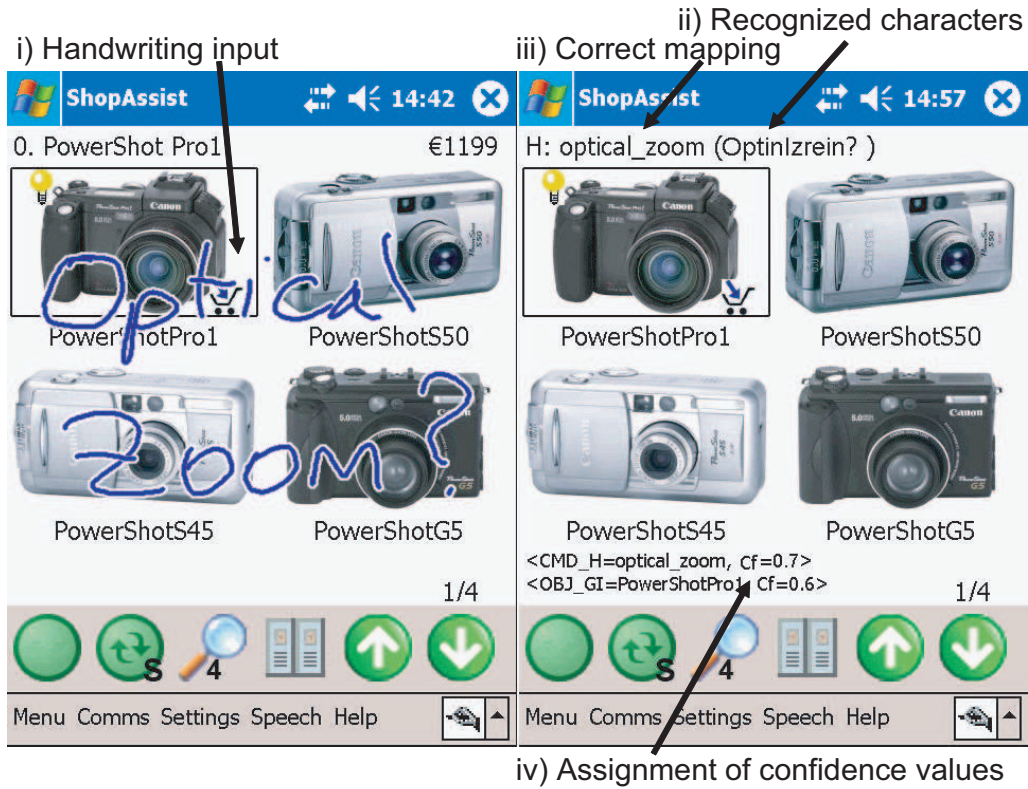


Figure 4.9: An example of how handwriting input (i) is recognized by the character recognizer (ii), and then mapped to a valid grammar entry (iii) and given a confidence value (iv) (Wasinger et al., 2004).

a confidence value exceeding 1.0, it is rounded back down to conform to the range of 0.0 to 1.0. During the character matching process, case, punctuation, and white-space are all disregarded. As shown by Equation 4.1, the total number of characters in the grammar entry, $T_{GrammarEntry}$, must fall within the total number of characters in the user's input plus or minus three characters:

$$T_{UserInput} + 3 \geq T_{GrammarEntry} \geq T_{UserInput} - 3 \quad (4.1)$$

The 3-best list of results generated for the handwriting confidence values $Cf_{Hn=1}$, $Cf_{Hn=2}$, and $Cf_{Hn=3}$ can be expressed by the formula:

$$Cf_{Hn} = \frac{M_n}{T_{UserInput}} + F \quad (4.2)$$

where,

Cf_{Hn} = The handwriting confidence value for each n^{th} best result.

M_n = The number of correctly matched characters in the grammar entry for each n^{th} best result.

$T_{UserInput}$ = The total number of characters in the user's input.

F = The conditional first letter bias, set to 0.1 when the first letter is correctly recognized, otherwise set to 0.0.

The sliding character match is demonstrated by the example illustrated in table 4.2, where the grammar entry ‘optical zoom’ (Cf=0.65) is determined to be the best match for the user input ‘Optinlzrein’.

Grammar Entry	No. of Chars	No. of Matched Chars	First-letter Bias (+0.1)	Confidence Value, Cf
brand	5	—	—	0.0
name	4	—	—	0.0
megapixels	10	4 (p,i,e,l)	No	0.36
optical zoom	11	6 (o,p,t,i,l,z)	Yes	0.65
focal length	11	5 (o,l,e,n,t)	No	0.45
f stop	5	—	—	0.0
...
PowerShot Pro1	13	4 (p,o,r,t)	No	0.36
PowerShot S50	12	4 (p,o,r,t)	No	0.36
PowerShot S45	12	4 (p,o,r,t)	No	0.36
...
brand PowerShot Pro1	18	—	—	0.0
...
PowerShot S50 brand	17	—	—	0.0
...

Table 4.2: An example of the sliding character match algorithm used for the input ‘Optinlzrein’ (11 chars long). The grammar entries denoted by ‘—’ are immediately discounted due to their lengths being either too long (i.e. >14) or too short (i.e. <8) when compared to the user’s input.

Intra-gesture: Intra-gestures in the MSA occur when a user interacts with the mobile device’s graphical display, for example by pointing at feature and object referents. Object referents refer to camera products such as the ‘EOS 300D’ and are available to the user in the form of graphical images, while feature referents refer to keywords such as ‘price’ and are available to the user in the form of scrolling text displayed on the PDA (see figure 2.16).

Depending on the user’s current viewing mode, nine, four, two, or one object rectangles are displayed on the PDA’s display (see figure 2.15). As shown in figure 4.10A, confidence values for object referents are generated by drawing a rectangle around the user’s point-coordinates equal in size to the graphical image rectangles currently being displayed. The intersection between this active area and each of the image rectangles is then calculated and used as the intra-gesture confidence value ($Cf_{GI_n=1}$, $Cf_{GI_n=2}$, and $Cf_{GI_n=3}$) such that:

$$Cf_{GI_n} = AA \cap IR_n \quad (4.3)$$

where,

Cf_{GI_n} = The intra-gesture confidence value for each n^{th} best result.

AA = The active area surrounding a user’s x,y coordinate point.

IR_n = The image rectangle used in calculating the n^{th} best result.

Figures 4.10B, 4.10C, and 4.10D show that if the user points to an image rectangle at perfect centre, the rectangles line up and the confidence value is 1.0. If the rectangles only half line up (side by side), the value will be 0.5, and if the user points to a corner, the value will be 0.25. Areas that extend off the display are discounted, as shown in figure 4.10E. The minimum and maximum range for each of the three best results is outlined in table 4.3, where it is shown that the confidence values for the N-best list entries are split over a maximum of four different rectangles together totalling 1.0. For example, if a user selects the centre of an image, N=1 will have a confidence value equal to Cf=1.0 and N=2 and N=3 will both have values equal to Cf=0.0 (see the maximum value for N=1 and the minimum values for N=2 and N=3 in table 4.3).

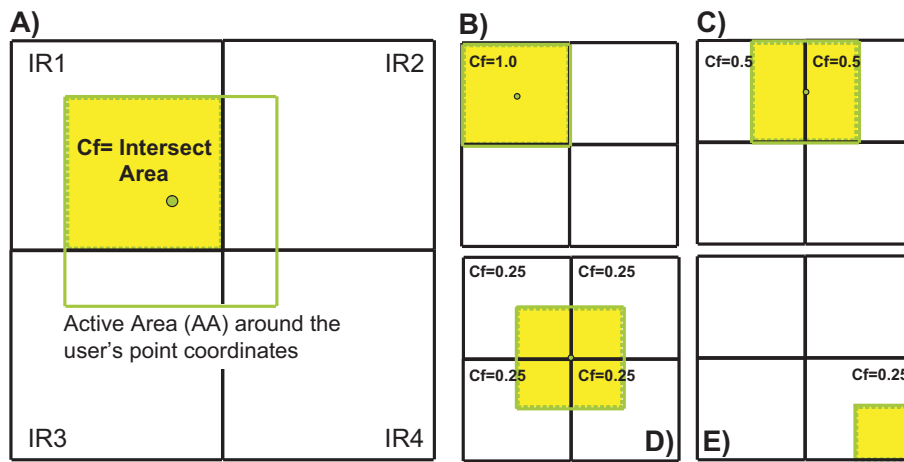


Figure 4.10: Confidence value generation for intra-point object resolution showing A) the intersect between AA and IR, and four example confidence values B), C), D), and E).

Cf_{GI}	Minimum Value	Maximum Value
N=1	0.25	1.00
N=2	0.00	0.50
N=3	0.00	0.25
N=4	0.00	0.25

Table 4.3: Minimum and maximum range of values for the intra-gesture N-best list of confidence values.

Experiments were conducted to determine the benefit of using exponents to remap intra-gesture object confidence values onto the range from 0.0 to 1.0, but although the values then shared the same minimum and maximum range as the modalities of speech and handwriting, the resulting confidence values were seen to less appropriately reflect the actual accuracy of the modality, and this was particularly visible for the lower values like 0.25 which were remapped to the value 0.0. A small study outlined in the following section (see section 4.1.3) shows that intra-gesture is a very accurate modality to begin with, and this is perhaps also reinforced by the relatively small number of images available for selection on the PDA's display at any one time, when compared

with the larger number of object referents that could be spoken or written about at the same point in time.

The selection of features when using the MSA modality of intra-gesture is based on a user pointing at keywords displayed in the lower of the two lines of text making up the visual-WCIS scroll bar at the bottom of the PDA's display, as shown in figure 4.7. The keyword that is selected by the user is assigned the N-best list entry N=1, while N=2 and N=3 are respectively assigned to the keywords left and right of the selected keyword. Confidence value generation for this communication mode is such that the speed of the scrolling text (which moves from right to left on the display) influences the overall confidence of a particular entry. As shown in table 4.4, each character/second speed increase from stationary equates to a drop of 0.02 confidence points, i.e.:

$$Cf_{GI_{n=1}} = 1 - 0.02 * \frac{chars}{sec} \quad (4.4)$$

The value of 0.02 was chosen such that Cf=1.0 at speed 0 (0 chars/sec), and Cf=0.0 at speed 10 (50 chars/sec) where it can be seen that a character scrolls on and off the display within one second. The second- and third-best results in the N-best list are assigned confidence values based on that of the first best result, i.e.:

$$Cf_{GI_{n=2}} = Cf_{GI_{n=3}} = \frac{Cf_{GI_{n=1}}}{2} \quad (4.5)$$

The length of each keyword is not considered in the current implementation, and studies to verify the accuracy of the implemented approach have only been conducted on an earlier version of the algorithm in which N=1 was always assigned Cf=0.8, and N=2 and N=3 were always assigned Cf=0.4 and Cf=0.2 respectively, i.e. without regard for changes in speed (see section 4.1.3).

visual-WCIS Speed	Confidence Value, Cf_{GI}		
	N=1	N=2	N=3
0 (0 chars/sec)	1.0	0.50	0.50
1 (5 chars/sec)	0.9	0.45	0.45
2 (10 chars/sec)	0.8	0.40	0.40
3 (15 chars/sec)	0.7	0.35	0.35
4 (20 chars/sec)	0.6	0.30	0.30
5 (25 chars/sec)	0.5	0.25	0.25
6 (30 chars/sec)	0.4	0.20	0.20
7 (35 chars/sec)	0.3	0.15	0.15
8 (40 chars/sec)	0.2	0.10	0.10
9 (45 chars/sec)	0.1	0.05	0.05
10 (50 chars/sec)	0.0	0.00	0.00

Table 4.4: Confidence value generation for intra-point feature resolution showing the N-best confidence values based on different given scrolling text speeds.

Another type of intra-gesture in the MSA is the slide action that is used to increase and decrease the speed of the visual-WCIS scroll bar, and it is worth mentioning how such an intra-slide gesture is distinguished from an intra-point gesture in the MSA application. The critical factor during this gesture recognition is that a slide gesture is based on a series of unidirectional coordinate

points in comparison to just a single point or several fairly similar-valued points. To demonstrate, when a user performs an intra-gesture, it is first categorized to be either a 'point', a 'slide', or an 'unknown'. If two or less coordinate points exist it is classified as a point action, while if a consecutive number of points greater than five and all with the same horizontal component of direction exist, the action is considered a slide gesture, otherwise the action is considered to be unknown. When an intra-slide gesture is identified, the direction is calculated based on the start and end point coordinates (i.e. left or right), and thus the system can determine whether the user intended for the speed of the visual-WCIS scroll bar to be increased or decreased.

Extra-gesture: Extra-gestures in the MSA are assigned a static confidence value of $C_f=1.0$ for $N=1$, and $C_f=0.0$ for $N=2$ and $N=3$. Supporting this static assignment is the excellent accuracy that was observed during studies into the use of extra-gesture interaction (see section 4.1.3).

In the current implementation of the MSA application, the RFID readers and antennas only permit the detection of an object as being either in or out of a particular container. The advantage of this is that a product can be picked up from one location and put back in another location, without this affecting the accuracy of the extra-pickup and extra-putdown gestures. A second benefit of this approach is that the gestures are identified in real-time. A limitation to this implementation is however that it is not possible to pinpoint the exact location of a product on a shelf. Such location information would for example provide the ability to assign confidence values based on objects that are left, right, above, or below the object that was selected and would additionally make it possible to resolve many locative references based on the descriptors just described. In (Butz et al., 2004; Spassova, Wasinger, Baus, & Krüger, 2005), a fully-implemented module to automatically detect the location of products in the MSA using optical marker recognition is described, however this detection process is not yet available in real-time. Traditional real-world shops circumvent the issue of product location by assigning a set physical position to the product along with a paper placeholder for name and price information.

In summing up this section, the generation of adequate confidence values for use by a multimodal system is still an area of ongoing research. Although speech recognizers are nowadays built on top of a great wealth of statistical data arising through decades of trial-and-error and experience, there is still limited statistical data for determining how best to rate the accuracy of other modalities such as gesture, particularly within constantly changing environment contexts like that of shopping and navigation. When comparing confidence values between different-type and even same-type recognizers (see section 4.2), the confidence weightings generated may also not be easily comparable due to factors like different statistical models used to train the recognizers. One solution as described in section 5.3.3, is to re-weight the confidence values (based on the accuracy of the results) in an attempt to balance out discrepancies between the different recognizers.

4.1.3 Accuracy, Interaction Times, and Scalability

The MSA has been demonstrated at numerous public events including the "DFKI Language Technology Summit" and "Empower Germany" in Saarbrücken in 2004, "Voice Day" in Bonn in 2005, and the "CeBIT 2005" and "CeBIT 2006" exhibition fairs in Hannover, Germany. The CeBIT fair is considered a difficult environment for demonstrating language technology products even at the best of times, firstly because of the significant amount of background noise generated by the many people visiting the fair and secondly due to the technical constraints that typically arise

when setting up and demonstrating real systems in untested environments. However, for exactly these reasons, it was decided to perform a series of rigour-tests on the MSA system during the CeBIT 2006 fair as part of the DFKI/MTI stand¹².

The studies that were conducted were designed to test the accuracy of several modality combinations available in the MSA, based on the digital camera dataset of 13 products, each associated with 12 attributes. Of particular importance was the relationship between accuracy and the confidence that each recognizer assigned to the results, while a second focus was to determine average interaction times needed when using individual modality combinations over a longer period of time.

The data that was gathered originated from a single experimenter, who carried out controlled sequences of interactions. Due to the large number of available modality combinations in the MSA, it was decided to test only a select few of the combinations. These modality combinations were chosen based on the results from two prior usability studies outlined in chapter 6 and entailed those combinations that were rated by users as being preferred and intuitive to use.

Careful attention was placed in testing the MSA in a configuration that allowed for completely complementary interaction. Although only a select few modality combinations were studied, all of the other modality combinations were also functional during the tests. Testing the system under this condition (rather than switching non-used modality combinations off) increases the number of interaction possibilities and the size of the active grammars, which can affect the overall accuracy of a system. Such a configuration is however a prerequisite for systems providing complementary interaction. Unimodal modality combinations like SS and HH were also designed such that feature and object information could be entered either during a single action or during separate actions and in any order (i.e. 'feature' then 'object', and 'object' then 'feature').

Table 4.5 shows the three most preferred modality combinations in the laboratory and real-world studies, as outlined in chapter 6. With the addition of the modality combination HH, these combinations represent the complete set of modality combinations that were rated in the studies as being significantly intuitive. It is also this set of modality combinations that was used as the basis for the study, i.e.: SS, SGE, SGI, HH, HGI, and GIGI.

Ranking	Laboratory Study	Real-world Study
1	SGE	GIGI
2	SS	HGI
3	SGI	SGI

Table 4.5: *The three modality combinations rated by users to be most preferred during laboratory and real-world usability studies.*

Testing was conducted such that each of the 12 feature attributes (e.g. brand, name, price) were combined with each of the 13 products and then tested using each of the 6 modality combinations mentioned in table 4.5. A single round thus totalled $6 \times 12 \times 13 = 936$ interactions (see figure 4.11). A total of 1,161 interactions were logged over two days during CeBIT 2006. The number of possible combinations could have also risen dramatically had semantic order been taken into consideration (i.e. 'feature' then 'object', or 'object' then 'feature'). For example, when distinguishing between

¹²DFKI: Deutsches Forschungszentrum für Künstliche Intelligenz, MTI: Mensch-Technik Interaktion

$H_F H_O$ and $H_O H_F$, the number of combinations for HH alone would equate to 312 (12 features x 13 objects x 2 semantic sequences). Although such interaction is possible in the MSA, the accuracy of events based on semantic order was not considered.

One aspect of the study affecting the recognition accuracy is the size of the grammars that were used. The speech grammars consisted of 171 words, while handwriting (as also shown in figure 4.11) consisted of a total of 47 words. Intra-gesture (which is based on the handwriting grammars) consisted of 25 individually selectable keywords/phrases (with a mean length of 9.6 characters) and 13 graphical objects that can be browsed through on the PDA's display. Extra-gesture consisted of 13 unique objects represented as camera images printed onto passive RFID tags (8cm x 5cm in size). In addition to grammar size, the accuracy of intra-gesture feature selection is also affected by speed because a user is required to select moving entries from the visual-WCIS scroll bar. For the purposes of the study, the scrolling text moved from right to left at 15 characters per second. A total of 50 characters can be displayed on a single line on the display, and these characters are refreshed five times a second (i.e. every 200ms). Confidence values for intra-gesture feature resolution were statically configured to be Cf=0.8 for N=1, Cf=0.4 for N=2, and Cf=0.2 for N=3.

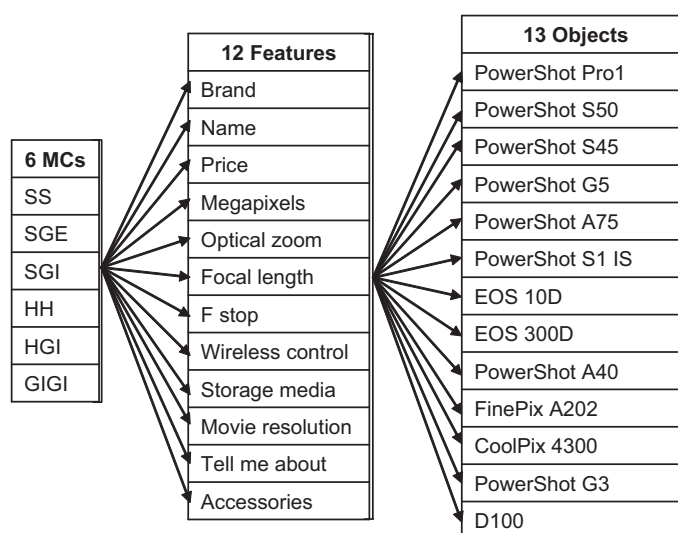


Figure 4.11: The range of interaction combinations that were used as the basis for the study: 6 Modality Combinations (MCs) x 12 Features x 13 Objects.

4.1.3.1 Modality Accuracy

The 1,161 interactions that were logged over the two days at CeBIT 2006 are categorized by semantic type (feature, object) and modality (speech, handwriting, gesture), as seen in table 4.6. This table shows the average confidence value and average accuracy rates obtained over the total number of interactions for feature and object information in general, and then for feature and object information based on each of the provided modalities. The table further shows the total number of occurrences within each category and the total number of errors that were recorded in each of the categories.

Several important aspects can be drawn from the results in the table. Perhaps most notable

	Cf_{Av}	Accuracy	Errors		Cf_{Av}	Accuracy	Errors
FTR	0.63	94.49%	64 in 1161	OBJ	0.76	93.37%	77 in 1161
FTR_S	0.43	95.14%	27 in 556	OBJ_S	0.29	98.62%	3 in 218
FTR_H	0.82	93.83%	28 in 454	OBJ_H	0.88	68.67%	73 in 233
FTR_GI	0.80	94.04%	9 in 151	OBJ_GI	0.84	99.83%	1 in 593
				OBJ_GE	1.00	100.00%	0 in 117

Table 4.6: Confidence value (Cf) and accuracy rate averages recorded during the study, where the first row in the table indicates the averages for feature (FTR) and object (OBJ) recognitions and the remaining rows show the averages for feature and object recognitions based on modality.

is that while the majority of accuracy rates for feature and object recognition are around 94% or more, the use of handwriting for providing object information resulted in a notably lower rate of accuracy (68.67%). This is thought to be due to the similarity of many of the object names like ‘PowerShot S50’ and ‘PowerShot S45’, where a single misrecognized character can easily lead to an incorrectly matched grammar entry. The table also outlines that the recorded confidence value averages for each group (Cf_{AV}) do not reflect their respective accuracy averages (Accuracy).

Figure 4.12 plots the percentage of feature and object confidence value occurrences as generated by the recognizers over the range from 0.0 to 1.0 (on the left of each graph). The figures further show the associated accuracies for each of the given confidence values, which were provided as feedback into the system by the experimenter during the course of the study (on the right of each graph). The accuracy rates for each group are summarized below:

- Speech (FTR_S=95.14% and OBJ_S=98.62% accuracy):** Almost all values with a $Cf > 0$ were correct for both feature and object selection (see figure 4.12A and 4.12B). Despite a high proportion of results being given a confidence value of $Cf=0$ (26% of occurrences for feature selection and 35% of occurrences for object selection), only a comparatively small proportion of these results were found to be incorrect (15.07% for features and 3.9% for objects). These findings imply that recognition results accompanied by a confidence value above zero are very likely to be accurate, while results accompanied by a confidence value of 0 are more likely to be correct than incorrect. Thus a Cf value of 0.0 generated by the employed PDA speech recognizer is not alone depictive of an incorrect recognition.
- Handwriting (FTR_H=93.83% and OBJ_H=68.67% accuracy):** Unlike with the modality of speech, figure 4.12C and 4.12D show that handwriting errors occur over the whole range of generated confidence values, with some higher confidence values having an unexpected lesser accuracy than the lower confidence values. The range of accuracy values is also more dispersed when compared to speech, ranging from 0% accuracy to 100% accuracy. For features, the accuracy of handwriting results that are accompanied by higher confidence values is comparable to that generated by the speech recognizer. In comparison, the error rate for object selection is very large and is dispersed over a wide range of confidence values, with the value $Cf=0.9$ returning the highest proportion of all errors occurring in the category of OBJ_H recognitions (13% from 31.3%).
- Gesture (FTR_GI=94.04%, OBJ_GI=99.83%, and OBJ_GE=100.00% accuracy):** Gesture error rates for object selection can be seen to be negligible, while gesture error rates for

feature selection averaged 5.96%. Because the confidence value for feature selection was at the time of testing fixed to the value $Cf=0.8$, only general information about the modality's accuracy can be obtained, rather than information regarding individual confidence values.

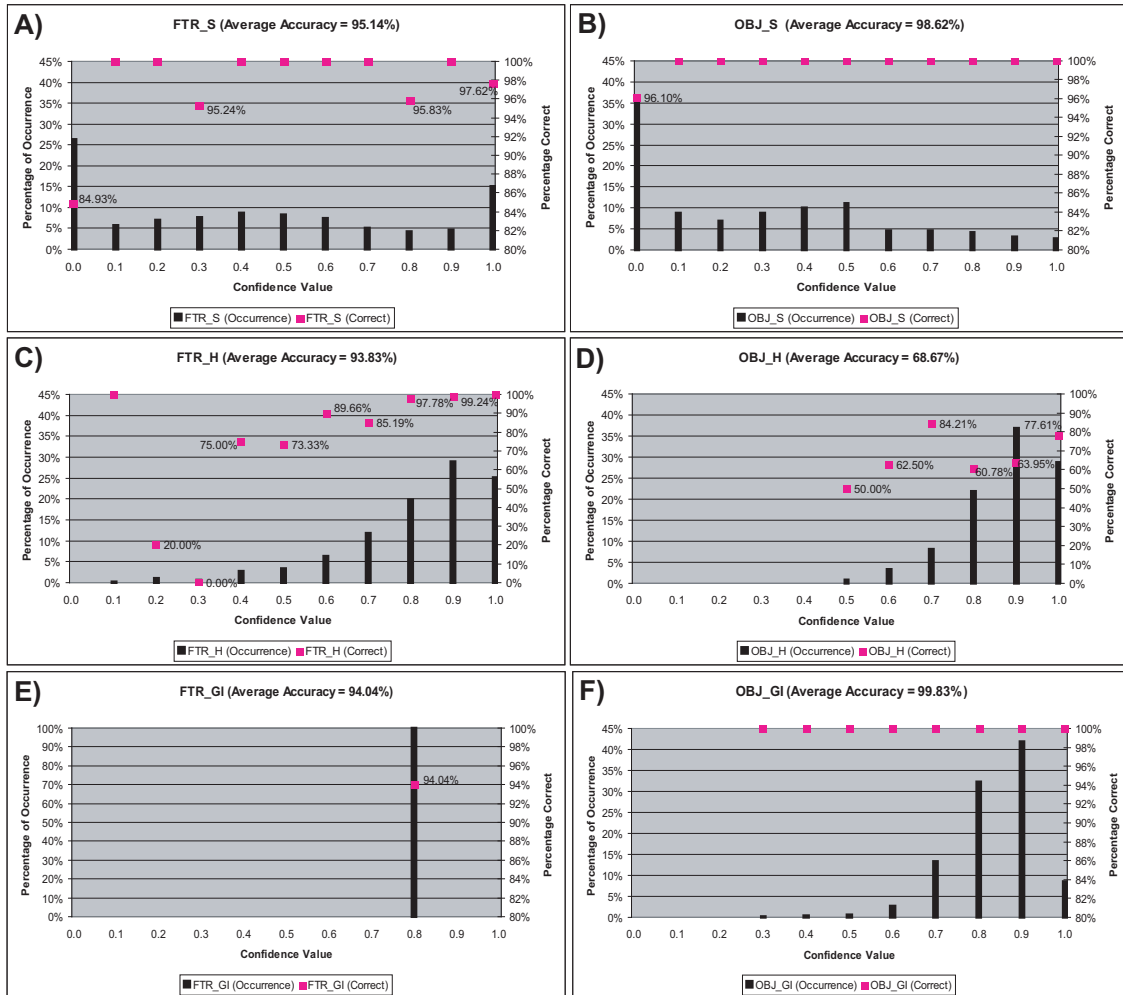


Figure 4.12: Accuracy rates for the recorded confidence values generated by the recognizers, for speech (A, B), handwriting (C, D), and gesture (E, F). The bar-columns indicate the percentage occurrence over the total number of interactions, while the line-markers indicate the percentage of correct interactions per individual confidence value.

Some general observations can be noted when comparing confidence values across different modalities. For example, it can be seen that recognizers assign confidence values differently, so that a $Cf=0.9$ for OBJ_H is much less reliable than a $Cf=0.9$ for OBJ_S. Even within the same modality, the confidence values assigned to user input can differ in terms of accuracy, e.g. FTR_H (accuracy of 93.83%) and OBJ_H (accuracy of 68.67%), and these differences are magnified further when compared over only a subset of confidence value assignments, like those between $Cf=0.8$ and 1.0 for handwriting, in which inaccuracies for object selection are high in comparison to feature selection.

To counter the differences observed here with regards to confidence scoring and accuracy, methods are required to rebalance the confidence values based on information over the complete set of available communication modes. This is the topic of section 5.3.3.

4.1.3.2 Modality Interaction Times and Scalability

Due to the large number of interactions that were recorded during the CeBIT 2006 usability study, it was possible to generate results on the average length of time it took for a user to communicate whilst using a particular modality combination. These results are outlined in table 4.7, and it can be seen that the modality combinations SGI and SS were the fastest, closely followed by HGI and SGE. The modality combinations HH and GIGI were much slower (i.e. more than twice as slow), but are the only two listed combinations that permit a user to interact privately with the system. In chapter 6, usability study results show that users are sometimes willing to trade speed for privacy, particularly in public environments. One reason why the modality combination SGI is so fast is that a user can provide speech and gesture input to the system in parallel, thus optimizing temporal aspects of the interaction.

Modality Combination	Total Time (HH:MM:SS)	Number of Interactions	Average Time (HH:MM:SS)
SGI	00:13:37	94	00:00:09
SS	00:17:01	103	00:00:10
HGI	00:17:10	97	00:00:11
SGE	00:15:36	73	00:00:13
HH	00:32:34	103	00:00:19
GIGI	00:30:35	96	00:00:19

Table 4.7: Time statistics for the interactions recorded during the usability study.

Some modalities are more scalable than others. In chapter 6 for example it is outlined that speech and handwriting scale better than intra-gesture for large feature and object databases, because it is easier to speak out the name of a physically visible product than it is to first manually find the product on a small mobile display and then point to it via intra-gesture. Such a benefit is however only applicable in instances where the interaction vocabulary is known by the user, either because it is intuitive (e.g. the products are displayed in front of them), or because the information is available to the user in an easy to understand fashion (e.g. via audio feedback, or as in the case of the MSA via the visual-WCIS scroll bar). Listed below is a brief discussion on the scalability of the MSA communication modes - speech, handwriting, and gesture - for user enquiries into object and/or feature information.

- **Speech and handwriting:** Speech allows for the selection of many features and objects all in quick succession. While handwriting is slower than speech, it also allows a user to naturally and flexibly request information on shopping products by writing on the PDA's display. Large vocabularies affect the recognition accuracy of speech and handwriting, particularly when the recognition is constrained to resource-limited mobile devices. Current state-of-the-art systems can minimize the accuracy effects of large vocabularies by creating different dialogue states in which only subsets of the entire grammar are active at a given time. This

is one of the tasks assigned to dialogue management. In contrast to speech, where a user of the MSA system provides requests by speaking out complete phrases, handwriting input requires the user to only enter keywords, thus improving on what is otherwise a very slow communication mode.

- **Gesture:** Gesture relies on a user selecting features and/or objects by either pointing to them on the display, or physically interacting with their tangible counterparts in the real-world. Because only a maximum of nine objects are displayed on the screen at any one time, and a maximum of around six features are displayed in the visual-WCIS scroll bar, recognition accuracy is high. The trade-off is that a user must first search for the relevant object in a potentially very large dataset, or they must wait for the relevant features to scroll by them on the display. As more and more objects and/or features are introduced to the system, this searching process will also increase in complexity. Similar to speech and handwriting, the effects of large datasets can however be minimized through dialogue-management procedures, in which a user might first be asked to select a particular product type like ‘digital camera’ or ‘language technology’ before the specific product vocabularies are activated (see figure 2.14). Extra-gesture suffers the problem that although it is a highly accurate modality, even for very large datasets, it requires (at least in the MSA implementation) the presence of physical real-world objects and is thus only of use if the products are within arms reach.

Section 6.2.5 extends on this outlined work on accuracy, speed, and scalability, by detailing a broad range of modality qualities that are considered important for multimodal interaction, including: comfort, enjoyment, familiarity, speed, accuracy, scale, accessibility, privacy, intuition, and complexity.

4.2 Multimodal Interaction

This section categorizes multimodal interaction in terms of its temporal and semantic synchrony, and in terms of the degree of (semantic) overlap between input originating from the recognizers. The effects of linking modalities to individual semantic constituents are discussed, as too are the wide range of modality combinations that are available in the systems described in this dissertation. Discussion is also extended to include analysis on different recognizer configurations that can be used for capturing semantically overlapped multimodal input.

4.2.1 Temporal and Semantic Synchrony of Multimodal Interaction

For the purpose of this dissertation, multimodal input is classified by the temporal synchrony and the semantic synchrony of the encompassed modalities. A similar approach is used by Caelen (1994), where four different types of interaction context are defined according to what the author classifies as the ‘use of modes’ and ‘information dependence’. The resulting multimodal integration types used in the approach are termed: exclusive, concurrent, alternate, and synergistic, but are generally distinguishable from one another through their temporal and semantic interrelations. Another not too dissimilar classification is given in (Alexandersson et al., 2004), where four basic interaction patterns are outlined - redundant, concurrent, complementary, and contradictory interaction. In this dissertation, multimodal interaction is classified by the groups: temporally non-overlapped, temporally overlapped, semantically non-overlapped, and semantically overlapped.

4.2.1.1 Temporal Synchrony of Multimodal Interaction

A prominent contribution to the temporal classification of multimodal interaction is outlined in (Oviatt, DeAngeli, & Kuhn, 1997; Xiao, Lunsford, Coulston, Wesson, & Oviatt, 2003), where input is said to occur sequentially or simultaneously with respect to time. In particular, ‘sequential interaction’ is defined to occur when multiple modalities are separated by a time lag, while ‘simultaneous interaction’ is defined to occur when multiple modalities are temporally overlapped. The definition of sequential and simultaneous interaction are used in a number of studies that identify multimodal integration patterns of different user groups including children (Xiao, Girand, & Oviatt, 2002), adults (Oviatt et al., 1997), and the elderly (Xiao et al., 2003). Sequential and simultaneous interaction as defined above, correspond 1:1 with the terms ‘temporally non-overlapped’ and ‘temporally overlapped’ interaction as used throughout this dissertation. Similar definitions regarding the temporal synchrony of input further exist within the W3C EMMA Working Draft, where the term ‘sequential’ is used to refer to multiple interpretations in which “the end-time of an interpretation precedes the start-time of its follower”, and the term ‘overlap’ is used to refer to multiple interpretations in which “the start-time of a follower interpretation precedes the end-time of its precedent”. In (Cohen, Coulston, & Krout, 2002), the definition of simultaneous interaction defined in (Oviatt et al., 1997) is extended for use in a study in which 20 minutes of video containing speech and gesture interaction from four people is analysed. The temporal relationship between simultaneous modalities is defined in this paper as being either ‘nested’ (where gesture occurs within speech), ‘contains’ (where speech occurs within gesture), or ‘staggered’ (where gesture follows speech before the speech component is ended, or vice-versa).

Temporal synchrony in this dissertation refers to the temporal relationship that exists between input that is received over multiple modalities. Modalities exhibit different characteristics, and one such characteristic is the ‘temporal duration’ of a modality, for example speech and handwriting are generally slower modalities than point-gesture. In the MSA, the time required to provide a ‘feature’ or an ‘object’ referent via gesture typically lies within the 1 to 3 second time range, while 5 seconds are required to provide a referent via speech, and around 10 seconds for handwriting. Durations of such differing time lengths have implications for systems that process multimodal input in real-time, in that differing multimodal input will take differing amounts of time to provide by a user. In the MSA, the modality fusion process is triggered as soon as a ‘feature’ has been identified (e.g. ‘price’). Depending on the modality used for the accompanying ‘object(s)’ (e.g. ‘PowerShot S45’), the system might decide to wait either 1 to 3 seconds for the object input, or up to 10 seconds for such input. Getting the temporal duration wrong will result in either a sluggish system (if too much time is provided by the system) or a miss-recognition (if the input is cut short by the system). To counter this, systems must be able to identify which modalities are currently active. Often, user tendencies are such that only a subset of the overall multimodal interaction possibilities are actually used, thus allowing a system to also better predict how long a user will take when communicating with the system. Table 4.8 outlines the temporal durations assigned to a range of modality input combinations in the MSA, which are formed by assigning a modality to the ‘feature’ and ‘object’ values in an interaction (e.g. ‘SGI’ represents feature=“speech” and object=“intra-gesture”). Note too that the temporal duration of an interaction is shorter when all of the semantic constituents are provided in a single user-turn in contrast to the individual constituents being provided in multiple user-turns. This is particularly evident for the unimodal combinations HH and SS where the time required to start and stop the modality adds to the overall duration of the user interaction. Allowing for interactions that span either single or multiple user-turns, as is

possible in the MSA, requires a flexible architecture that is able to check the status of a user's interaction to determine when it is complete.

Speed	Temporal Duration	Modality Input Combinations <Feature><Object>
Fast	Approximately 1 to 3 seconds	S _G I, G _I G _I , H _G I S _G E, G _I G _E , H _G E
Average	Approximately 5 seconds	S _S , G _I S, H _S
Slow	Approximately 10 seconds	H _H , G _I H, S _H

Table 4.8: Temporal durations for modality input combinations in the MSA, where each combination represents a modality for the feature and for the object input (e.g. <Feature modality="S"><Object modality="GI">).

The 'temporal order' in which multimodal input can occur is also important. For example, during an interaction, one modal input can be said to occur 'before', 'during', or 'after' another modal input. In the case of the MSA, multimodal input can be provided in any order, e.g. speech before gesture, gesture before speech, handwriting before gesture, and so on. Semantic constituents in the MSA (e.g. feature="price", object="PowerShot S45") can also occur in any temporal order. The temporal order of modal input and of semantic constituents is illustrated by the following example, where the semantic constituents are represented by F (Feature) and O (Object) and the modalities are represented by S (Speech) and H (Handwriting):

$F_H O_S$: <FTR_H="price"><OBJ_S="PowerShot S45">
 $O_S F_H$: <OBJ_S="PowerShot S45"><FTR_H="price">

As shown in figure 4.13, such input may or may not overlap with respect to time. Sequential multimodal input refers to input that is *temporally non-overlapped*, while simultaneous multimodal input refers to input that is *temporally overlapped*. Temporally overlapped input occurs in the MSA when, for example, gesture and/or handwriting is provided during the same time-interval that a speech utterance is provided. The flexibility exhibited by the MSA is achieved in that the individual recognizers work independently to one another and are linked only by a central blackboard upon which interaction nodes are stored and later interpreted by the modality fusion component. The importance of a flexible architecture with regard to temporal order is supported by studies on sequential and simultaneous integration, in which it is shown that different user groups are habitual. For example, with regards to map-based tasks, 66% of adults were in (Oviatt et al., 1997) shown to be sequential integrators, while 77% of children and 80% of the elderly were shown to be simultaneous integrators. Studies further show that a user's multimodal integration pattern is resistant to change, even when high error rates of 40% are applied to force a switch in integration pattern (Oviatt et al., 2003).

Temporal duration and temporal order can be calculated by analysing the start and stop times of individual modality segments that are provided to the system. Some systems like SmartKom (Wahlster, 2002b, 2003), QuickSet (Cohen et al., 1997), and MATCH (Johnston et al., 2002) also record timestamp information for individual words as they are recognized by the system. This is important in applications where the actual timestamp of words (e.g. demonstrative pronouns like 'this' and 'that', and locative adverbials like 'here' and 'there') are needed to be linked to the

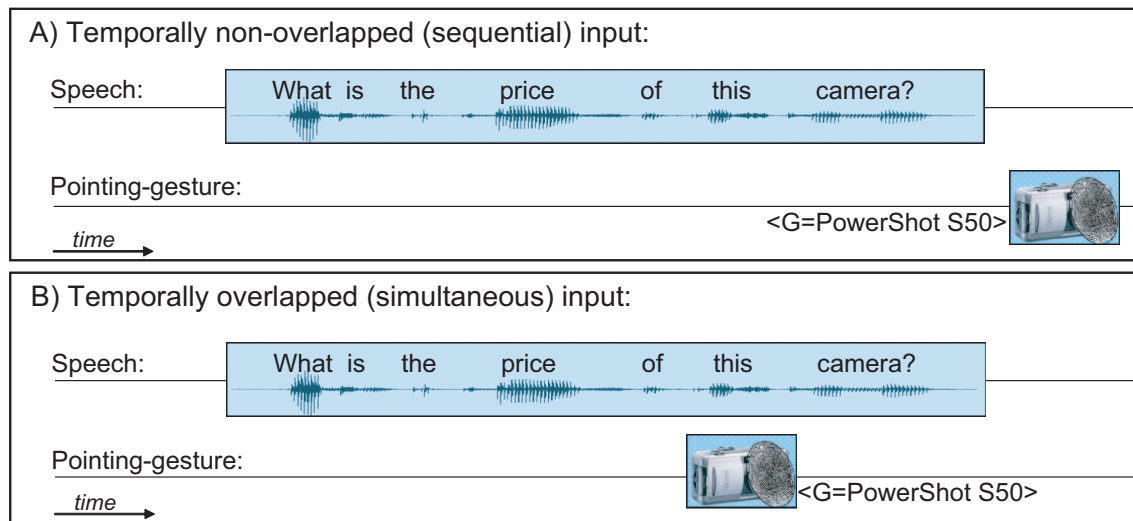


Figure 4.13: Temporal relationships of multimodal input: A) temporally non-overlapped input and B) temporally overlapped input. The communication modes are in this case speech and gesture, which are represented in the figure by the visualization of an audio signal and a fingerprint.

timestamps of their counterparts in a different modality (e.g. for error resolution). To illustrate, a user might in the scenario of the MSA provide the following spoken input: “Compare this camera to this one”, but provide only a single point-gesture rather than two. In such an instance, knowing the timestamps of the speech segments ‘this camera’ and ‘this one’ might identify for which segment a gesture reference was missing, thus simplifying the error handling of the system. Unfortunately, to the best of the author’s knowledge, no embedded speech recognizer to date has an API that provides access to word-level timestamp information. Thus, systems like the MSA have to suffice with information on the temporal order and temporal duration of input when interpreting speech that was processed locally on the mobile device.

4.2.1.2 Semantic Synchrony of Multimodal Interaction

Semantic synchrony refers to the semantic relationship that exists between multiple modalities. Studies have shown that different modalities are better suited than others for different types of semantic constituents. For example, within a map-based task in which users were asked to select real-estate, based on attributes such as location and price (Oviatt, 1996), spoken input was identified as being best suited for ‘subject’ and ‘verb’ constituents, while pen input was identified as being best suited for ‘spatial-location’ constituents. Similarly, in (Wasinger et al., 2005; Wasinger & Krüger, 2006), it was found that within a shopping task, speech is best suited for enquiring about an object’s ‘features’ (e.g. ‘price’, ‘brand’), while gesture is best suited for selecting an actual ‘object’.

Semantic constituents can occur in a particular order, which, alongside the temporal order of input as described above, can be referred to as the ‘semantic order’ of the constituents. In the MSA, semantic constituents like ‘feature’ and ‘object’ can be provided in any order, e.g. feature before object and vice-versa, and quite often relevant semantic information spans multiple user-turns, as in the case when a user refers to a referent in a previous user-turn (i.e. anaphora) or in a

future user-turn (i.e. cataphora). To cater for endophora (see section 2.3), systems like the MSA must store a list of previously and presently active referents along with potential future referents.

As shown in figure 4.14, multimodal input may also overlap with respect to the underlying semantics of the input. When the semantic constituents within a user interaction do not overlap (as in the examples above where ‘feature’ and ‘object’ information are provided rather than ‘feature’ and ‘feature’ information) the input is called *semantically non-overlapped*, while if the semantic constituents provided within a user interaction do contain overlapped (or redundant) semantic information, the input is called *semantically overlapped*. Semantically overlapped input occurs in the MSA when for example the modalities speech and gesture are both used within a single interaction to provide the same information about a referent.

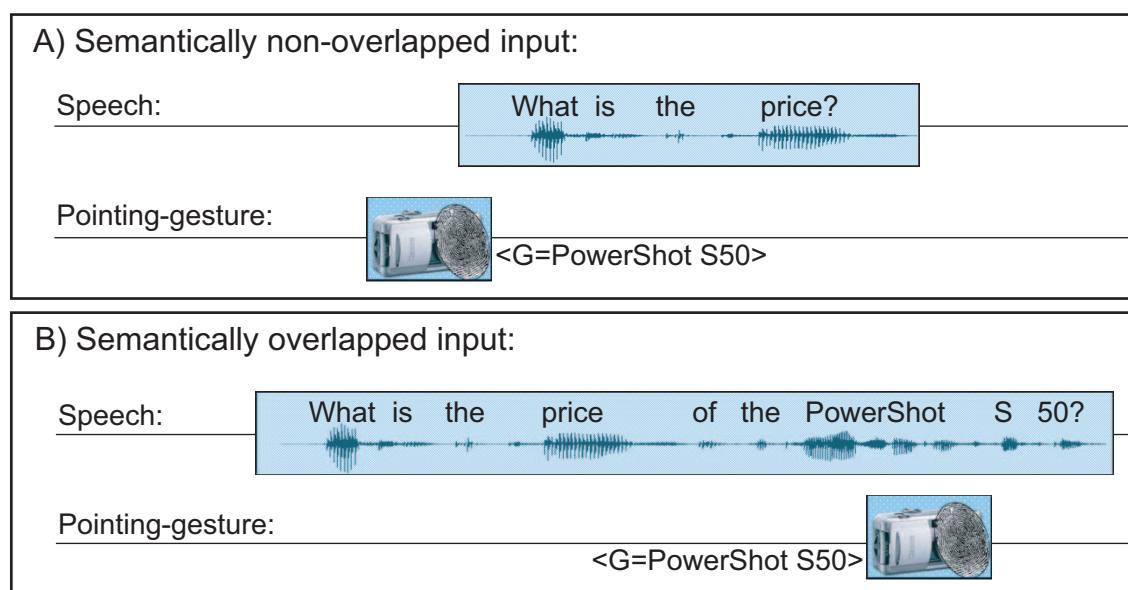


Figure 4.14: Semantic relationship of multimodal input: A) semantically non-overlapped input and B) semantically overlapped input.

Wasinger and Krüger (2004) categorize semantically overlapped input into the three groups: non-overlapped, overlapped and non-conflicting (constructive), and overlapped and conflicting (destructive). In particular, ‘non-overlapped’ input occurs when input that is provided to the system does not contain any overlapped semantic information, while ‘overlapped and non-conflicting’ input is said to occur when semantic information that has been provided multiple times does not conflict, and ‘overlapped and conflicting’ input is said to occur when semantic information has been provided multiple times and does conflict. These three categories are illustrated by the examples shown in table 4.9.

In contrast to temporal synchrony, which is a well-established field of research, semantic synchrony is still a developing field and the terms used to classify semantically overlapped multimodal input (where it is classified at all) are varied. For example, in (Oviatt & VanGent, 1996) it is described as “simultaneous redundant spoken and written input”, and in the W3C Ink Markup Language Working Draft (W3C-InkMarkup, 2004) it is termed “crossmodal redundancy”. In this dissertation ‘semantically (non)overlapped input’ is often shortened to just ‘(non)overlapped input’ to conserve space, especially in the figures illustrating 23 different modality combinations in

Semantic Input	Modality	
	Speech	Gesture
Non-overlapped	What is the price?	PowerShot S50
Overlapped and non-conflicting	What is the price of the PowerShot S50?	PowerShot S50
Overlapped and conflicting	What is the price of the PowerShot S50?	EOS 300D

Table 4.9: Three classes of semantically overlapped input (Wasinger & Krüger, 2004).

chapter 6. ‘Temporally overlapped input’ always retains its full name when used in this dissertation.

4.2.2 Semantic Overlap

When multimodal input is temporally overlapped, the modalities are said to share a common ‘time space’, while when multimodal input is semantically overlapped, the modalities are said to share a common ‘semantic space’. Just as input can be overlapped by differing temporal amounts (e.g. one input can be partially overlapped or completely overlapped by another input in time), input can also be overlapped by differing semantic amounts, which in this dissertation is called the *degree of semantic overlap*, or simply ‘partial semantic overlap’. For example, the degree of semantic overlap for a single gesture like $\langle G = \text{“the PowerShot S50”} \rangle$ is less than the same gesture accompanied by the spoken utterance ‘this camera’ or ‘the PowerShot S50’. The degree of semantic overlap is best seen by the hierarchical ontology demonstrated in figure 5.3 where a ‘PowerShot S50’ is known to be a ‘Canon’-‘digital’-‘camera’-‘shopping product’-‘object’. For multiple references to the object referent ‘PowerShot S50’, a high degree of semantic overlap is said to occur if the references are close to the actual leaf node (i.e. ‘PowerShot S50’) rather than the parent node (i.e. ‘object’). In this dissertation, a referent is classified as being either: ‘unidentifiable’, ‘type identifiable’, or ‘uniquely identifiable’. Table 4.10 illustrates these terms for the case where speech is used to provide the object referent ‘PowerShot S50’. This classification of referents applies equally well to other modalities like handwriting and gesture (when accompanied with an appropriate set of graphical objects).

Referent Classification	Example
Unidentifiable	What is the price [of this]?
Type identifiable	What is the price of this camera?
Uniquely identifiable	What is the price of the PowerShot S50?

Table 4.10: The degree of semantic overlap based on referent expressiveness.

Semantically overlapped information is often useful in reaffirming a system’s accuracy, as is the case in figure 4.14B where two modalities provide the same semantic input. It is also possible that a user provides only partial semantic information, for example, a user might say “What is the price of the camera?” while looking at a display showing a set of objects in which only one is of type ‘camera’. In this case, the minimal user input is still sufficient to resolve the referent, but requires the fusion of speech input and knowledge over which objects are currently visible on the display. Grice’s ‘Maxim of Quantity’ (Grice, 1975) supports the notion that users provide

only minimal information. The author states that people “make a contribution as informative as required, but not more so”. One work bearing some resemblance to the categorization in table 4.10 is (Gundel, Hedberg, & Zacharski, 1993; Gundel, 2003), in which a ‘Givenness Hierarchy’ containing six cognitive statuses that referents can have is outlined (i.e. in focus, familiar, uniquely identifiable, referential, and type identifiable). As described at the end of section 5.3.5, although partial semantic overlap is in principle possible in the MSA/BPN, it does not occur in the system as the user is asked to select only a single product type on the shelf during synchronization (see figure 2.14).

4.2.2.1 Linking Individual Modalities to Individual Semantic Constituents

Having now defined multimodal input with respect to time and semantics, it is useful to also incorporate discussion on its ties to the individual modalities. User interaction (even with regards to a single user-turn) rarely encompasses only a single semantic constituent, and (as described above) even a single semantic constituent can be expressed in a multitude of different ways. More commonly, user interaction is formulated by a set of grammars consisting of rules that contain multiple semantic constituents, e.g.:

```
<Rule>      = what is the <Feature> of <Product>.
<Feature>   = price | optical zoom.
<Product>   = the PowerShot S50.
```

As a result, a modality is not just used to express an entire interaction, but rather the individual semantic constituents contained within the interaction, and the same modality need not be tied to all contained semantic constituents.

The linking of individual modalities to individual semantic constituents is an area in which little direct research has been conducted. This may be due to the presumption that such an approach would result in a modality allocation that is highly specific to an individual grammar or language specification and thus not portable to other applications. The modality-free grammars created for use within the MSA (e.g. <Feature><Object>, <Feature><Object><Object>) have however been shown to be equally applicable to other scenarios including the COMPASS Smart Dining Service (Aslan et al., 2005). The lack of research by the community might also be due to the belief that the benefits applying to modality usage on the whole, are directly transferable to modality usage for individual semantic constituents. The author however believes this also to be incorrect for a variety of reasons. Usability studies on the MSA have, for example, shown that the combination of some modalities are more, or less intuitive to use. The combination of on- and off-device modalities (e.g. handwriting together with extra-gesture), and fast and slow modalities (e.g. speech together with handwriting) are two groups of combinations shown to be less intuitive (see chapter 6). Results on user preference for a range of different modality combinations also support this in that users prefer to represent ‘feature’ attributes in the MSA with the modality of speech, and they prefer to represent actual objects (which often had names that were difficult to pronounce like EOS 10D and FinePix A202) with modalities like (selection) gesture and handwriting. The results from user studies further indicate that the relationship between modality usage and the type of semantic constituents being referred to may be dependent on aspects like scenario (e.g. spatial tasks versus numeric tasks, see Oviatt and Olsen (1994)), object (e.g. socially sensitive products like clothing undergarments in contrast to less sensitive products like digital cameras), user (e.g.

individual user preferences), environment or context (e.g. background noise and public/private environments), and device (e.g. handwriting on a PDA's small display).

The W3C EMMA Working Draft specifies two terms in which the relationship between modality and underlying semantics are described (within single turns of user input): supplementary and complementary multimodality. The terms are described in the documents as being “two fundamentally different uses of multimodality”:

- **Supplementary multimodality:** ‘Supplementary multimodality’ refers to the use of modalities in which every interaction can be carried through to completion in each modality as if it was the only available modality. Such functionality enables a user to select at each time the modality that is best suited to the nature of the interaction and the user’s situation.
- **Complementary multimodality:** ‘Complementary multimodality’ refers to the use of modalities in which the interactions available to the user differ per modality. Complementary use of multimodality is said to occur if interactions in one modality are used to complement interactions in another. For applications that support complementary use of different modalities, the W3C EMMA Working Draft points out that particular care needs to be placed in identifying to a user what modality (or modalities) can be used at each particular time.

Based on this W3C classification, it can be seen that a system’s multimodal capabilities might be 1. supplementary (or unimodal) for some or all modalities, 2. complementary (or multimodal) for some or all modality combinations, or 3. supplementary and complementary for some or all encompassed modalities and modality combinations. The flexible architecture of the MSA corresponds to the third choice, in that the modalities speech, handwriting, and gesture can each be used in a supplementary (or unimodal) fashion for all product interactions in the MSA, and the modality combinations (9 in total for interactions consisting of a feature and an object referent) can be used in a complementary (or multimodal) fashion for all product interactions in the MSA.

4.2.2.2 Modality Input Combinations in the MSA

In this dissertation, the term *modality input combination* is used (in contrast to the more general term ‘modality’ or ‘modalities’) to define interaction in which individual modalities refer to individual semantic constituents. To illustrate, in the example shown earlier in figure 4.14A, the multimodal interaction $\langle G = \text{“PowerShot S50”} \rangle \langle S = \text{“What is the price?”} \rangle$ can be classified as a Speech-Gesture (SG) modality input combination, in which the ‘feature’ is provided by the modality of speech and the ‘object’ is provided by the modality of gesture. Using the term modality input combination to reiterate the W3C terms supplementary and complementary multimodality, one can see that ‘supplementary multimodality’ (4.15A) refers to the use of unimodal modality input combinations like SS, HH, and GG (for the simplified modality-free language $\langle F \rangle \langle O \rangle$). Applications like the MSA and the BPN are indeed supplementary as they allow for complete interactions to be performed in each modality supported by the system (speech, handwriting, and (intra) gesture for the MSA; speech and (intra) gesture for the BPN). In comparison, ‘complementary multimodality’ (4.15B) refers to modality input combinations that are comprised of different modalities such as SG and HG.

Applications are often designed to cater for only a limited set of modality input combinations. The MUST system for example (Almeida et al., 2002) caters for only speech-gesture (pointing) combined interaction, rather than supporting all the modality input combinations speech, gesture

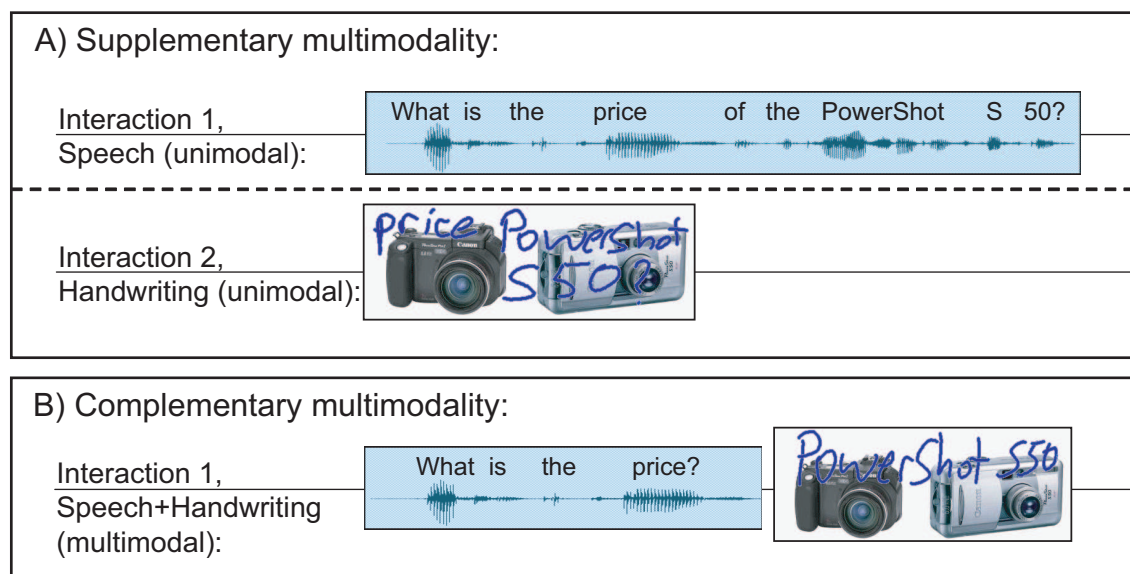


Figure 4.15: Supplementary (A) and complementary (B) modality combinations as applied to the MSA.

(pointing), and speech-gesture combined interaction. Past research on spoken dialogue systems has indicated that a common problem for unacquainted system users is knowing what they can say to the system (both in terms of how to formulate an input and in terms of knowing the extent of the application’s functionality). The author believes that a similar problem will begin to exist for multimodal applications as more and more applications are fitted with a broader range of modality input combinations and where the decision on which combinations to make available to a user is often based on presumption rather than a thorough study of the benefits and disadvantages of each individual combination in a given environment context.

In comparison to supplementary multimodality, where each interaction can be carried out in each modality, complementary multimodality (as defined in the W3C Working Draft) does not require that each interaction be carried out in each individual modality input combination, rather only a subset of all combinations. Indeed a far greater challenge (and one of the challenges of this dissertation) is to study the complete set of complementary multimodalities, i.e. the set of all encompassed modality input combinations. As an example, in the modality-free language supported by the MSA (see section 5.1.3.2), it can be seen that a typical interaction consists of a query+feature and an object (e.g. “What is the *price* of the *PowerShot S50*?”). For three modalities (speech, handwriting, and gesture) and two semantic constituents (feature and object), the number of modality input combinations total 9 (see figure 4.16 left), and extending this to an example containing three semantic constituents (e.g. “*Compare this camera to this one.*”) subsequently increases the number of modality input combinations to 27 (see figure 4.16 right). This increase in combinations does not take the degree of semantic overlap into account (e.g. ‘this camera’, ‘the PowerShot S50’), nor the combinations in which only some of the semantic constituents are overlapped (e.g. a non-overlapped feature combined with overlapped object information). In the case of the MSA, the blackboard architecture does however cater for such semantic constituent combinations without the need to define specific templates for each different modality input combination.

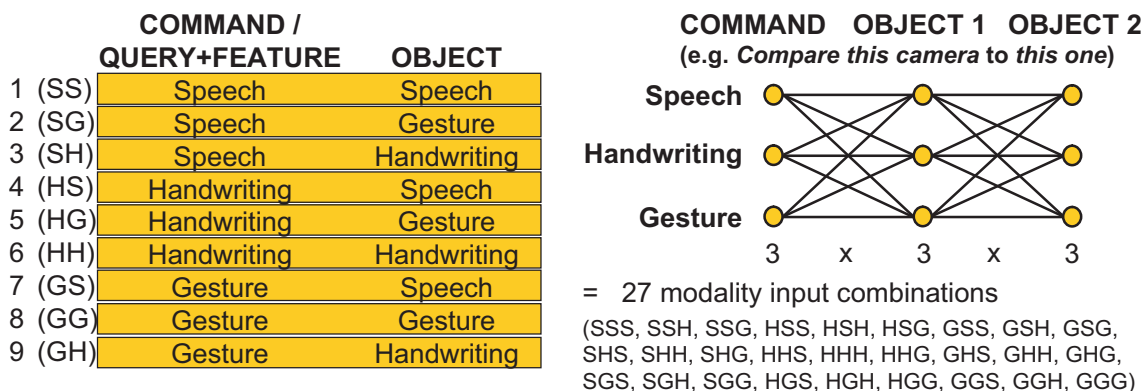


Figure 4.16: MSA modality input combinations showing modality input combinations for two semantic constituents (left) and three semantic constituents (right).

Modality Input Combinations					
	Feature	Object		Feature	Object
1	S	S	13	S,H	*
2	S	H	14	S,GI	*
3	S	GI	15	H,GI	*
4	S	GE	16	*	S,H
5	H	S	17	*	S,GI
6	H	H	18	*	S,GE
7	H	GI	19	*	H,GI
8	H	GE	20	*	H,GE
9	GI	S	21	S,GI	S,GI
10	GI	H	22	S,GI	S,GE
11	GI	GI	23	S,GI	GI,GE
12	GI	GE			

Table 4.11: The 23 modality input combinations analysed in the MSA. 1 to 12 are semantically non-overlapped, while 13 to 23 are semantically overlapped: overlapped feature (13 to 15), overlapped object (16 to 20), and overlapped feature and object (21 to 23). The ‘*’ is used as a wildcard to denote the use of any modality or modality input combination.

In the MSA, a total of 23 different modality input combinations form the basis of interaction. These include 12 semantically non-overlapped combinations and another 11 semantically overlapped combinations, as shown in table 4.11. The number of non-overlapped combinations (12, i.e. 3 communication modes for the feature x 4 modes for the object) differs to the 9 illustrated in figure 4.16 (left) because gesture interaction for an object referent (e.g. 'PowerShot S50') can be further categorized as being either intra-gesture (interaction with objects on the PDA display) or extra-gesture (interaction with tangible physical-world products). The 11 overlapped combinations were chosen such that the effect of using a wide range of overlapped modality combinations on the semantic constituents 'feature', 'object', and 'feature+object' could be analysed. These 11 overlapped combinations do not cover the complete set of overlapped combinations, but are in the author's opinion a fair representation of this set. For the overlapped 'feature+object' combinations, which are more complex and have a longer temporal duration than all of the other types of combinations that were studied, only the combinations based on speech and gesture were analysed in detail. This set of overlapped 'feature+object' combinations was chosen based on preliminary requirements analysis and literature studies that showed speech and gesture to be a viable option for overlapped combinations. The viability of the overlapped speech and gesture combinations (in contrast to other overlapped combinations) are supported by the iterative tests and summative evaluations resulting from the usability studies on user preference and modality intuition as outlined in chapter 6.

4.2.3 Multiple Recognizers and their Contribution to Semantically Overlapped Input

This section analyses the benefit of different recognizer configurations for the capturing of multimodal interaction.

4.2.3.1 The For and Against on Semantically Overlapped Input

In the previous section, it was shown that semantically overlapped input arises when for a particular interaction, a user provides the same information about a referent through the use of different modalities like speech and gesture, e.g.:

<S="What is the price of the PowerShot S50?"><G="PowerShot S50">

Such semantically overlapped interaction was further stated to contribute either in a constructive manner (i.e. overlapped and non-conflicting) or in a destructive manner (i.e. overlapped and conflicting). The main advantage of semantically overlapped input is that it increases the robustness of human-computer interaction in the form of higher multimodal recognition accuracy rates. This increase in robustness is due to the unique composition that different communication modes and their associated processing techniques have, often giving one communication mode a natural advantage over another mode. Gains in robustness due to the processing of dual-input signals is documented not only for computer interpretation of input interaction but also for human interpretation of output presentation (e.g. audio and visual) (McLeod & Summerfield, 1987, 1985; Sumbly & Pollack, 1954).

Opponents to semantically overlapped information note that "the content of multimodal input is usually not redundant between modes" (Hura & Owens, 2003) and that it is a myth that "multimodal integration involves redundancy of content between modes" (Oviatt, 1999). The main drive

behind these opponents is not so much the question on whether semantically overlapped input is useful, but rather that users do not provide such input to start with, be this due to the increased complexity and temporal duration of such interaction, user preference, or other. Oviatt (1999) in particular makes the point that the “dominant theme in a user’s natural organization of multimodal input is complementarity of content rather than redundancy”.

In the usability studies described in chapter 6, two of the 11 overlapped modality input combinations - SGI (object overlapped) and SGE (object overlapped) - were shown to be ‘not significantly non-intuitive’ to use in a real-world setting, while five of the 11 (including the above mentioned two) were rated not significantly non-intuitive in the laboratory setting. Clearly this intuition rating is quite distant from the ideal case, i.e. that the modality combinations be significantly intuitive to use, but proving that the combinations are at least not significantly non-intuitive is a solid starting point. A better result was obtained for user preference in a laboratory setting, with the above two overlapped modality input combinations being rated 5th (SGE, object overlapped) and 8th (SGI, object overlapped) best out of all 23 modality input combinations, and thus higher than even some non-overlapped modality combinations (see figure 6.6A). The results from these studies show that although users may not particularly like semantically overlapped modality combinations, the results are not sufficiently significant to discount all of the combinations based on user preference and modality intuition. It is the author’s opinion that semantically overlapped input should not be prematurely discounted as being a viable form of input, particularly because a user’s preference for a modality input combination may be affected by the context in which it is used, and most studies do not incorporate the aspect of application context when testing semantically overlapped modality input combinations.

Applications expected to benefit from semantically overlapped multimodal input combinations include for example those with a focus on error minimization. Aviation (e.g. flight controllers), health (e.g. data entry in hospitals), and the military (e.g. coordinating the location of military units) are three application contexts where error-minimization (also with respect to time) is a critical aspect. Education (e.g. tutoring and self-learning) is another area that would likely benefit from semantically overlapped input, for example students using multiple modalities like speech and handwriting to reinforce the learning of a topic (similar to writing and thinking aloud at the same time). Scenarios where the user is on-the-go and in adverse environments (e.g. containing differing types of noise relating to the individual communication modes), and even certain medial tasks like credit card entry might benefit from semantically overlapped input, where Oviatt (2000a) outlines that the benefits of mutual disambiguation are “greatest for relatively brief or impoverished monosyllabic content”, a property common to single-digit numbers found on a credit card.

Proponents for allowing users (but not forcing users) to provide semantically overlapped input, including the author, would finally argue that partial semantic overlap (see section 4.2.2) is also a form of so-called ‘redundant’ information, and that contrary to myth 6, “Multimodal integration involves redundancy of content between modes” (Oviatt, 1999), does commonly occur during multimodal interaction. Such partially overlapped input also has little additional complexity associated with it, for example the user interaction: <S=“What is the price of this camera?”><G=“PowerShot S50”>, where one modality identifies the object type while the other modality identifies the actual object and thus also the object’s type.

4.2.3.2 Passive Collection of Semantically Overlapped Input

For semantically overlapped input to be willingly provided by a user (rather than forcedly provided), the interaction must be intuitive, simple, and time efficient. One way to access such input is if it can be passively collected from the user. ‘Passive input’ is defined in (Oviatt, 2000a) as being “input that requires no explicit user command to the computer” and contrasts to ‘active input’, like speech and handwriting, where a user does intend for the interaction to be issued to the system. A prominent source of passive input can be derived from vision technologies that track and interpret gaze, head position, body location, body posture, facial expressions, and manual gestures (Tan, Shi, & Gao, 2000; Turk & Robertson, 2000; Myers et al., 2002). Another source of passive input can be derived from sensing technologies. In the project BAIR for example (Wahlster et al., 2004b), biosensors are being used to infer a user’s affective state, which is achieved by measuring biophysiological data like electrocardiogram and electrodermal activity through sensors applied to a user’s body. In the BPN, a variety of sensors like GPS, magnetic compass, and 3-axis attitude sensor arrays (pitch, roll, yaw) were used to determine a user’s location and facing direction, while in the MSA, a user’s facing direction could be determined based on infrared beacons that provide each shelf’s ID when connecting to a shelf.

One particularly good modality combination in terms of active and passive input, is that of speech and lip movement, where overlapped input can be captured not from two different communication modes (e.g. speech and handwriting as described in the previous section) but rather from a single communication mode (i.e. speech). This is achieved through the use of two compositionally different recognizers that process speech and lip movements in parallel to interpret the input signal. In (Bernstein & Benoit, 1996), a third input device for recognizing speech, called a face-glove, is also mentioned in addition to a microphone and a camera, thus permitting for the recognition of speech via the three senses hearing/audio, sight/vision, and touch. One source of complementary information achieved through the processing of speech and lip movement is in relation to the visemes and phonemes used for vowel articulation. In particular, Robert-Ribes, Schwartz, Lallouache, and Escudier (1998) state that ‘vowel rounding’ is better conveyed visually and ‘vowel height and backness’ is better conveyed auditorally. Research prototypes that combine speech and lip movement can be found in (Rubin et al., 1998) and (Stork & Hennecke, 1995), and IBM has also created a commercial product called the IBM “Audio Video Speech Recognition System”¹³, which uses a microphone, camera, and an infrared light placed in the mouth piece of a wireless headset to capture audio and to read lip movements (see figure 4.17).

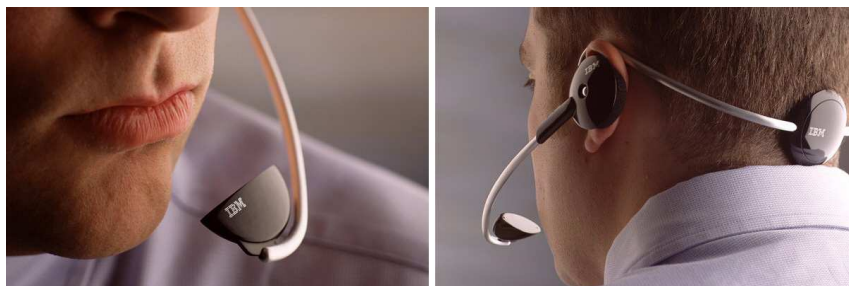


Figure 4.17: IBM audio video speech recognition system (AVSRS).

¹³IBM AVSRS, http://www.ibm.com/technology/designconsulting/port_headset.html

A similar approach is also illustrated in the SmartKom project where three independent recognizers are used for processing the single modality of speech in parallel (Wahlster, 2003), though rather than the three recognizers being focused on different senses associated with speech interpretation, like audio, visual, and touch, the recognizers focus on the interpretation of speech, based only on the user's speech signal. In particular, the speech signal is sent in parallel to one recognizer for speech-to-text purposes, to a second recognizer to identify clause boundaries based on speech prosody, and to a third recognizer to identify emotions in the user's speech.

4.2.3.3 Same-type Recognition as a Source of Semantically Overlapped Input

A further dimension to the collection of semantically overlapped input is that of same-type recognition. Extending on the SmartKom project where the single modality of speech is used for three different purposes (i.e. speech recognition, boundary prosody, and emotional prosody), systems also stand to gain from processing single modalities by multiple recognizers of the same type. 'Same-type recognizers' are defined in this dissertation to be any set of two or more recognizers that have the same function, for example two speech recognizers employed for the recognition of the same speech signal for speech-to-text purposes. Same-type recognizers need not be located on the same device, which would be near to computationally impossible for resource-limited mobile applications like the BPN and MSA. As shown in figure 4.18, same-type recognizers can be located both on the client PDA device and on a remote server, thus contributing to the notion of 'Always Best Connected' (ABC) (Wahlster, Krüger, & Baus, 2004a) in that the system is able to select the 'best' recognition result from those tendered by the individual recognizers.

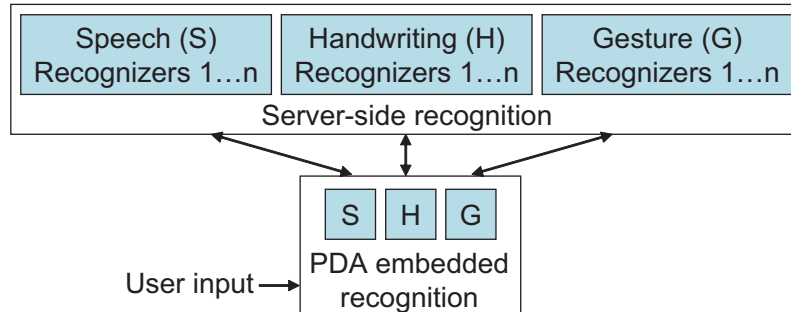


Figure 4.18: Architecture supporting same-type recognition with recognizers located both on the client PDA device and on a server.

Using the recognition of speech as an example, the accuracy of a recognition result is generally affected by aspects like the acoustic models upon which the recognizer is built (e.g. models designed for the office, car, or outdoors), as well as the size and flexibility of the defined vocabulary. Due to memory and computational restrictions, mobile devices like PDAs can at best only support a subset of the capabilities (i.e. smaller vocabularies and lower recognition accuracy rates) than would otherwise be available through the use of resource-intensive Network Speech Recognizers (NSRs) and Distributed Speech Recognizers (DSRs) (Paliwal & So, 2004). In comparison to embedded speech recognizers, NSR configurations transmit entire speech signals from a client device to a server that then performs the recognition task, while in DSR configurations, feature extraction of the speech signal is performed locally on the client device and only these features are transmitted to the server for the recognition task.

By allowing each available speech recognizer to tender results on a user's speech input, an application is more able to decide which result suits best. Limitations in network accessibility and network bandwidth utilization (e.g. bandwidth of phone networks and accessibility to WLAN and LAN networks) will at times render NSR and DSR results inadequate, particularly if the transmission to and from a server takes a lot longer than real-time human-computer interaction. At other times, the NSR and DSR results may be better suited than those from an embedded recognizer, and indeed recognizers using acoustic models that fit an environment context will also have a natural advantage over those that do not. For example, the recognizer used in the MSA and BPN uses an acoustic model designed specifically for the automotive industry and will in an office setting, therefore, be less suited to a recognizer that uses an acoustic model designed specifically for office contexts. As described in section 4.1.1.1, the MSA architecture does support multiple same-type recognizers.

This section has demonstrated that semantically overlapped input can occur in a number of configurations that are not too complex for a user to use, while at the same time contributing to the overall robustness of human-computer interaction. These configurations as summarized in table 4.12.

	User Input Modality	Recognition Device	Example, based on an MSA object referent
Active collection of overlapped input:	1. Speech 2. Gesture	1. Speech 2. Gesture	U: S=PowerShot S50 U: G=PowerShot S50
Passive collection of overlapped input:	1. Speech	1. Speech 2. Lip-movement	U: S=PowerShot S50
Multiple recognizers as a source of overlapped input:	1. Speech	1. Speech 2. Speech	U: S=PowerShot S50

Table 4.12: *Different configurations illustrating how semantically overlapped input can arise in a user-friendly manner.*

4.3 Direct and Indirect Interaction: Anthropomorphization

In this section, the concepts of direct and indirect interaction and anthropomorphization are described with relation to their implementation in the MSA. Particular focus is placed on the language grammars, the product personalities, and the state-based object models that define when objects may initiate dialogue interaction with a user.

4.3.1 The Role of Anthropomorphization in the MSA

Ubiquitous environments are becoming notably more complex and instrumented. When mobile users interact with ubiquitous environments rather than with a desktop screen, there is a need for communication with a multitude of embedded computational devices. For human-environment interaction with thousands of networked smart objects, Nijholt, Rist, and Tuinenbreijer (2004) state that “a limited animistic design metaphor” might be appropriate. ‘Animism’ is the belief that “everything is alive, everything is conscious or everything has a soul” (Wikipedia, 2006a).

The term is used to refer to the belief in which “the world is a community of living persons, only some of whom are human” and to “the culture of relating to such persons, be that human, rock, plant, animal, ancestral, or other”. Animism is one of the oldest ways of explaining how things work when people have no good functional model, and Reeves and Nass (1996) point out that people do often treat objects similar to humans. One common instantiation of animism is that of anthropomorphism.

Anthropomorphism is the tendency for people to think of inanimate objects as having human-like characteristics (Wasinger & Wahlster, 2006). Although there are various product designs that use an anthropomorphic form, like the Gaultier perfume bottles that have the shape of a female torso¹⁴ (see also figure 4.19), in this dissertation anthropomorphization refers solely to the pretended conversational abilities of the products. Since a shopper’s hands are often busy with picking up and comparing products, the most natural mode to ask for additional information about a product in many situations is through the use of speech. When a product talks and answers a shopper’s questions, the product is said to be anthropomorphized.

There is a longstanding debate among HCI researchers regarding the usability of anthropomorphic representations (Shneiderman & Maes, 1997; Don, Brennan, Laurel, & Shneiderman, 1992). It is argued that they create false user expectations, interfere with system predictability, and reduce user control in certain scenarios, and this has led to user-interface design taboos like “don’t use the first person in error messages”. People are however used to dealing with disembodied voices on the telephone, and empirical user studies conducted on the MSA have provided evidence that most shoppers have little concern about speaking with items such as digital cameras (see section 6.2.4). The world of TV commercials have also reinforced the use of anthropomorphism in a shopping context, where shoppers have for years been subjected to anthropomorphized products like ‘Mr Proper’¹⁵, a liquid cleaning product that is morphed into an animated cleaning Superman, and the animated ‘M&M’¹⁶ round and colourful chocolates, as shown in figure 4.19. Indeed an abundance of examples also exist within science fiction novels like ‘The Hitchhiker’s Guide to the Galaxy’ (Adams, 1979), where a computer called Deep Thought has human-like characteristics and is programmed to determine “the answer to the ultimate question of life, the universe, and everything”, and ‘2001: A Space Odyssey’ (Clarke & Kubrick, 1993), where the conscious computer HAL 9000 becomes paranoid and afraid for his life, causing him to eventually run rampant.

Anthropomorphized interaction can often be irritating or misleading if implemented without careful consideration, but the MSA is designed in such a way that it presents its limitations frankly. This is achieved in that the visual-WCIS mechanism guides the user in a decision-oriented dialogue and makes it clear that the system has only restricted but nonetheless very useful communication capabilities. It is contended that anthropomorphism can be a useful framework for interaction design in ubiquitous and instrumented environments if its strengths and weaknesses are understood. Ben Shneiderman, as a prominent critic of anthropomorphized user interfaces stated at a panel discussion documented by Don et al. (1992), “I call on those who believe in the anthropomorphic scenarios to build something useful and conduct usability studies and controlled experiments”. With the implementations described in this section and the usability study results described in section 6.2.4, this is exactly what has been done as part of the research defined by this dissertation.

Figure 4.20 shows an instantiation of the MSA system during usability testing at Conrad Elec-

¹⁴Jean Paul Gaultier, <http://www.gaultier.com>

¹⁵Procter and Gamble, Mr Proper, <http://www.eu.pg.com/ourbrands/mrproper.html>

¹⁶M&Ms, <http://www.mms.com>



Figure 4.19: Different commercial instantiations of anthropomorphized objects, including the Gaultier perfume (left and centre), M&M chocolates (top right), and the Mr Proper cleaning product (bottom right).

tronic in Saarbrücken, where a user (U) can be seen interacting with an object (O). In the figure, the user picks up a particular object and is then automatically spoken to by the product: i) O: “Hi there, I’m the new camera from Canon with 7 megapixels”. Following this, the user queries the object: ii) U: “What is your price?”, and the object responds with: iii) O: “My price is €599”.

During the field studies outlined in (Wasinger & Wahlster, 2006) and in section 6.2.4, subjects were asked to interact with anthropomorphized digital camera objects, and this was then followed by a written questionnaire covering not only aspects regarding interaction with anthropomorphized digital cameras, but also a range of other anthropomorphized objects based on product categories like cosmetics (soap), electronics (personal computer), and automotive (cars). One simplification that was imposed on the field study to keep the results from each subject consistent was that a single voice type was used for all digital camera products. The system is however capable of providing a limited range of different synthesized voice types to a complete set of products. As described in (Schmidt, 2005; Schmitz, Baus, & Schmidt, 2006), the MSA system also forms the foundation for a number of extensions to anthropomorphized object interaction. In particular, the authors describe their intentions to associate not just different voices but also different personalities to each object, by controlling certain speech attributes, speech behaviours, and affect. The term ‘affect’ is in this case used to describe the change of the subjective condition of a person, particularly emotions, moods, and motivations, and is important because interaction with consistent and pleasant personalities has shown to evoke sympathy in the interaction partner. The authors furthermore describe their intentions to provide the objects with the ability to generate acoustic

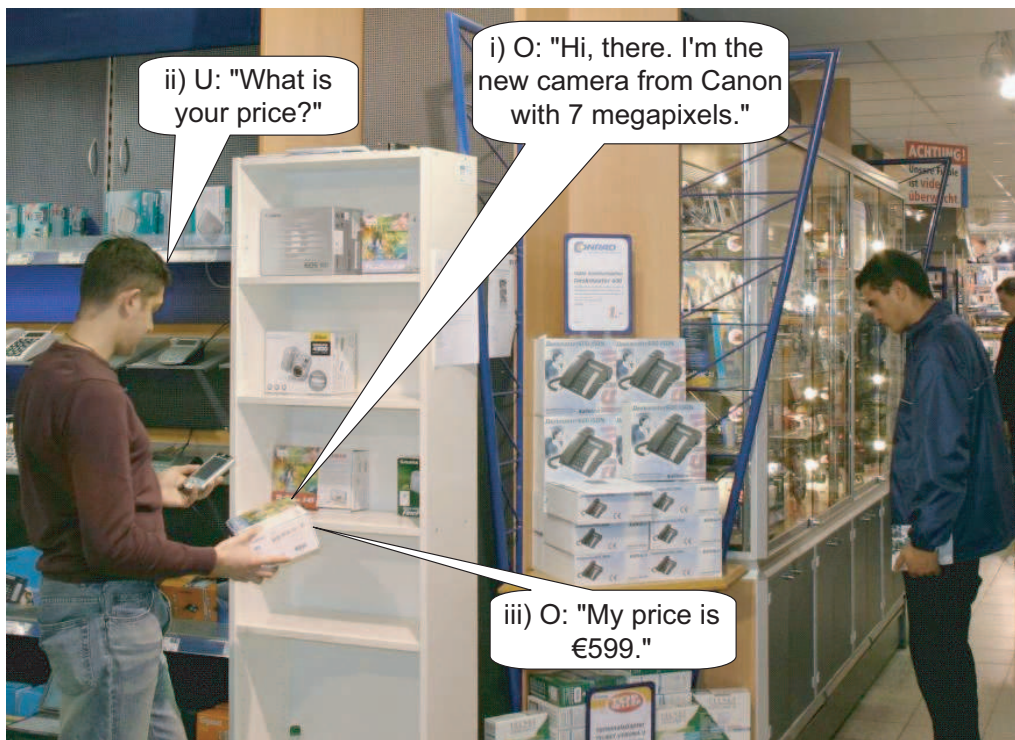


Figure 4.20: Anthropomorphized object interaction during the MSA usability study conducted at Conrad Electronic in Saarbrücken, Germany (Wasinger & Wahlster, 2006).

and haptic output through the use of Motes¹⁷ smart-dust sensors. In this fashion, a product might for example react by sounding an acoustic cue to indicate a mismatch of chosen products, like a digital camera and an incompatible camera battery.

4.3.2 Adding Human-Like Characteristics

In the MSA shopping scenario, users may choose to interact either directly or indirectly with the shopping products, and the shopping products will in return also respond accordingly. The terms *direct* and *indirect* interaction are derived from the mode of reference being made to the person segment of a dialogue. In English for example, there exist the tenses: ‘first person’ (the person speaking), ‘second person’ (the person being spoken to), and ‘third person’ (the person being spoken about). From an input perspective, direct interaction refers to the 2nd person (e.g. “What is your price?”) and indirect interaction refers to the 3rd person (e.g. “What is the price of this/that camera?”). From an output perspective, direct interaction (as used by the anthropomorphized objects) takes the 1st person (e.g. “My price is €599”) and indirect interaction takes the 3rd person (e.g. “The price of this/that camera is €599”). Direct and indirect interaction are termed multimodal ‘interaction modifiers’ because, as shown in figure 4.21, these interaction types are independent of the underlying multimodal interactions that have been discussed thus far. To demonstrate, a user may interact multimodally regardless of whether he/she chooses to interact directly or indirectly with a product, and this can be seen in the following two utterances: <S=“What is your

¹⁷CrossBow Technology, ZigBee Mote-Kit, <http://www.xbow.com/Products/productsdetails.aspx?sid=105>

price?"><G="PowerShot S50"> and <"What is the price of this camera?"><G="PowerShot S50">.

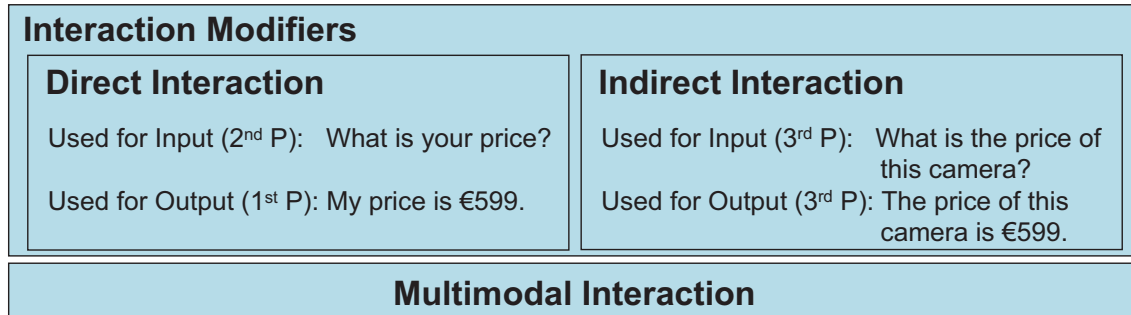


Figure 4.21: Direct and indirect interaction, shown as a modifier of multimodal interaction.

Anthropomorphization requires a slightly different set of language grammars, and the system output generated for product descriptions is also slightly different to that used during indirect interaction between a user and the MSA. Speech is the only modality in the MSA in which complete sentences can be used, and this contrasts to the communication modes handwriting and gesture where interaction is predominantly keyword-based. As shown earlier in figure 4.3, three forms of speech input are defined in the language grammars, namely ‘keyword’ (i.e. speaking only the keyword, e.g. “price”), ‘indirect’ (e.g. “What is the price of <Product>?”), and ‘direct’ (e.g. “What is your price?”). For the generation of anthropomorphized responses by the system, additional formatting information correlating with the product data is also provided, e.g. “My price is <Data>”, where <Data> corresponds to the database entry for a given product’s price.

In the default configuration, the MSA personalizes each product with one of five different formant synthesizer voice profiles (three male, two female, and all adult). These are based on parameters such as gender, head size, pitch, roughness, breathiness, speed, and volume. A limitation of this approach is that five different voices can not provide each product in a shelf - let alone an entire store - with a unique voice. By dynamically assigning voice profiles to products, the MSA would be able to at least assign unique voices to the first five products that a user interacts with, but this would limit the ability to predefine individual personalities for each object. From a commercial perspective, the most scalable alternative would be to use pre-recorded audio samples for each product. Although this requires different magnitudes of storage space, current mobile devices are expected to be able to handle such data in the configuration of the MSA system, where only products from a single shelf are synchronized at a single point in time.

4.3.3 State-based Object Model

A novel feature of the anthropomorphized objects in the MSA is their ability to initiate interaction with the user when in a particular state (see figure 4.22). These states are based on variables such as a product’s location, extra-gesture events, and an elapsed period of time. The location of a product may be either ‘in a shelf’, ‘out of a shelf’, or ‘in a shopping trolley’. Extra-gesture events that can alert an object to initiate a dialogue with the user include: ‘pickup’ and ‘putdown’ events. Thus, the physical acts of the user like ‘Pick_Up (product007, shelf02)’ and ‘Put_Down (product007, trolley01)’ are mapped onto dialogue acts like ‘Activate_Dialogue_With (product007)’ and ‘Finish_Dialogue_With (product007)’ respectively. In the given example, the putdown action reflects

a positive buying decision because the product is placed inside the trolley, but the product could just as equally have been put back on the shelf, thus reflecting a negative buying decision.

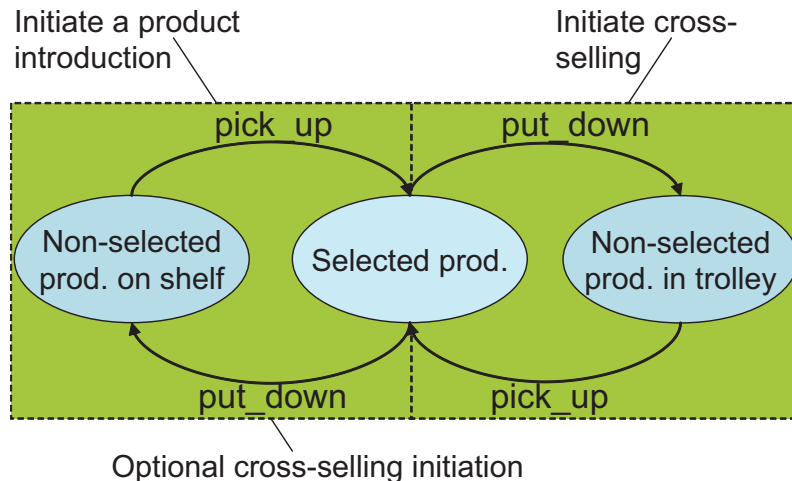


Figure 4.22: Product states in the MSA that are used for object-initiated interaction (Wasinger & Wahlster, 2006).

As described in (Wasinger & Wahlster, 2006), the MSA is actually a mixed-initiative dialogue system because both a product and a user can start a dialogue or take the initiative in a sub-dialogue. For instance, when the product is picked up - and no accompanying user query is issued - the product will introduce itself. Another system-initiated dialogue phase is that of cross-selling, which occurs when a product is placed into the shopping trolley. ‘Cross-selling’ and ‘up-selling’ are two frequently used terms used in the field of marketing to define the method in which the value of a sale can be increased by suggesting accompanying products or higher-valued products to the customer. In the MSA, such a dialogue might give advice on accessories available for the product, for example: “You may also be interested in the battery-pack NB-2LH, situated in the camera accessories shelf”. When a product is picked up from the shelf for the first time, object-initiated dialogue interaction occurs if no user interaction is observed within a five second timeframe. Silence as a powerful form of communication is well documented (Knapp, 2000), and in our case such silence forces the product to introduce itself (see (i) in figure 4.20).

In section 6.2.4, an empirical field study on user interaction with anthropomorphized objects is described. The goal of the study was to test the hypothesis that people prefer interaction with anthropomorphized objects (i.e. direct interaction) over indirect interaction. A second goal was to analyse the effects that product type (e.g. cosmetics, electronics, automotive), user-product relationship (e.g. buyer, seller), and gender (male, female) have on a person’s preference for direct interaction.

4.4 Symmetric Multimodality and Presentation Output Planning

Section 4.1 defined the communication modes used in the MSA/BPN, including speech, hand-writing, and gesture recognition. This section now looks at the flip side of these input modes. The concept of symmetric multimodality is defined, and this is followed with a description of the

output modalities used in the MSA/BPN and the degree to which such output can be configured in the system.

4.4.1 Symmetric Multimodality in the MSA/BPN

Symmetric multimodality is defined in (Wahlster, 2003) to mean that “all input modes (e.g. speech, gesture, and facial expression [in the case of the SmartKom project]) are also available for output, and vice versa.” A particular challenge for symmetric multimodal systems is stated to be that they must not only be able to understand and represent a user’s multimodal input, but also their own multimodal output. Coordination and fission of the output communication modes, also known as ‘modality fission’, provides the inverse functionality of its modality fusion counterpart since it maps a communicative intention of the system onto a coordinated multimodal presentation (Wahlster, 2002b). As outlined in chapter 3 and in particular table 3.1, most previous multimodal systems do not support symmetric multimodality since they focus either on multimodal fusion or multimodal fission. The work in this dissertation follows the principle outlined in (Wahlster, 2003): “only true multimodal dialogue systems create a natural experience for the user in the form of daily human-to-human communication, by allowing both the user and the system to combine the same spectrum of modalities”.

Symmetric multimodality in the MSA/BPN differs to the implementation observed in the SmartKom project due primarily to the different communication modes that are catered for. In contrast to the SmartKom project, where a virtual character is used to communicate to the user via speech, pointing-gesture, and facial expression, the MSA/BPN does not make it an objective to use a virtual character in its implementation of symmetric multimodality. The result of this is that whereas SmartKom models the actual physical actions via a virtual character, the MSA models just the result of the action (i.e. the resulting audio output, written text, or referent selection). The MSA does however take symmetric multimodality one step further by incorporating real-world interaction into the design, for example a user may select an object by physically picking it up, and the system may in response select an object by casting a spotlight on it. Such interaction is classified as ‘off-device interaction’ and contrasts to the more typical ‘on-device interaction’, both of which are explained in section 4.1.1.3.

The top half of figure 4.23 shows the communication modes used for providing input in the MSA/BPN: speech, handwriting, and gesture (both intra- and extra-gesture). All of these communication modes and the wide range of combinations that can be created by fusing them together are however only one side of the interaction equation. The flip side encompasses the output communication modes used by the MSA/BPN when replying to the user, and these can be seen in the bottom half of figure 4.23. Between these two sets of communication modes are the interaction manager and the presentation output planner, whose responsibility it is to coordinate the input and output modes and to carry out the process of modality fusion and fission.

Catering for interactions in the physical real-world creates a number of new challenges for symmetric multimodal systems. In (Kray, Wasinger, & Kortuem, 2004), the need to accommodate multiple collocated users is outlined to be a challenge, and in (Wasinger et al., 2003) a range of parameters categorized into different groups - environment context, user model, and device resources - are described to have an affect on presentation output planning. A summarized list of such parameters, taken from Wasinger et al. (2003), can be seen in figure 4.24 where one can see that some of the parameters remain static over time (e.g. gender) while other parameters are highly dynamic (e.g. emotions), and some parameters are easier or harder to detect than others.

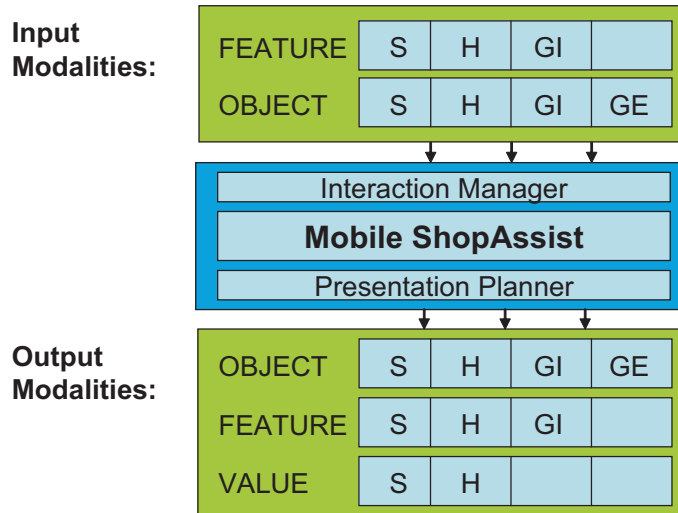


Figure 4.23: Symmetric multimodal input and output matching of the communication modes speech (S), handwriting (H), intra-gesture (GI), and extra-gesture (GE) (Wasinger & Wahlster, 2006).

To demonstrate the effects that such parameters might have on a system, a noisy or crowded environment will for example place constraints on the type of interaction that is best suited to a dialogue, and a user that is mobile will prefer the use of certain modalities over others, and device requirements will further affect which modalities should best be used.

Environment Context:	User Model:	Device Resources:
<ul style="list-style-type: none"> Noise-level Light-level Crowdedness Weather conditions (outdoors) 	<ul style="list-style-type: none"> Role of user (tourist, business person) Age (young, middle-aged, elderly) Gender (male, female) Walking speed (stationary, slow, normal, fast) Eye sight and special needs Emotions (anger, distress, happiness) Cognitive load and time pressure User interests and preferences 	<ul style="list-style-type: none"> Speaker volume Screen size and contrast CPU speed, working memory, and storage

Figure 4.24: Parameters that can affect the planning of output in mobile scenarios (Wasinger et al., 2003).

Figure 4.25 illustrates the symmetric use of modalities in the MSA/BPN. The MSA/BPN is in fact capable of presenting information both on and off the mobile device, thus being able to utilize public devices contained in the surrounding instrumented environment like public speakers and displays. Speech output for example is presented to the user via an embedded synthesizer, and/or via a remote synthesizer located on the server. ScanSoft’s RealSpeak Solo¹⁸ is used for both embedded and server-side concatenative speech synthesis, and IBM Embedded ViaVoice is used for formant synthesis. Whereas the concatenative synthesizer sounds more natural, the formant synthesizer sounds more robotic but requires much less memory (2MB per language rather

¹⁸ScanSoft RealSpeak Solo, <http://www.scansoft.com/realspeak/mobility/>




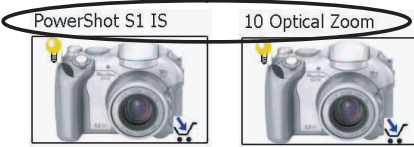



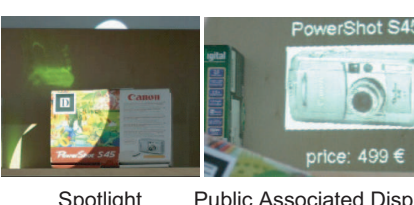
	User Input	System Output
Speech:		
Handwriting:		
Intra Gesture:		
Extra Gesture:		

Figure 4.25: The symmetric use of modalities in the MSA/BPN (Wasinger & Wahlster, 2006).

than between 7 and 15MB per language for a single voice) and provides far greater flexibility in manipulating voice characteristics like age and gender, which is important for anthropomorphization (see section 4.3). Handwriting output takes the form of system text that is displayed either on the PDA's display or in the ambient environment via PADs (Public Associated Displays, see Spassova et al. (2005)). Intra-gesture output for object selection is achieved by drawing a border around the selected object, while intra-gesture output for feature selection is achieved by highlighting the active keyword within the visual-WCIS scroll bar that scrolls across the bottom of the PDA's display. Finally, extra-gesture output is made possible through the use of a steerable projector that can place real-world products under a spotlight.

4.4.2 Presentation Output Planning in the MSA/BPN

In the MSA/BPN, the user not only has full access to the input communication modes, but also to the output communication modes that are used by the system. In particular, the user can request that different semantic information (e.g. F: Feature, O: Object, and V: Value) be presented in specific communication modes (S: Speech, H: Handwriting, G: Gesture), on, off, or both on and off the mobile device.

Elting and Michelitsch (2001) define several common tasks for a presentation planner, including content selection and organization, the selection of appropriate output modalities, and the coordinated distribution of information among the output modalities. These aspects are largely regarded as future work in the MSA/BPN, however the MSA/BPN is capable of synchronizing certain modality output combinations like speech and extra-gesture (i.e. spotlight), and is also able to

format presentation output according to its encompassed semantic element types like feature (F), object (O), and value (V). The following spoken utterance templates illustrate the different output possibilities that the system has in replying to a user query such as “How many megapixels does the PowerShot S50 have?”. In this case, the spoken system output (which may also be accompanied by information in other modalities) contains the information: <Feature=“megapixels”> (defined as part of the template), <Object=“PowerShot S50”>, and <Value=“5”>, i.e.:

F+O+V: “The <Object> has <Value> megapixels.”

O+V: “The <Object> has <Value>.”

F+O: “Megapixels of the <Object>.”

F+V: “It has <Value> megapixels.”

One aspect where the MSA/BPN excels, is with regards to the ability to configure the system’s use of output communication modes. Modifying the output settings can be done in two ways. First, as shown in figure 4.26A, the MSA/BPN contains a global modality manager in which one can select the class of modality used to present all information back to the user. Second, as shown in figure 4.26B, the user can more specifically define which modalities are used for what purpose according to a set of predefined templates. The user is additionally able to create his or her own personalized templates that can be loaded from and saved into XML format. The predefined presentation output planning templates available in the MSA/BPN are described below.

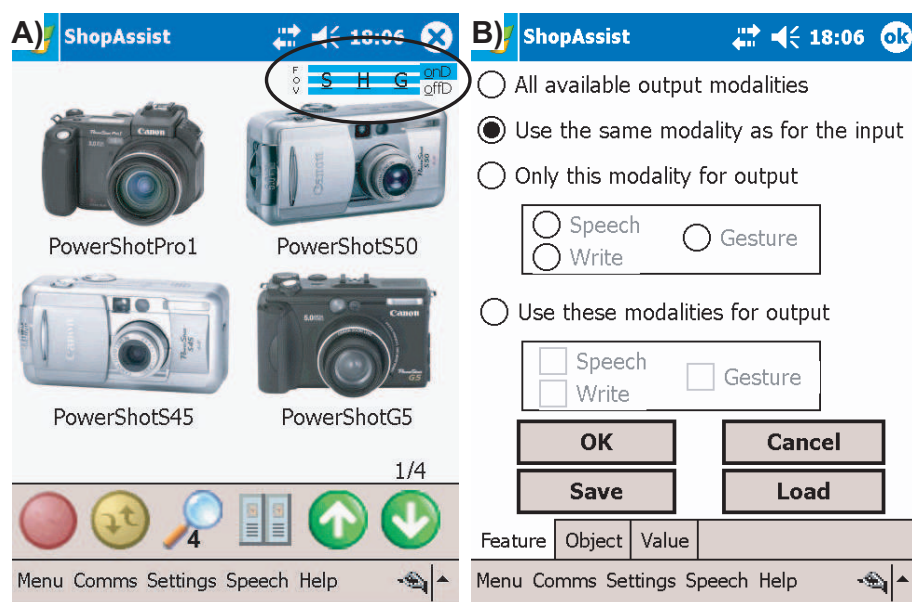


Figure 4.26: Modifying the output communication modes in the MSA/BPN via A) the global modality manager (top right) and B) the specialized user-settings interface.

- **All available output modalities:** Information is presented to the user in all available communication modes (speech, handwriting, and gesture). The user must still decide whether the information should be presented on-device, off-device, or both on- and off-device, for example depending on whether privacy is required or whether the user is shopping alone or as part of a group.

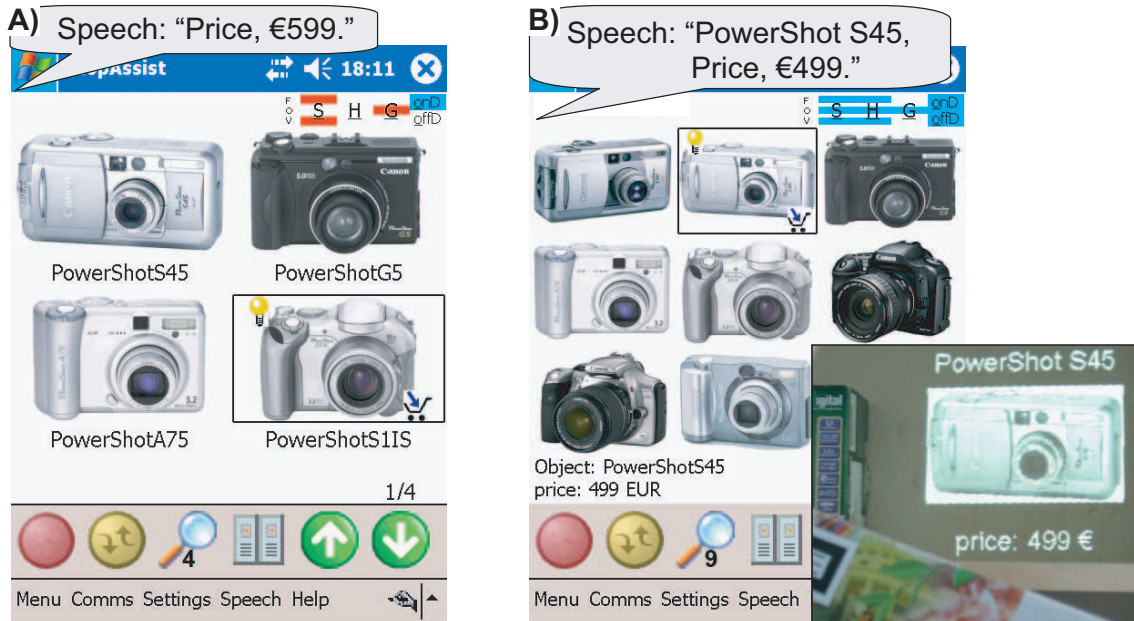


Figure 4.27: Two examples demonstrating the configurability of the MSA/BPN's presentation output planner, showing A) the result of the system mimicking the user's input modalities, and B) a user-defined configuration consisting of on- and off-device output for the modalities of speech, handwriting and gesture.

- Mimic:** In this mode, the system replicates the semantic-modality categories used by the user. Figure 4.27A shows an example in which the user has provided the following interaction: $\langle S = \text{"What is the price?"} \rangle \langle GI = \text{"PowerShot S11S"} \rangle$, and the system can be seen to reply by highlighting the 'PowerShot S11S' object on the display while simultaneously issuing feature and value information via speech: "Price, €599". A simpler derivative of this mode would be to mimic only the communication modes that were used (rather than individual semantic-modality categories) by providing all information in all the provided modalities.
- Unimodal:** In this mode, just speech or just handwriting might be selected depending for example on whether the mobile device is situated in a backpack (in which case speech would be best suited as the output means), or if the environment is particularly loud and noisy (in which case handwriting might be better suited).
- User-defined:** This last mode allows the user to specifically define which semantic types (feature, object, and value) should be communicated in which particular communication modes (speech, handwriting, and gesture). Figure 4.27B demonstrates one such configuration where speech and handwriting are used to provide feature, object, and value information, and gesture is used to provide only object information, i.e. $F_{SH}O_{SH}G_{V_{SH}}$. In this example, the information is presented both on- and off-device, such that speech output is additionally sent to a set of public speakers while handwriting and gesture information is presented via the PAD, as seen in the insert in the figure.

4.5 Multiple Users and Multiple Devices in the MSA Instrumented Environment

In earlier sections of this chapter, much was discussed with regards to modal and multimodal interaction. The goal of this section is to discuss the issues that arise when such interaction occurs in a scenario allowing multiple users to interact with multiple devices, at the same time, and with a common set of applications, as also outlined in (Kray et al., 2004; Wasinger, Kray, & Endres, 2003). Such a scenario is expected to be the norm for instrumented environment contexts of the future including shopping and sightseeing. Many of the issues that are described in this section are based on the experiences gained from designing the MSA/BPN system and can be used as guidelines for designing future systems. It is not the goal of this section to provide answers to all of the issues that are raised, rather only to indicate their significance for mobile users in instrumented environments. Topics that are discussed include that of situated interfaces, the relationship between multiple users and multiple devices, and the aspect of device control and sharing. The motivation for this section originates from the initial scenario outlined in section 2.4.2; however, rather than users being equipped and interacting with only their own personal mobile devices, they may now also be interacting with other users and with a range of devices provided as part of the public infrastructure. Shops, museums, airports, and even living rooms of the (not so distant) future are application areas where discussion on multiple users and multiple devices in instrumented environments are expected to become important.

4.5.1 Situated Interfaces

Prior to analysing the challenges to multi-user multi-device interfaces, it is important to define what exactly constitutes an instrumented environment interface and how this differs from traditional interfaces. An ‘interface’ in this context comprises all means employed by one or more users to access functionality provided by a computer system. Interfaces are embedded in a physical space known as an ‘environment’, within which ‘interactions’ representing the actions through which users communicate their goals and intentions to the system take place and the physical entities used to interact with the application, called the ‘devices’, may be found (see figure 4.28, left).

One property that sets multi-user multi-device interfaces apart from other types of interfaces is the relationship that they have with the environment. Unlike the traditional setup of a single user interacting with a personal computer, interactions involving multiple users and devices are inherently very closely linked to the state and affordances of the surrounding environment. Figure 4.28 illustrates this link via a schematic overview of the corresponding relationships (left) and shows that multiple users interact with a user interface that is comprised of several devices, in order to access functionality in the form of services or applications. In contrast to traditional graphical user interfaces, intelligent user interfaces may be largely transparent to the user, for example when a user interacts through the use of a microphone or by picking up an instrumented object.

In (Kray et al., 2004), four different types of interface are described, based on the number of people and the number of devices involved. As shown in figure 4.28 (right), the first quadrant identifies ‘single-user single-device’ interfaces, which are commonly used to interface personal devices like a walkman. ‘Multi-user single-device’ interfaces on the other hand cater for a larger audience, for example a group of people watching a pantomime on TV or listening to music on

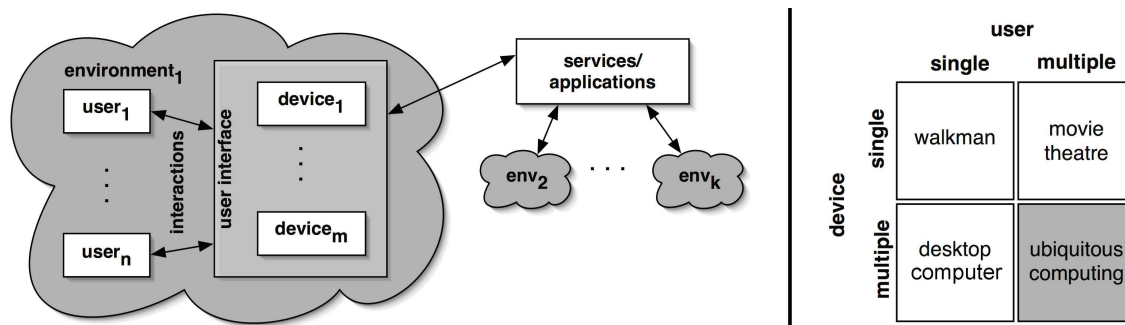


Figure 4.28: Situated interfaces (left) and categorization of different interface types (right) (Kray et al., 2004).

a radio. An example of a ‘multi-device single-user’ interface is that of the traditional desktop computer - a single user interacting with a keyboard, display, and mouse. Finally, a ‘multi-user multi-device’ interface corresponds to several people using multiple devices, as is the case when users are interacting with the same set of products in a shop. Multi-user multi-device interfaces are challenging because each transition from a less complex type of interface to a more complex one introduces new issues that need to be addressed. For example, moving from a single-user single-device interface to a multi-user single-device interface entails questions such as who controls the device and how the device can best be shared. Similarly, moving from a single-user single-device interface to a single-user multi-device interface will introduce the need of fusing multiple inputs together.

4.5.2 Multiple Users

Multiple users create a range of interesting challenges for systems operating in instrumented environments, particularly with respect to how they interact. Two forms of interaction are that of collaborative and independent interaction. ‘Collaborative’ interaction occurs when multiple users are using the same devices to achieve a common goal, e.g. two people going shopping together. ‘Independent’ interaction in comparison occurs when multiple users use a set of devices (perhaps even the same set of devices) to achieve their own separate goals, e.g. two people using a public display, one to shop for a digital camera while the other to shop for a mobile phone. Combining both collaborative and independent users results in a third type of interaction where some people collaborate while others interact independently with the system, as would commonly occur inside a store full of customers. Another distinction in this context is that users of a system may be collocated, distributed (located at different sites), or again a combination of both. Figure 4.29 summarizes the characteristics of users in a multi-dimensional graph that spans the design space.

The MSA system supports multiple users each accompanied by their own mobile PDA device. This is a particularly effective solution because, as this dissertation demonstrates, the PDA is capable of providing a rich platform for many different interaction types, and with the exception of extra-gesture recognition and the storage of the store’s product database, all processing is conducted directly on the mobile device. Multiple MSA users can each access the product database at the same time to download shelf contents, and interactions conducted through the mobile device (e.g. speech and handwriting) can easily be assigned to belong to a particular user. Assigning products that are physically picked-up and putdown on the shelves would also be possible if ad-

ditional IDs were assigned to users, perhaps based on RFID-technology in the form of a tag/ring that one wears around a finger, to link each extra-gesture event with a tangible interaction in the physical real-world.

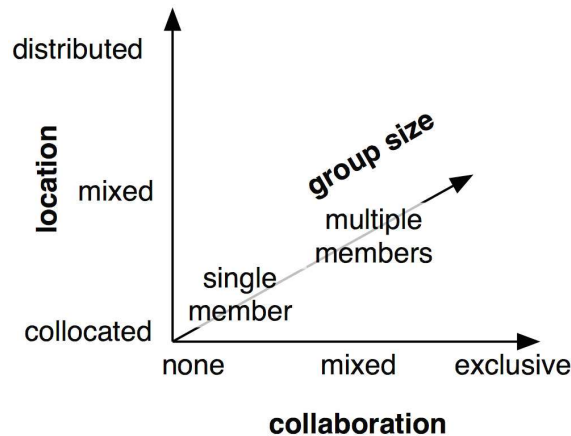


Figure 4.29: Characteristics of users (Kray et al., 2004).

4.5.3 Device Taxonomy

In order to access the functionality of an application, a user (or group of users) utilizes various devices such as a microphone and a touch screen display. While one can distinguish between the use of a single device and the use of multiple devices, the use of multiple devices (e.g. mouse and keyboard) is far more common. However, it should also be considered that multiple devices are harder to coordinate and the use of a single device may well still be necessary, for example when a large number of people are all competing for a small number of devices that must ultimately be shared.

When the MSA is used in an instrumented environment, a shopper is, in addition to the PDA, able to interact with instrumented devices like shopping trolleys and shelves, a variety of different products located on the shelves or within a trolley, public speakers, and public displays. The shopper is also able to call upon specialized services of the instrumented environment such as virtual characters that can introduce the shopper to the store or to a particular product and a public spotlight that can be used by individuals to quickly locate products in the store or on a shelf. The public displays that are available to shoppers in the MSA scenario include two different types; large plasma displays (see figure 4.30H) and Public Associated Displays (PADs) (Spassova et al., 2005). Microphone arrays connected directly to the instrumented environment are a further possibility that would provide shoppers without access to a PDA with a highly flexible form of interaction. The list of different public devices available for user interaction in the MSA scenario is shown in table 4.13, and the actual devices found in the MSA instrumented environment can be seen in figure 4.30.

Devices typically allow for 'input', 'output', or both 'input+output', and for 'private' or 'public' use. For example, microphones only support input, while speakers only support output, and touch screens support both input and output. Headphones privately transmit their output to a single user, while a public loudspeaker does not. One can furthermore distinguish between devices

Public Devices	Reference to Figure 4.30
Projector, PAD text	A
Projector, PAD images	B
Projector, virtual characters	C
Projector, spotlight	D
Shelves	E
Physical product objects	F
Shopping trolley and display	G
Plasma wall-display	H
Speakers	I

Table 4.13: Public devices in the MSA instrumented environment.

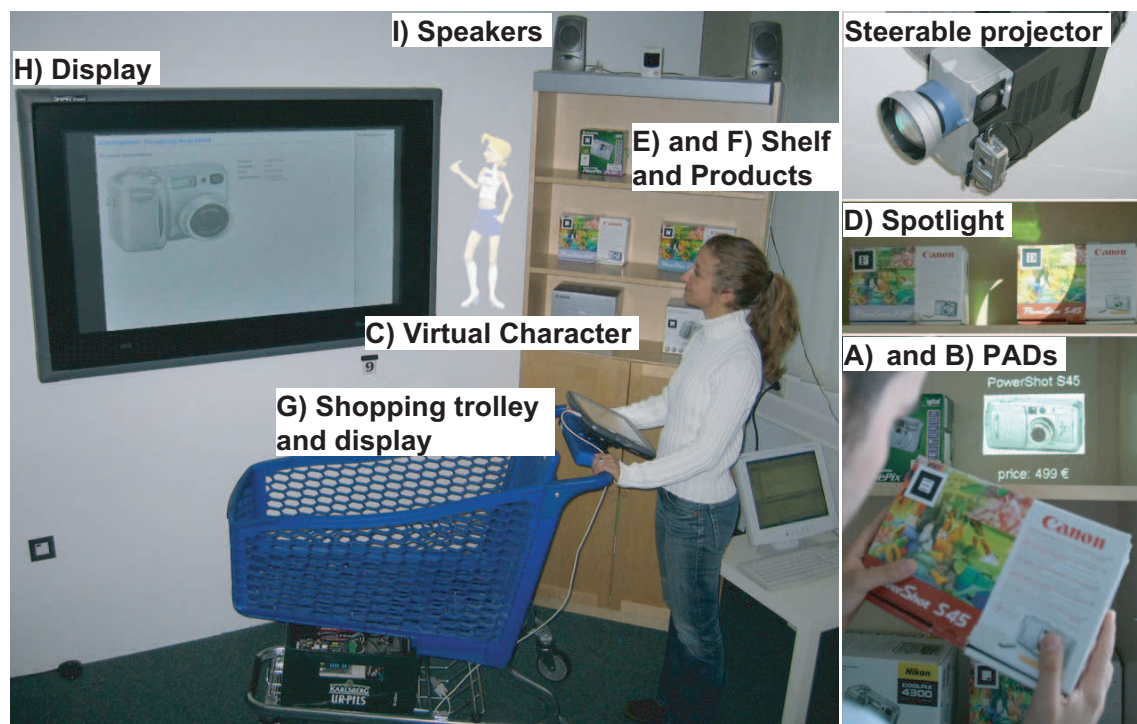


Figure 4.30: Public devices located in the instrumented environment created under the projects REAL/READY and FLUIDUM, showing the PAD text and images (A and B), a virtual character (C), the spotlight (D), the shelf and physical products (E and F), a shopping trolley (G), plasma wall-display (H), and public speakers (I).

that afford ‘shared use’ and those that only afford ‘non-shared use’. A large public display is an example of a device offering shared use, in contrast to the display on a PDA.

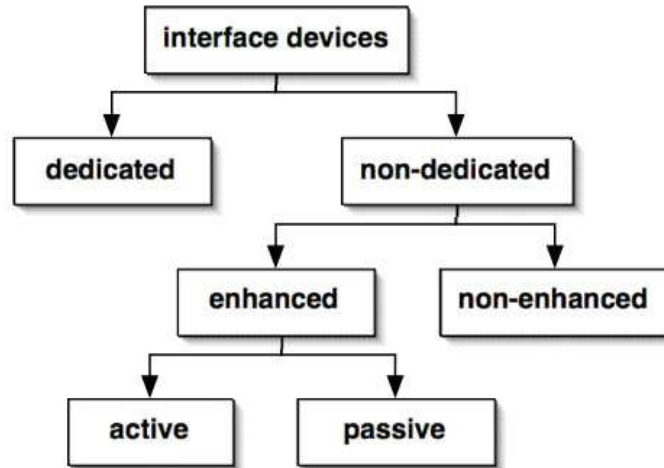


Figure 4.31: Device taxonomy in the MSA instrumented environment (Wasinger et al., 2003).

In an instrumented or ubiquitous environment, one can distinguish between several classes of interface device depending on their function and capability. On the one hand, there is the group of devices that are primarily dedicated to the handling of input and output such as displays, microphones and cameras. On the other hand, there are devices that fulfil other functions in everyday life such as shelves and products. This latter group of non-dedicated devices can be further partitioned, based on whether or not they have been augmented or enhanced. For example, one can attach a sensor, such as an RFID tag, to an object like a digital camera, to enable a ubiquitous environment to better perceive it and to facilitate its identification. Enhanced devices may be passive in that they require the environment to detect their presence, or they can alternatively be active in that they pursue interaction with their environment such as a weight-sensitive table. Figure 4.31 depicts device types as they are categorized for the MSA instrumented environment.

4.5.4 Device Control

The control of multiple devices by users and applications in an instrumented environment can be categorized into the groups: device allocation, device sharing, and device release. These forms of control are however influenced by a multitude of factors characteristic of a dynamically changing environment. There is, for example, a need for constant re-evaluation and re-adaptation of the allocation of resources due to fluctuations in users and devices as they move in and out of instrumented environments. In this section, the terms user-, system-, and mixed-initiative device control are outlined, where ‘control’ refers to the allocation, sharing, and release of devices as shown in figure 4.32. This is followed with a guideline on some implications for device control as required by tangible user interfaces, including social issues and spatial/temporal constraints relating to multiple users and multiple devices.

In an instrumented environment, both users and the system may request control of a device and in different ways. Taken from the more common dialogue-strategy categorization (Cohen, Giangola, & Balogh, 2004), device control can be classified as user-, system-, or mixed-initiative.

During ‘user-initiated’ control, a user might directly specify which device(s) should be used. There are however several ways to specify a device, ranging from spoken commands (“Show me the PowerShot S50 specifications on the plasma wall-display.”) to multimodal references (“Show me the PowerShot S50 specifications on that [pointing gesture] device.”). The user might also request device control on a more abstract level, for example by asking the system to communicate in the visual but not auditive communication mode (see section 4.4.2). Another form of request is ‘system-initiated’, in which the system automatically allocates a set of devices to a given user. The resulting assignment may however displease the user, even if multiple situational factors are taken into account, and the user may also feel controlled by the system. In this case, a ‘mixed-initiative’ approach might be preferred, in which the user directly specifies some devices, while the system automatically selects others. While this may combine the problems inherent to both approaches it may also remedy some. For example if users can specify at least some devices, they may feel less likely to have lost control.

Figure 4.32 shows that the control of a device may be either exclusive or shared. During ‘exclusive’ control a single person uses the device, while during ‘shared’ control several users may access the device either cooperatively or in parallel. In principle, the methods for device allocation also apply to device sharing, with the exception that not all devices are shareable (e.g. a headphone). The considerations presented for device allocation also apply to releasing the control of a device, i.e. device release, in that either the user or the system may explicitly or implicitly release control of a device. In addition, there may be a strong spatial/temporal component in the process, such as when a user simply walks away from a set of devices or does not use a device for a longer period of time, in which case the control of the device should also be implicitly released.

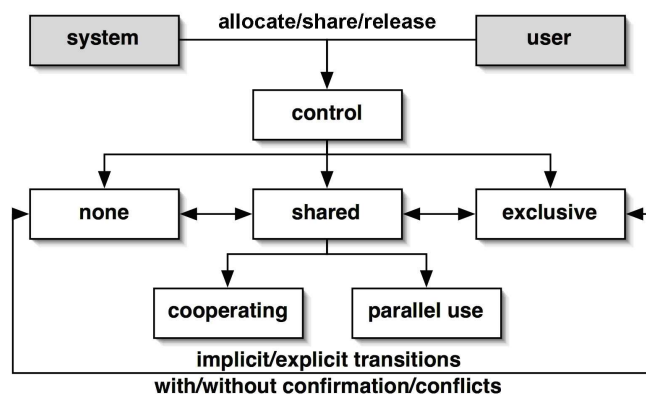


Figure 4.32: Assignment of device control (Wasinger et al., 2003).

The control of multiple devices by a system and its users in an instrumented environment presents many implications (Wasinger et al., 2003). Both a system and the users may for example have preferences for different types of device. One way to accommodate for this is through device modelling, by listing the properties of each device like shareable/non-shareable and public/private, and making the model accessible to both the system and the user. An added level of complexity arises when a device is only partly shareable. For example, touch screen displays are shareable in terms of output presentation but many are currently still difficult to share in parallel in terms

of input (the DiamondTouch¹⁹ table is one such exception). Another factor is that of resource limitation. If the devices that a system requires are no longer available, the system will have to either consider using different devices, redistributing the already allocated devices, or informing the user of an expected waiting time. Such a redistribution of devices may be classified as resource adaptation.

Similar to devices, users must also be modelled if the system is to best understand their needs, and this information must be merged with any prerequisites that the user may currently have. An important issue is that users need to be provided with system resources in a fair manner and must also ‘feel’ that this is the case, especially in times of device conflict. The system must be able to make distinctions between the desired needs of a user (soft prerequisites) and the required needs of a user (hard prerequisites). For example, a distinction may be made between a user who desires a large screen to view a single camera compared to a user who needs the large screen to compare many cameras. Distinctions may even be required to classify the value of a user for example a repeat-customer compared to a first-time customer.

In contrast to single-user scenarios, multiple users also require certain social aspects to be considered when allocating the control of devices such as privacy, contribution of background noise, and urgency. Social implications can affect either the users themselves (e.g. users desiring privacy), cooperating users (e.g. does one input device such as a microphone dominate over another input device such as a touch screen), or bystanders (e.g. the use of public audio).

Spatial influences can also have a large affect on allocating device control to multiple users. While a system must try and distribute users to areas that best support the requested functionality, the system must also consider any desires of the user and try not to force the user to move too far away from his or her current position. Spatial concerns become more complex when devices are already in use by other users, as the system must then try and predict the optimal allocation of resources not just for the present time but also for the future. Temporal influences include for example the urgency with which a user requires a specific functionality or set of devices. Temporal conflicts may arise when there are too few devices and may require decisions to be made by the system as to how long a user must wait before either an alternative user is disrupted or until users are relocated.

With regards to multiple users and multiple devices, the topics discussed in this section, including situated interfaces, multiple users, device taxonomy, and device control, are included to highlight several aspects that are expected to be relevant to mobile applications of the future.

¹⁹MERL DiamondTouch, <http://www.merl.com/projects/DiamondTouch/>

Chapter 4 discussed a fundamental requirement for multimodal systems, i.e. the modes of interaction. In this chapter it is discussed how multimodal input provided by a user through different modes of interaction can be effectively modelled and interpreted by a system.

5.1 Multimodal Input Modelling and Knowledge Representation

Section 5.1.1 describes several basic concepts like syntax, semantics, and world knowledge. Following this is a discussion on the role of knowledge sources and the modelling of user input in the MSA/BPN. In section 5.1.2, a variety of multimodal representation languages are summarized, and the contribution that ontologies make to such systems is also outlined. Finally, in section 5.1.3, multimodal input and knowledge representation are discussed with respect to the data and method attributes and the communication acts defined in the MSA/BPN.

5.1.1 Background to Input Modelling and Knowledge Representation

An important aspect for many multimodal systems is the representation of knowledge sources, like ontologies, and the modelling of user input during recognition, semantic interpretation, and unification. Whereas knowledge sources are used to define the domain of context for a system and its users, the modelling of user input is required as a communications interface to system modules and occurs at different processing stages such as when input is written to the multimodal blackboard, as it is interpreted and unified to form an unambiguous and modality-free result, and during transit between different modules. This section starts with a discussion on several limitations that are common to stand-alone mobile systems when modelling input. A few basic concepts such as syntax, semantics, and world knowledge are then defined, and following this, the role of knowledge sources and the modelling of user input at different processing stages in the MSA/BPN are discussed.

5.1.1.1 Mobile Device Limitations to Modelling Input

Mobile systems are in many respects more limited than their stationary desktop counterparts, and this is accentuated when embedded software is to run stand-alone on the mobile device rather than in a distributed or remote fashion. While applications that work in a distributed or remote fashion are capable of outsourcing processing power, rich knowledge bases, and indeed core functionality

such as language processing, these applications are inherently reliant on a connection to external networks and/or the Internet. Such a connection is generally dependent on the surrounding infrastructure and is based on technologies like WLAN, Bluetooth, infrared, and mobile phone networks. Fast and cheap Internet coverage for mobile users situated outdoors and in shopping centres is still a developing market and fast WLAN coverage is, even in larger cities, well below 100%. Without a suitable connection, distributed and remote applications can be considered useless.

The MSA and the BPN were designed to work offline as embedded applications. This removes the dependence on connectivity but also limits the functionality of the application to the capabilities embedded on the mobile device. The primary mobile device limitations with respect to knowledge representation include the speed and size of the internal memory, the processing power, and the available software. Software, and in particular SDKs and APIs, is often operating system dependent, meaning that available desktop software is commonly incompatible with that of mobile devices. The mobile applications described under this dissertation are, as a result, only capable of modelling closed domains that are restricted in size and expressiveness by the predefined language models.

5.1.1.2 Syntax, Semantics, and World Knowledge

Multimodal knowledge representation, and indeed the underlying meaning of user requests, are based on formal languages that define language syntax, semantics, and world knowledge. From a linguistic point of view, *syntax* refers to how words can be put together to form correct sentences and what structural role each word and each phrase plays in the parent sentence. Allen (1995) states that most syntactic representations of language are based on the notion of context-free grammars, which represent sentence structure in terms of what phrases are subparts of other phrases.

While syntax is used for defining the structure that input can take, it does not reflect the input's meaning, which is referred to as *semantics*. The benefit of modelling the semantics is that the information is then understandable by both computers and humans alike. The *logical form* encodes possible word senses and identifies the semantic relationships between the words and the phrases. These relationships are often captured by feature structures or frame-based structures that contain semantic slots to be populated with values taken from a user's input. In the case of the MSA, frame-based structures are used to define the different communication acts that the system is capable of interpreting, including wh-queries, yn-queries, and commands (see section 5.1.3.2).

World knowledge allows a system to reason further about its application domain and is used to map the meaning of words to values that are semantically correct for a given application context. For example, a user issuing the utterance: "Tell me about the Cool Pix", might, if buying a camera, be interested in more information about Nikon's CoolPix 4300 digital camera. Alternatively, the user, if looking at a set of fancy photos on a display, might be asking for more information regarding how the photos were taken. Such world knowledge can be used to restrain semantic interpretations of a user's input.

5.1.1.3 Knowledge Sources and their Use in the MSA/BPN

Merriam-Webster (1998) defines 'knowledge' as "the fact or condition of knowing something with familiarity gained through experience or association", and Ackoff (1989) defines knowledge as a

scale of understanding, in which data, information, knowledge, and wisdom respectively signify increased levels of understanding (see figure 5.1). According to Ackoff, ‘data’ represents symbols but does not itself have meaning. ‘Information’ on the other hand represents data that has been processed. It may or may not be useful and exhibits meaning by way of relational connections. Extending this, ‘knowledge’ represents the application of data and information and has the intent of being useful. ‘Understanding’ is defined to be the process by which knowledge can be taken and synthesized to form new knowledge. It may build upon currently held data, information, knowledge, and even understanding itself. AI systems can be seen to possess understanding in the sense that they are able to synthesize new knowledge from previously stored information and knowledge. Finally, ‘wisdom’ is an extrapolative, non-deterministic, and non-probabilistic process that is used to reason about questions to which there are no easily achievable answers or perhaps no answer at all, including questions on morals, ethics, and judgements between right and wrong and good and bad.

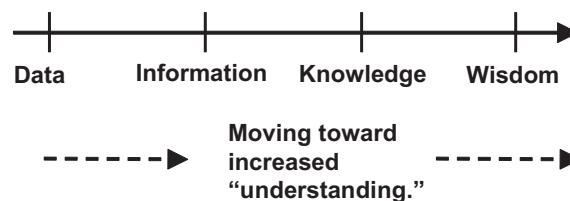


Figure 5.1: Understanding: Data, information, knowledge, and wisdom (Ackoff, 1989).

Applications like the MSA/BPN can be seen to model data, information, and knowledge. Such applications can also be seen to use simple reasoning techniques (i.e. understanding) during the interpretation of user input, for example when parsing anaphora, in which past utterance information is used to resolve vagueness in a current utterance. The MSA/BPN gains its knowledge from the following sources:

- The SQL database of objects including shopping products like digital cameras.
- The finite-state grammars that are used for input recognition and are derived from the database of objects.
- The ontology that defines product types and their relationships (e.g. a ‘PowerShot S50’ is a ‘camera’ is a ‘shopping product’ is an ‘object’) and a thesaurus for describing synonyms of object attributes (e.g. ‘price’ is a ‘cost’ is a ‘worth’).

The SQL database of objects in the MSA application is a good example for showing the modelling of data, information, and knowledge. In particular, the database entries represent the data, while the database itself, containing relationships between the data segments, represents information. Finally, the use of this information in the system to interpret user input and to then provide relevant output represents knowledge. For example, the system knows that when a user asks for the price of a product, it should retrieve information in the database regarding the product’s price and then present this information back to the user.

Two of the above mentioned knowledge sources, the SQL database of objects and the ontology, are handcrafted in the MSA/BPN, while the language models defined by the finite-state grammars are dynamically generated based on the other knowledge sources. This allows for the flexible

addition and removal of product objects within the database without the need to recompile any source code. As shown in figure 5.2, the MSA grammars are created based on the product types contained within each particular shelf, meaning that interaction with different shelves generates different grammars.

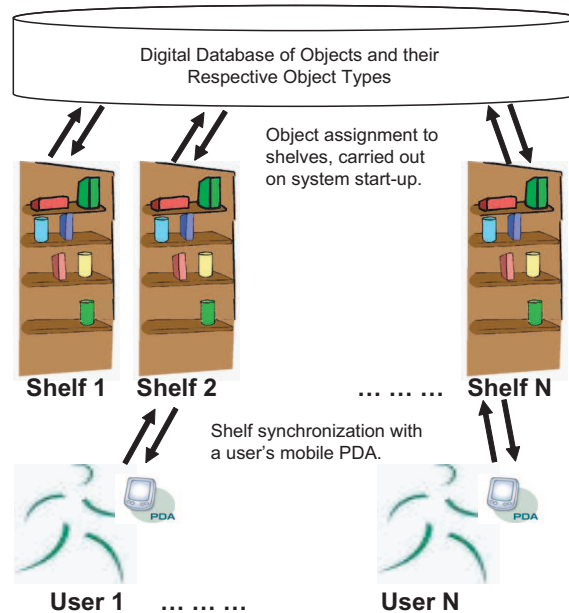


Figure 5.2: Shelf synchronization in the MSA.

5.1.1.4 Modelling of User Input at different Processing Stages

User input may be represented by a system in different forms depending on the stage of processing for which it is required. As an example, during communication with the server in the MSA application (e.g. when extra-gesture events are recognized), unicode text is represented in XML. This data is then read by an XML parser, located on the mobile device, and stored as an object tree that is based on the Document Object Model (DOM), which provides a convenient interface for loading, accessing, and manipulating XML documents. To better suit modality fusion, user input in the MSA is then stored as nodes on the modality fusion blackboard, where each node represents a ‘semantic constituent’ in the user’s input (e.g. command, feature, object) and the associated ‘modality type’ (e.g. speech, handwriting, gesture).

Communication acts are used in the MSA/BPN to define the different tasks that a user can perform with the system, like queries (e.g. “What is the price of the PowerShot S50?”), commands (e.g. “Compare the PowerShot S11S to the PowerShot S50”), and assertions (e.g. “That is incorrect”, in response to an incorrect system output). These communication acts can be likened to the concept of frames. A ‘frame’ is a data structure commonly used for knowledge representation in AI systems (Minsky, 1975). Each frame (also more commonly referred to as a ‘class’ when referring to object-oriented programming) contains properties called slots. The ‘slots’ describe a frame’s ‘attribute-value pairs’, where a value may be a primitive such as a string, an integer, or a reference to another frame. The communication acts used in the MSA/BPN represent an elementary version of the frame structures described above. This is because frames permit for concepts

like inheritance and procedural attachment, which are not needed to model the small number of communication acts available in the MSA/BPN (see section 5.1.3.2). It can however be noted that object-oriented programming techniques, including the use of inheritance and procedural attachment, are used in the MSA/BPN for other purposes, for example when modelling landmark and shopping product instantiations.

Described below are several areas where input representation is required in multimodal systems like the MSA/BPN.

- **User input recognition and input annotation:** User input that is to be accepted by a system must first be represented in a language model. Language models are modelled differently by different recognizers and recognizer types (e.g. speech and handwriting). Two of the most common methods for modelling spoken language are that of formal language theory and stochastic (or n-gram) language modelling (Gorrell, 2004). The MSA/BPN uses the formal language theory approach for modelling spoken user input. In particular, finite-state regular grammars are used. These are classified by the Chomsky Hierarchy as a type 3 formalism (Chomsky, 1956; Wikipedia, 2006b). The Backus-Naur Form (BNF), created by John Backus and Peter Naur as part of the ALGOL 60 specification in the mid 1950's (Naur, 1963), is the standard used for representing formal language. In the MSA/BPN, a deviation of BNF called the Speech Recognition Command Language (SRCL) is employed. This is a file format developed by IBM for defining grammars that are compatible with their commercially available embedded ViaVoice speech engine.

Within the MSA/BPN, recognized input is stored on a blackboard that is essentially a linked-list of nodes. Each of these linked-list nodes contain the raw and parsed data that is recorded for an interaction segment, together with the type of information that it represents (e.g. 'object' or 'feature' information). In addition, each node contains information on the method in which the input was provided including the modality type, the recorded confidence value, and the start and finish timestamps for the user interaction.

- **Semantic interpretation and unification of user input:** After a multimodal input interaction has been recognized and written to the blackboard by the appropriate recognizers, it must be interpreted to retrieve the expected meaning of the input. To do this, the individual information segments that are generated by the interaction are sorted by relevancy, matched to a relevant communication act, and then unified to obtain the resulting modality-free result. The type of data structures that are used by multimodal systems for modelling the semantic interpretation of user input generally rely on frame-based representations (Minsky, 1975) or a descendant of this called feature structures (Carpenter, 1992). These representations are able to represent possibly partial data in a record-like fashion that encompasses attribute and value pairs. This is in turn ideally suited for use in unification-based formalisms. In the MSA/BPN, communication acts like $\langle Q_{wh-yn} \rangle \langle F \rangle \langle O \rangle$ are composed of individual frames (e.g. query, feature, and object frames) and each frame's corresponding attribute-value pairs (e.g. SemanticConstituent, ModalityType, and SourceType attributes and their values).
- **Storage and communication between modules:** User interaction may also require re-representation when it is passed from one module to another. This occurs in the MSA/BPN when partial input segments are recognized by a remotely located recognizer and then sent

to the mobile device for interpretation and unification. Such representation also occurs after a multimodal user interaction has been unified by the modality fusion component, at which point the interpreted interaction is written to a locally stored file on the mobile device and also made available to registered third party services such as SPECTER (Schneider et al., 2006; Kröner et al., 2006), which can be used to keep a personal journal of a user's actions and affective states. The presentation planner used in the MSA/BPN has access to the internal data structures used by the interaction manager, and thus a re-representation of data for use by the presentation planner is not necessary. The representation used for the communication of input interaction segments is W3C's XML standard (W3C-XML, 2006).

5.1.2 Multimodal Input and Knowledge Representation Standards

In the previous section, a range of standardized representations were described, all of which have become mainstream building blocks for dialogue systems, including BNF and SRCL for modelling formal grammars, linked-list data structures that can be used to store recognized input, frame-based and feature structures for use during semantic interpretation and unification, and XML for the storage and communication of information between different program modules. While these standards have remained mostly unchanged over years past, more specific formalisms regarding the representation of multimodal interaction (both for the method and for the data) have been a topic of significant discussion. The incorporation of ontology models has been another area of recent development, focusing on providing a broader level of closed-domain knowledge and now also extending towards the Semantic Web and open domains. This section starts with a discussion on state-of-the-art ontology modelling representations and then describes representations designed for the communication of multimodal input like W3C's EMMA (W3C-EMMA, 2005), which is currently being standardized, and SmartKom's M3L (Herzog et al., 2004) and MIAMM's MMIL (Kumar & Romary, 2003), which have been specifically created as part of larger multimodal systems.

5.1.2.1 Ontology Modelling Protocols

An ontology is a formal explicit description of concepts and their interrelationships in a domain of discourse. Ontologies are usually organized in taxonomies and typically contain modelling primitives such as classes, relations, functions, axioms, and instances (Gruber, 1993). McGuinness (2002) describes how ontologies have emerged from academic obscurity to mainstream business and practice, and she also states that enormous gains exist in representing knowledge/ontologies in a format that is not just human-readable, but also computer-understandable. A range of different ontology markup languages for encoding meaning currently exist. These are generally based on XML and as described below include RDFS, DAML+OIL, and OWL (see Heckmann (2005) for more information).

- **XML:** The eXtensible Markup Language (W3C-XML, 2006) is designed to serve weakly structured data as an interchange format, and it provides rules and syntax for structured documents but imposes no semantic constraints on the meaning of such documents.
- **RDF:** The Resource Description Framework (W3C-RDF, 2004) was developed by the W3C as part of its Semantic Web effort and became a W3C Recommendation in 1999. Whereas XML concerns itself with syntax, RDF deals with semantics by providing a clear set of rules

for simple descriptive information. RDF is based on the notion of ‘resources’, which are described by statements in the form of 3-tuple sets of resource attributes: subject (resource), predicate (property), and object (value).

- **RDFS:** RDF Schema, also known as the RDF vocabulary description language (W3C-RDFS, 2004) is a semantic extension of RDF providing mechanisms for describing groups of related resources and the relationships between these resources. Whereas RDF is a data model used to express semantics, RDFS is a schema that can be used to constrain and describe data based on an RDF data model.

Several ontology languages build on RDFS. Two of these include DAML+OIL and OWL:

- **DAML+OIL:** DAML+OIL is a semantic markup language that attempts to fuse the goals of two separate markup languages OIL (Ontology Inference Layer) (Fensel et al., 2000) and DAML (DARPA Agent Markup Language) (Hendler & McGuinness, 2000). It provides representation and inference support and combines the widely used modelling primitives from fame-based languages with the formal semantics and reasoning services provided by description logics.
- **OWL:** The Web Ontology Language (OWL) (W3C-OWL, 2004) is a second such semantic markup language that was developed as a vocabulary extension to RDF. It is a W3C Recommendation and was developed by the Web Ontology Working Group, incorporating lessons learnt from the design and application of DAML+OIL.

Although ontology modelling is not a primary focus of this dissertation, it is recognized that the use of ontologies within multimodal systems is important for allowing a system to conveniently infer and deduce information regarding things like object type and relation. SmartKom (Wahlster et al., 2001) for example uses an OIL ontology for modelling domain knowledge and even incorporates the ontology into the specification of its own markup language M3L. Mobile applications like the MSA/BPN can also profit from ontological information, and this is shown in figure 5.3 where the given ontology can be seen to cover object types (and object relationships) ranging from digital cameras to language technology products and buildings. The figure also shows a hypothetical example of reference resolution in which an ontology is used to resolve the word ‘camera’ in the user utterance “What is the price of the camera?”, by mapping the reference to the camera instance ‘PowerShot S50’. This is called crossmodal reference resolution because the object type is recognized through the mode of speech, while the actual object is identified through graphical image analysis. In the MSA/BPN, ontologies are used to define different alternatives that a user can say when referring to objects, as illustrated by the following nested representation of the object ‘PowerShot S50’: [PowerShot S50 [Canon [Digital [Camera [Shopping Product [Object]]]]]]. The words defined in this representation also need to be defined in the language models of the individual recognizers (i.e. the finite-state grammars) for them to be usable. Similarly, the synonyms making up the thesaurus in the MSA need to be defined as part of the grammar files, as shown below for the word ‘price’, which exhibits the same semantics as the words ‘cost’ and ‘worth’.

```
<ATTR_ID = "price", handwriting_value="price", speech_value=...>
<ATTR_ID = "price", handwriting_value="cost", speech_value=...>
<ATTR_ID = "price", handwriting_value="worth", speech_value=...>
```

In the MSA/BPN, communication acts are defined within the application's source code. The ontologies created for the MSA/BPN applications are limited in complexity and would require modification if the applications were to be developed for a commercial market, particularly because domains like shopping and navigation contain thousands of product types and landmark types. The MSA/BPN ontology is represented using XML and Data Type Definitions (DTDs) rather than RDFS, DAML+OIL, or OWL. The MSA/BPN is also restricted in that DTDs (Document Type Definitions (W3C-XMLSchema, 2004)) are not readable by the majority of PDA XML parsers, meaning that XML documents although being checked to see if they are well-formed (i.e. complying with the basic syntax and structural rules of the XML specification) are not checked for validity (i.e. complying to the rules defined in a DTD). This restriction is expected to be resolved in future generations of the Microsoft Windows Mobile Platform.

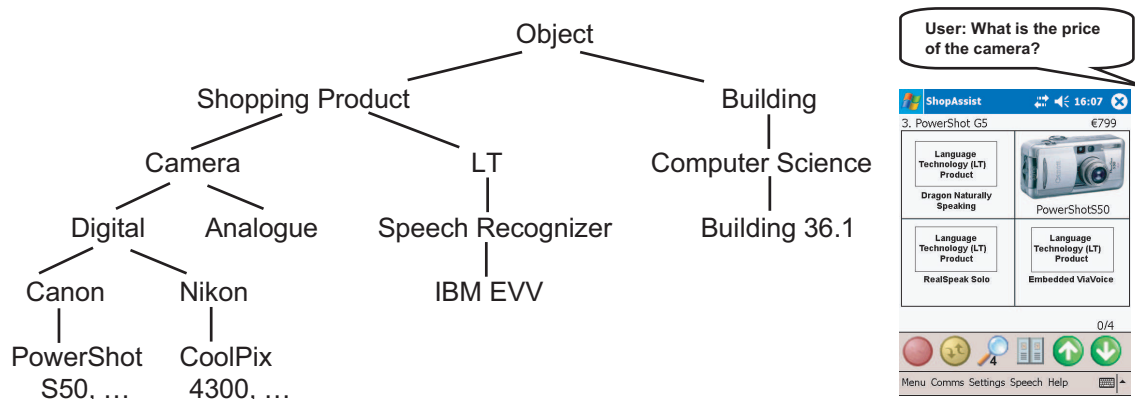


Figure 5.3: Partial ontology of object types and an example use of the ontology during reference resolution. Note the word ‘camera’ is mapped to the product ‘PowerShot S50’ because this is the only visible object of type ‘camera’.

5.1.2.2 Multimodal Communication Protocols

The specification of language representation protocols for communication between system modules has become increasingly important as multimodal systems expand in size and complexity. SmartKom (Wahlster et al., 2001) for example may be classified into over 30 different high-level modules ranging from face interpretation, to media fusion, presentation planning, and action planning (see figure 3.4 in chapter 3). SmartKom’s architecture is also complex in that the system is capable of running in a distributed fashion in which individual modules may be located on physically different computers. QuickSet (Cohen et al., 1997) is a second system designed to run in a distributed environment. Systems such as these, in which individual modules communicate over networks and potentially from different code bases (e.g. Java, C, Prolog), require a standardized communications protocol and a common language specification. Representations such as EMMA, NLSML, M3L, and MMIL (see below), all set about achieving this goal by providing a task-independent communications protocol that defines how input may be annotated with information such as timestamp and confidence, while at the same time allowing for the incorporation of task-specific templates or communication acts based on domain knowledge in an ontology. While EMMA and NLSML are W3C Working Drafts designed to be generic and thus usable by a

wide range of different applications, M3L and MMIL are representations designed (aside to being reusable) for use in specific applications, i.e. SmartKom and MIAMM respectively. VoiceXML (Voice Extensible Markup Language) (W3C-VoiceXML, 2004) is another specification, and although it was designed for creating audio dialogues, efforts are now being made to use this specification for multimodal applications, by merging it with graphical componentry to provide for speech and GUI interaction (Peters, 2006).

Some of the objectives of the MSA/BPN differ to those of QuickSet and SmartKom in that the MSA/BPN was designed to work as an embedded application that does not need to be connected to an external network. As a result, the individual modules reside locally on the mobile device rather than residing on powerful stationary computers distributed throughout the environment. This in turn removes much of the need for generating output represented in a standardized communications format because the modules all have access to the data structures stored in the working memory of the device. The third party representation languages described in this section are nonetheless very interesting in that they provide insight into the method attributes (e.g. modality type, confidence value, and start and finish timestamps) and the data attributes that they model.

EMMA: Extensible MultiModal Annotation markup language EMMA (W3C-EMMA, 2005) is a multimodal markup language that is currently being specified by the W3C Multimodal Interaction Working Group with the aim of enabling access to the Web using multimodal interaction. It is a Working Draft and is as such still under review by the W3C¹.

EMMA is used to represent the interpretations of a user's input (e.g. speech, keystrokes, pen input) together with various input method annotations (e.g. confidence scores, timestamps, input medium). The W3C expects the markup to be used as a standard data interchange format between components of a multimodal system, in which EMMA markup is generated automatically by the components rather than being directly authored by developers. Components that generate EMMA markup are expected to include recognizers (e.g. speech, handwriting), natural language understanding engines, and multimodal interaction components. Components that use EMMA are expected to include interaction managers and multimodal integration components. SmartWeb (see section 3.1.7.2) is one system that already makes use of the EMMA Working Draft specifications, and the use of the Working Draft in this system has resulted in an extension to EMMA called SWEMMA (Reithinger et al., 2005).

The EMMA language concerns itself with the interpretation of single inputs (e.g. contained in an individual natural language utterance) rather than input collected over the course of a dialogue. An EMMA document contains three different types of information; instance data that is task-dependent and contains the actual semantics of a user's interaction, optional data models that constrain the structure and content of an instance, and metadata, such as confidence and timestamp information, that annotates the data contained in an instance. The Working Draft is intentionally vague regarding the representation of the instance data and the data models, providing only recommendations that the instance data be specified in XML and the data models be based on a standard such as XML Schema (W3C-XMLSchema, 2004). The main focus of EMMA is thus on the metadata used to annotate user interpretations, rather than on the modelling of the actual interpretations.

¹W3C endorsement of EMMA will take place only after the specification has proceeded through the review stages of Working Draft, Candidate Recommendation, Proposed Recommendation, and Recommendation

The main metadata annotations defined in EMMA include: the medium, mode, and function of an input modality, as well as confidence value and timestamp information. A ‘medium’ is defined as the closed set of values defining the communication channel: acoustic, tactile, or visual. The ‘mode’ is defined as the specific mode of communication used on the channel (as seen from a user’s perspective), and includes voice (acoustic), touch (tactile), and visual appearance/motion (visual). A tactile medium is defined to include most hands-on input device types such as pen, mouse, keyboard, and touch screen, and an example mode within the tactile medium might include pointing and clicking on a graphical user interface. The ‘function’ is defined to be orthogonal to the mode, in that it defines the communicative function, for example speech can be used for recording (e.g. voicemail), transcription (e.g. dictation), dialogue (e.g. interactive spoken dialogue systems), and verification (e.g. identifying the user through their voiceprint). Confidence values may be represented in EMMA at different levels of processing (e.g. first by a speech recognizer and then by a natural language understanding component). Various different representations of timestamps are listed including start and end times, duration times, and time offsets based on references to prior interpretation timestamps. EMMA metadata also allows for successive processing of interpretations as new information becomes available and also allows for reference to prior interpretations during processing. Other data sources that may be referenced include the grammar used to derive the EMMA result, the original raw signal (including the signal’s MIME media-type as defined by RFC2046²), and the source of input (e.g. NC-61 microphone).

In summary, EMMA shows much potential with regards to being an expressive and flexible markup language. It is designed to be generic enough to support a wide range of applications that, like XML, will extend far beyond the Web itself. The main reasons why EMMA is not used within the MSA/BPN are as follows:

1. The EMMA specification is still a Working Draft.
2. Few components (e.g. speech and handwriting recognizers) are currently able to produce EMMA documents, and this is particularly the case for embedded recognizers running locally on mobile devices.
3. The MSA/BPN applications only require a limited degree of communication because all of the processing (with the exception of extra-gesture recognition) is performed locally on the mobile device. In particular, the speech, handwriting, and intra-gesture recognizers all communicate over the internal blackboard architecture. Once the modality events are written to the internal blackboard, the data is stored as interaction nodes in the system’s working memory, and this data is then accessible in this format by components such as the interaction manager and the presentation planner.
4. The XML interpreter used by the mobile device (i.e. the default parser also used by Microsoft Internet Explorer) is resource intensive and documented to have memory leaks in its current form, making the handling of large amounts of XML unviable. XML in the MSA/BPN is limited in use to the representation of data containers (required when synchronizing with a shelf) and for communicating recognized extra-gesture events.

By definition, the use of EMMA would require modification to all of the components generating the markup, such as the individual recognizers (speech, handwriting, gesture), and all of the

²RFC2046: Multipurpose Internet Mail Extensions (MIME), <http://www.ietf.org/rfc/rfc2046.txt>

components required to interpret or modify the markup, such as the modality integration components, interaction manager, and potentially also the presentation planner. Thus, a sophisticated communications protocol for the small amount of required communication is considered to be excessive, especially since almost all of the input is recognized and interpreted on the one embedded device. A compromise for the future might be to implement a subsection of the EMMA specification, leaving out aspects that are not required by the MSA/BPN. With the exception of a few attributes, like those relating to the annotation of words in a word lattice (i.e. functionality not available in the MSA/BPN embedded speech engine), many of the metadata attributes defined in EMMA are in fact also defined in the MSA/BPN.

NLSML: Natural Language Semantics Markup Language NLSML (W3C-NLSML, 2000) is a natural language semantics markup language being specified by the W3C Voice Browser Working Group for describing the meanings of individual natural language utterances, where the word ‘utterance’ is defined in the specification as being a meaningful user input in any modality supported by a particular platform. The objectives of NLSML are to accurately reflect the user’s intended meaning in terms of the application’s goals and to also account for vagueness and ambiguity. It is an XML markup language which is expected to be generated by components such as speech recognizers, natural language understanding, dialogue, and multimedia integration components. Components that use NLSML might for example include multimedia integration components and dialogue managers. The focus lies in the representation of single utterances rather than an entire dialogue, and NLSML is still a Working Draft that has not yet been endorsed by the W3C.

An NLSML document consists of a single result containing one or more interpretations. Each interpretation can be further decomposed into the elements: model, instance, and input. Similar to EMMA, the ‘model’ element represents a data model containing the application-specific semantics. This data model consists of groups which may contain other groups or simple types such as string, Boolean, number, monetary values, date, time of day, and duration. In contrast to EMMA, the model is fixed to a particular format, being that of the draft W3C XForms specification (W3C-XForms, 2006). XForms are similar to HTML forms except that they are represented in XML. The ‘instance’ element, also similar to EMMA, represents the instantiation of a data model for a given utterance. The third type of element in an interpretation is the ‘input’ element. This contains the textual representation of a user’s input and includes attributes such as timestamp-start, timestamp-end, mode (e.g. speech), and confidence. Aside from text, an input element may also be represented by the ‘noinput’ or ‘nomatch’ elements, which are defined as being relevant for multimodal integration.

In summary, the NLSML specification focuses on how one can express the semantics of a user utterance (i.e. the model and instantiation) rather than how one can annotate the input for multimodal integration as in EMMA. The specification’s focus on data modelling is in turn highly dependent on the draft W3C XForms specification. Although NLSML briefly tries to account for multimodal integration by defining several attributes for the input element such as time and confidence, the representation is weak compared to that defined in EMMA, leaving certain aspects largely undefined, including for example information on the actual device used for input. NLSML does not define in detail the representation of multimodal input, instead concentrating on speech-only input and only briefly on speech-dtmf (Dual Tone Multi Frequency) interaction, which may arise when talking on a telephone and pressing numbers on the telephone’s keypad. Overlapped

inputs are not discussed at all in the specification, where it is stated that “the representation of multimodal input is deferred until the specification for multimodal inputs is better defined”. The document further states that ambiguities can currently only be represented at an interpretation level (i.e. based on an entire utterance) rather than on a word level due to validation issues arising from the use of XForms, and that the source of ambiguities is also not able to be modelled in the language. These limitations will perhaps be resolved in future revisions of the Working Draft, the most recent of which dates back to the year 2000.

M3L: MultiModal Markup Language M3L (Herzog et al., 2004) was created as part of the SmartKom project. It was designed for the representation and exchange of multimodal content between the various input and output processing components such as speech recognition, gesture recognition, face interpretation, media fusion, and presentation planning. Similar to the EMMA specification, M3L allows for the representation of data model and method annotation information, but unlike EMMA the data models are actually specified as part of the M3L language. Based on XML, the M3L language is formulated as a set of around 40 schema specifications.

The 40 schema specifications are divided into groups labelled basic, extended, and complex. ‘Basic’ data types specify concepts like integer, Boolean, float, time, and string, while ‘extended’ data types build on basic types by, for example, specifying numbers to be either positive or negative and time to include attributes such as century, year, month, ..., and millisecond. ‘Complex’ data types are used to specify a variety of concepts such as lists, geometry, money, person, and address. Domain knowledge describing the intentions of both the user and the system is defined offline using the OIL (OIL, 2000) representation and notation framework. The M3L ontology comprises more than 700 concepts and about 200 relations, which describe the abstract objects needed to communicate about the whole set of functionalities (Reithinger et al., 2003). The conversion from OIL to M3L is automated by a tool called OIL2XSD (Gurevych, Merten, & Porzel, 2003), which transforms an ontology written in OIL into an M3L compatible XML Schema definition, capturing the hierarchical structure and a significant part of the semantics of the ontology. The XML schemas can be viewed as Typed Features Structures (TFS) and allow for automatic data and type checking during information exchange. M3L was not devised as a generic knowledge representation language (Herzog et al., 2004), which would require an inference engine in every component for the exchanged structures to be interpreted adequately. Instead, very specific element structures are used to convey meaning on the syntactic level. As such, only closed world reasoning is supported, in which everything that the user and the system can talk about is encoded in the ontology.

In contrast to EMMA and NLSML, which are being designed to be generic markup languages, M3L was built to cater specifically for the needs of a single multimodal dialogue system. This engineering-oriented approach is evident in that the application-specific interaction defined in the ontology also forms part of the actual markup language. The language does however clearly differentiate between application-dependent and application-independent information and is thus still reusable by other systems. This is also demonstrated by the fact that it is in use by two separate multimodal dialogue systems: SmartKom (Wahlster et al., 2001) and COMIC (Catizone, Setzer, & Wilks, 2003).

MMIL: MultiModal Interface Language MMIL (Kumar & Romary, 2003) is an interface language that is expressed in XML Schema and was initially designed for the project MIAMM

(MIAMM, 2004; Reithinger et al., 2003), but has since also been incorporated into the project OZONE (OZONE, 2004). It is used as the exchange format between modules such as the dialogue manager and multimodal input and output components, and it allows for the incremental integration of multimodal data to achieve a full understanding of the multimodal acts within the system (Kumar & Romary, 2003). MMIL is expected to be used by agents such as speech interpreters, multimodal fusion modules, action planners, and modality advisers, as well as response generation modules dealing with multimodal fission, speech synthesis, and visual feedback.

The interface language is stated to be a general format for representing multimodal content at lower levels of representation like linguistic analysis and higher levels of representation like communication within the dialogue manager. It contains both generic (i.e. task-independent) descriptors, relating to dialogue management and interaction concepts used within the system such as timestamp and confidence, and domain specific descriptors (i.e. task-dependent), relating to specific domain tasks.

MMIL is based on a flat meta-modal data representation that combines any number of two types of entities, namely events and participants. ‘Events’ describe temporal entities expressed by the user or occurring in the course of the dialogue (e.g. interaction events that are provided via speech or haptic input). ‘Participants’ refer to individuals or sets of individuals about which a user says something or the dialogue system knows something about (e.g. the user, or graphical objects). Events and participants are further described by two properties: restrictions (providing more precise information about the event or the participant) and dependencies (relating events and participants to one another) (Reithinger et al., 2005).

In contrast to the aforementioned representation languages, which were all generally seen to distinguish between ‘task-specific’ data models and ‘task-independent’ method annotations (e.g. confidence and timestamp), MMIL places its primary focus on the grouping of input into a further layer of complexity defined by ‘events’ and ‘participants’. This is stated to be necessary because the language is capable of representing information on a variety of different levels such as phone, word, phrase, utterance, and dialogue levels.

5.1.3 Multimodal Input and Knowledge Representation in the MSA/BPN

Multimodal user input provides information on both the *data* contained in the user’s input and the *method* in which the input was provided (e.g. speech or handwriting). The modelling of both these types of information is crucial for an accurate understanding of the user’s input during the process of modality fusion. The goal of this section is to specify the data and method attributes as well as the communication acts that are used in the MSA/BPN. In the MSA/BPN, data attributes are mapped to predefined communication acts, while method attributes provide information like timestamp and confidence value, required during the fusion of multimodal input. The set of communication acts used in the MSA/BPN forms a knowledge source. This knowledge source can be seen to also define constraints on the valid relationships between different attributes and attribute-value pairs. For example, ‘type constraints’ are used to restrict the values that a certain attribute may contain (e.g. string, integer) and ‘number constraints’ are used to limit the number of objects that a user can request during product comparisons.

5.1.3.1 Data and Method Attributes in the MSA/BPN

The data and method attributes used by the MSA/BPN are shown in table 5.1. ‘Data attributes’ store information on what the user is trying to communicate, while ‘method attributes’ store information on how the user is trying to communicate it. Both of these attribute sets are stored in interaction nodes that are generated when a user communicates with the application, for example through the action of speaking to the system or by picking up objects from a shelf. These nodes are stored on the multimodal blackboard where they are later parsed and unified based on predefined communication acts. The blackboard is modelled in program code as a doubly linked-list. Many of the attributes are stored directly in working memory and are only accessed by the system components in this format. This is the case for speech, handwriting, and intra-gesture input. Extra-gestures are in contrast recognized by the server rather than by the mobile device, and they are communicated to the mobile device as XML, over a WLAN connection. As shown in the extra-gesture XML segment below, the attributes supplied by the server are only a subset of the overall attribute listing and also differ slightly in name. This is because some of the attributes such as those concerning timestamp information are generated on the client mobile device upon receiving the events. The difference in attribute names is due to the different processing stages of the system, for example ‘modality_type’ in the XML segment below is represented as ‘ModalityType’ in table 5.1. The Document Type Definition (DTD) defining the permissible syntax and structure of extra-gesture events is part of a separate document and is not included in the XML segment below.

```
<EVENT>
<MFACTION mfid="mfid" modality_type="gesture" gesture_type=
"disappeared" object_name="PowerShot S45" sensor_type="RFID"
timestamp="timestamp" confidence="1.0"></MFACTION>
</EVENT>
```

The data and method attributes used within the MSA/BPN are grouped under the parent classes ‘data’ and ‘method’. The definition and the data type of each attribute (e.g. enumerator, string, double, long) are outlined below. Where appropriate, examples of each attribute’s associable values are provided, and it is also described where in the system these attributes are assigned a value, i.e. within what module the values are generated and where the attributes are later required by the system for modality fusion. A summary of the attributes can also be seen in table 5.1.

- **Data attributes:** Data attributes store information on what a user is trying to communicate. This set of attributes can be categorized by the type of semantic constituent (used to label a user’s input), three types of data variables (used to hold recognized and interpreted text, and interpreted object references), and the N-best lists (used to store text and object references and their associated confidence values).
 - **SemanticConstituent:** The semantic constituent attribute is an enumerator that is used for labelling a user’s input as being one of the set of values: unknown, query, command, feature, feature_descriptor, object, and object_set. Take as an example the user query: “What is the price of the PowerShot S50?”. After parsing, such an utterance would result in two entries on the modality fusion blackboard, one with a SemanticConstituent of type ‘feature’ (relating to the input ‘price’) and a second with a SemanticConstituent of type ‘object’ (relating to the input ‘PowerShot S50’). Such values are

Data Attributes		
Data Type	Name	Values (where appropriate)
enum	SemanticConstituent	UNKNOWN, QUERY, COMMAND, FEATURE, FEATURE_DESCRIPTOR, OBJECT, OBJECT_SET.
CString	DataStringRAW	
CString	DataString	
CObjectNode	DataObject	
CString	NBestDataStringValues[3]	
CObjectNode	NBestDataObjectValues[3]	
double	NBestConfidenceValues[3]	
Method Attributes		
Data Type	Name	Values (where appropriate)
enum	ModalityType	UNKNOWN, SPEECH, HANDWRITING, GESTURE.
enum	GestureType	UNKNOWN, EXTRA-POINT, EXTRA-PICKUP, EXTRA-PUTDOWN, INTRA-POINT, INTRA-SLIDE.
double	Confidence	
enum	SourceOrigin (SourceName)	PDA, SERVER.
CString	SourceDeviceName	
CString	SourceDeviceDesc	
long	SourceDeviceID	
SYSTEMTIME	TimestampRAW	
long	Timestamp	
long	TimestampStart	
long	TimestampFinish	
enum	TimeType	PRESENT, PAST, NONE.

Table 5.1: Data and method attributes used by the MSA/BPN application.

generated by a semantic interpreter when user input is parsed for application-specific keywords. The SemanticConstituent attribute is required at a later processing stage when deciding on and populating an appropriate communication act. Typical examples of the application-specific keywords that semantic constituents may refer to are listed below. This list would differ for each application context, and the context in this particular instance is that of shopping.

- * **Unknown:** The type of semantic constituent is unknown.
- * **Query:** One of the open set of values: {what is, how many, ...}.
- * **Command:** One of the open set of values: {compare, find, ...}.
- * **Feature:** One of the open set of values: {price, megapixels, optical zoom, ...}.
- * **Feature Descriptor:** One of the open set of values: {under 500 EUR, ...}.
- * **Object:** Further divided into:
 - **O_TypUnID:** A non-uniquely identifiable object reference: {thing}.
 - **O_TypPartialID:** A partially identifiable object reference: {product}.
 - **O_TypID:** A type-identifiable object reference: {camera, grocery, language technology product, ...}.
 - **O_ID:** A uniquely identifiable object reference: {PowerShot S50, ..., baked beans, ..., IBM EVV, ...}.
- * **Object.Set:** One of the open set of values: {things, products, cameras, ...}.
- **DataStringRAW:** An attribute of type string, used to hold recognized user input that is expressible in text format. This attribute's value is generated during the recognition of all speech and handwriting input as well as the recognition of some intra-gesture input,

namely intra-gestures occurring when a user points to a keyword as it scrolls past on the GUI. The recognition of object references such as cameras, conducted via intra- and extra-gestures, are mapped directly to internal object instantiations and stored in the variable `DataObject` instead. This attribute is used to keep a record of the user's recognized input before it is parsed by a semantic interpreter.

- **DataString:** An attribute of type string, generated by a semantic interpreter and based on the information contained inside the variable `DataStringRAW`. While `DataStringRAW` might contain the value “What is the price of the PowerShot S50?” or “What does the PowerShot S50 cost?”, `DataString` will contain the semantic mapping of this input. In the given example, both utterances will result in a feature node with `DataString`=“price” and an object node with `DataString`=“PowerShot S50”).
 - **DataObject:** A pointer to an object node of type `CObjectNode`. This attribute is generated when intra- and extra- gesture object references are recognized, and it contains a convenient link to a recognized object and much of its associated information.
 - **NBestDataStringValues[3]:** An array of strings used to store the 3-best list of interpreted data string values. The value of the variable `DataString` corresponds to the first entry in this array.
 - **NBestDataObjectValues[3]:** An array of `CObjectNode` pointers used to store the 3-best list of interpreted `DataObject` values. The value of the variable `DataObject` corresponds to the first entry in this array.
 - **NBestConfidenceValues[3]:** An array of doubles used to store the confidence values associated with the 3-best values defined in `NBestDataStringValues[3]` or `NBestDataObjectValues[3]`. Each blackboard node will only ever contain a single type of semantic constituent, and thus this array will either apply to a set of string values or to a set of object values, but never both.
- **Method attributes:** Method attributes store information on how a user is trying to communicate with the system. This set of attributes relates to the modality used during communication, the underlying input devices used, timestamps, and confidence values. Their primary use within the MSA/BPN occurs in the modality fusion component, where interaction nodes on the blackboard are filtered through a range of constraints before being fused to form a modality-free result, corresponding to one of the predefined communication acts for a particular scenario like that of shopping.
 - **ModalityType:** An enumerator used for labelling the mode of communication with one of the set of values: unknown, speech, handwriting, and gesture.
 - **GestureType:** An enumerator used to further label the ‘gesture’ mode of communication into one of the set of values: unknown, extra-point, extra-pickup, extra-putdown, intra-point, and intra-slide.
 - **Confidence:** A double ranging from 0.0 to 1.0. This value is used to store the confidence value associated with the current best data attribute (either data string or data object depending on the node at hand), where 0.0 represents ‘least confident’ and 1.0 represents ‘most confident’. This is a convenience variable and overlaps with the first double in the array `NBestConfidenceValues[3]`.

- **SourceOrigin:** An enumerator used for labelling the origin of the input with one of the set of values: PDA and server. The values PDA and server can also be seen in the MSA/BPN to respectively correspond to input that is derived locally or distributed/remote.
- **SourceDeviceName:** An attribute of type string, used to identify the module responsible for providing the data attributes. Example values for this variable include: ‘IBM EVV’, ‘Microsoft Transcriber’, ‘FEIG-Electronic RFID’.
- **SourceDeviceDesc:** An attribute of type string, providing additional description information on a particular source device, such as manufacturer, full name, and version number.
- **SourceDeviceID:** An attribute of type long, used to uniquely identify a source device.
- **TimestampRAW:** An eVC++³ structure representing date and time information that is stored in the members: year, month, day, weekday, hour, minute, second, and millisecond. This provides an absolute timestamp for each modality event written to the blackboard and contrasts to the Timestamp variable described below, which provides a relative timestamp based only on the current day. The attribute’s primary use is for logging parsed user interactions in the discourse history log file.
- **Timestamp:** An attribute of type long, used to represent timestamp information in milliseconds, based on the current hour, minute, second, and millisecond. Millisecond information is not readily available on current mobile devices and is thus set to 0. This timestamp discards year, month, and day information that is obtainable from the TimestampRAW variable. The timestamp also differs from traditional Unix and ANSI/Windows timestamps, which respectively use January 1 1970 and January 1 1601 as their epoch. The simplification is well suited to the mobile applications at hand because the multimodal interactions it is used for generally only take between a few seconds and several minutes. An example conversion into milliseconds would see the time 14:27:31:000 result in the timestamp 52051000. The timestamp is generated each time a modality event is written to the blackboard. Interactions that span over several seconds, such as speech and handwriting, have only their finish time stored in this attribute. See TimestampStart and TimestampFinish for further information.
- **TimestampStart, TimestampFinish:** Based on the Timestamp attribute, these variables are used by modalities in which interaction spans longer periods of time such as speech and handwriting.
- **TimeType:** An enumerator used for annotating modality input with a time classification belonging to the set of values: present, past, and none. References that are made during a current interaction (e.g. to objects on the display) are labelled ‘present’, and these are then downgraded to ‘past’ after the interaction has been parsed or after 30 seconds have elapsed. When modality input is removed from the blackboard and written to the discourse history log file it is labelled with a TimeType of ‘none’. The TimeType attribute is primarily required for anaphora resolution, where past references may carry on into future utterances.

³eVC++: the embedded Visual C++ programming language

5.1.3.2 Communication Acts in the MSA/BPN

The previous section outlined the data and method attributes in use by the MSA/BPN mobile applications. These attributes contain information on a user's multimodal utterance and in particular on what the user uttered and on how it was uttered. Stored within interaction nodes on the modality fusion blackboard, the attributes are categorized by the type of semantic constituent (i.e. query, command, feature, feature_descriptor, object, object_set) and the associated modality (i.e. speech, handwriting, gesture). It is the role of the modality fusion module to filter and select nodes from the blackboard that best fit one of a range of predefined communication acts. These communication acts outline valid user-system interactions. A communication act, as implemented in the MSA/BPN, can be seen to represent one or more frame-based structures that are used to define a particular task such as querying a feature of a particular object. Each frame's contained elements (or slots) define variables whose values are to be ascertained (i.e. attribute-value pairs) as shown in figure 5.4. In the MSA/BPN, communication acts for shopping and navigation are defined in the program code of the application. This contrasts to systems such as MIAMM (MIAMM, 2004), where much of the user-system interactions are defined in communication acts that may be expressed in XML and configured externally. A range of communication acts, including examples of their use in the MSA shopping scenario, are described in table 5.3. These acts also demonstrate the flexibility and extent that such a modality-free language can cater for.

The modality-free communication acts specified for the MSA consist of the following elements, based upon which it is possible to define the 11 communication acts outlined in table 5.2:

<SM>	= Sentential Mood		
	= <Q> = Query	= who/how... (Q_{wh}), yes/no (Q_{yn}).	
	= <C> = Command	= compare (C_c), find (C_f),	
<F>	= Feature	= price, megapixels, optical zoom,	
<FD>	= Feature_Descriptor	= under 500 EUR,	
<O>	= Object = O_TypUnID	= thing.	
	O_TypPartialID	= product.	
	O_TypID	= camera, groceries, nlt.	
	O_ID	= PowerShot S50, ..., baked beans, ..., IBM EVV,	
<O _{set} >	= Object_Set	= things, products, cameras,	

wh-yn query	find command	compare command
<Q _{wh-yn} ><F> <O>	<C _f > <O>	<C _c > <O><O>
<Q _{wh-yn} ><F> <O> ⁺	<C _f ><F FD> <O _{set} >	<C _c ><F> <O><O>
<Q _{wh-yn} ><F> ⁺ <O>	<C _f ><F FD> ⁺ <O _{set} >	<C _c ><F> ⁺ <O><O>
<Q _{wh-yn} ><F> ⁺ <O> ⁺		<C _c ><F> ⁺ <O><O> ⁺

Table 5.2: Summary of the communication acts.

The division of the communication acts into queries and commands relates to each utterances sentential mood. An utterance may be used to either 'assert', 'query', or 'command', and the way in which a sentence is used is known as its 'mood'. There are several sentential moods:

Communication Act			Example				
Q, C	F, FD	O	Query (Q), Command (C)	Feature (F), Feature Descriptor (FD)	Object 1 (O ₁ , O _{set})	Object 2 (O ₂)	
Q _{wh-yn}	F	O	✓	Does (O ₁)	have a wireless control?	the PowerShot S50	
Q _{wh-yn}	F	O ⁺		What is	the price of	the PowerShot S45	and the PowerShot S50 ...?
Q _{wh-yn}	F ⁺	O		What is	the price of, and how many megapixels does (O ₁) have?	the PowerShot S45	
Q _{wh-yn}	F ⁺	O ⁺		What is	the price of, and how many megapixels do (O ₁) and (O ₂) have?	the PowerShot S45	and the PowerShot S50
C _f		O	✓	Find (O ₁).		the PowerShot S50	
C _f	F FD	O _{set}		Find	(O _{set}) with a price less than 500 EUR.	cameras	
C _f	[F FD] ⁺	O _{set}		Find	(O _{set}) with a price less than 500 EUR, and more than 4 megapixels.	cameras	
C _c		OO	✓	Compare		the PowerShot S45	to the PowerShot S50.
C _c	F	OO		Compare	the price of	the PowerShot S45	with the PowerShot S50.
C _c	F ⁺	OO		Compare	the price of, and the number of megapixels of	the PowerShot S45	with the PowerShot S50.
C _c	F ⁺	OO ⁺		Compare	the price of, and the number of megapixels of	the PowerShot S45	with the PowerShot S50, and the PowerShot S70.

Table 5.3: Examples of the use of the aforementioned communication acts, with the slots: Query (Q, Q_{wh-yn}=who/how..., yes/no), Command (C, C_f=find, C_c=compare), Feature (F), Feature Descriptor (FD), and Object (O, O_{set}=Object set, e.g. ‘cameras’). The fourth column containing the symbol ‘✓’ denotes the acts implemented in the MSA.

assertion (or declarative), query (or interrogative, e.g. yes/no and wh-questions), and command (or imperative) (Allen, 1995). The language models in the MSA make use of all of these moods, for example a user might say (in reply to a recognizer miss-recognition) “That is incorrect” (assertion), or “Find me the EOS 10D” (command), or “What is the price of the EOS 10D?” (query).

Table 5.3 provides some examples demonstrating the use of each of the 11 different communication acts. The examples are represented using natural language speech input, but may equally have been represented in the form of keywords using the modalities of handwriting or gesture, for example [$\langle Q_{wh-yn} \rangle \langle F \rangle \langle O \rangle =$ “Price, PowerShot S50” or $\langle C_c \rangle \langle F \rangle \langle O \rangle \langle O \rangle =$ “Compare, Price, PowerShot S50, PowerShot S70”. The abbreviations that are used are taken from the elements defined above. Elements in parenthesis such as (O₁) are used to link the semantic constituents in an utterance whose word order does not flow from left to right in the table. The repeated phrases operator ‘+’ is also employed in the table. This symbol is common in grammars that define arbitrarily long sequences of phrases, and can in tables 5.2 and 5.3 be defined as meaning two or more occurrences of the preceding non-terminal.

The 11 defined communication acts represent a theoretical specification. Three communication acts were implemented in the MSA to demonstrate the principles and concerns of modality fusion. These are marked with a ‘✓’ in table 5.3 and are illustrated in figure 5.4.

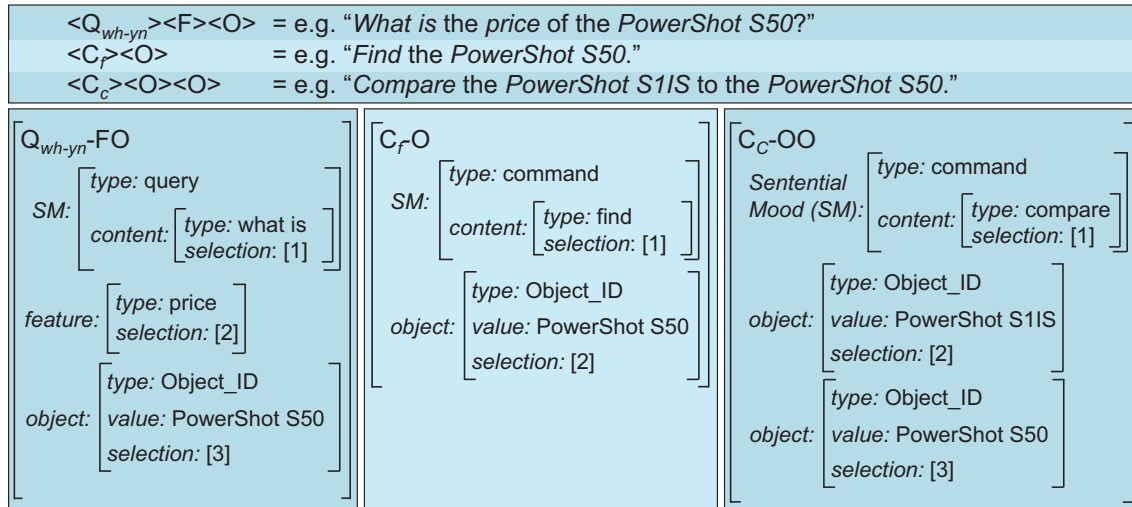


Figure 5.4: Three communication acts typically used in the MSA for queries and commands, also showing the inherent underlying frame-based structures. The ‘selection’ pointer in the structures links to additional information on the data and method attributes for each node.

5.2 Timing, Timeframes, and Saliency in the MSA/BPN

This section discusses three particular timing aspects that are relevant to the processing of multimodal input in the MSA/BPN, i.e. when to activate the modality fusion component, how much time to allocate to a user before considering a current user interaction to be complete, and the use of saliency to incorporate references that fall outside the time bound of a current user interaction.

5.2.1 Activating the Modality Fusion Component

In the MSA/BPN, the modality fusion blackboard acts as an interface between the individual recognizers running in parallel and the modality fusion component whose job it is to fuse multimodal interactions. Multimodal systems based on spoken dialogue communication often only initiate the modality fusion process based on information received in the primary modality, i.e. speech. A more ideal approach to initiating the process of modality fusion would be to create a program thread that constantly checks for elements on the blackboard to try and fit these to the set of available communication acts. Such an approach is ideal because activation of the modality fusion component is then not based on any single modality. The approach is however also resource intensive, especially if the timeouts for checking the blackboard are to occur frequently.

In the MSA/BPN, a hybrid approach to activating the modality fusion component has been implemented to cater for the limited resources of the mobile device. Rather than activation being tied to a single specific modality or to a dedicated timeout thread, the modality fusion process is initiated based on the type of semantic input written to the blackboard. Table 5.3 in section 5.1 outlines a wide range of communication acts applicable to the MSA/BPN shopping scenario, and it can be seen that each communication act consists of either a query combined with a feature, henceforth called ‘query+feature’, or a ‘command’. Each time information is written to the modality fusion blackboard, it is checked whether either of these two semantic structures exist. This approach allows a user to browse objects without triggering the modality fusion component and has the added

advantage that it does not bias one modality over another. ‘Query+feature’ information is also often represented as only ‘feature’ information, particularly in modalities like handwriting and gesture where the ‘query’ component is omitted, leaving only the feature information available for capture by the system (e.g. “[What is the] price?”).

5.2.2 Allocating an Appropriate Timeframe for Terminating a Current User-turn

Once the modality fusion process has been triggered, the system must determine an appropriate period of time in which it allows the user to finish entering input. State-of-the-art multimodal systems differ significantly with respect to this aspect. COMIC (Boves et al., 2004) for example provides a user with a system-initiated fixed time window in which to enter speech and pen input following the end of a system prompt. User studies on this system revealed however that subjects found interaction difficult when being dictated both when to start and stop an interaction, and such interaction resulted in low recognition rates. In (Chai et al., 2004), interaction by the user is initiated by the user, while the timeout is based on a two second period of inactivity. This approach does not however take modality differences into account, and as a result if gesture input occurs more than two seconds before speech input, the interaction is not correctly recognized. Such an approach would be detrimental to the MSA because the use of the communication mode gesture can be seen to be particularly prevalent for selecting and browsing products at a user-defined speed. In MATCH (Johnston et al., 2002) and later in the EMBASSI project (Elting et al., 2003), so-called ‘intelligent timeouts’ are used to identify the end of a user interaction. These timeouts are conditional on the other input modes in that the system identifies which recognizers are still active after the user begins interacting. If no additional devices are pending results, the interaction is considered complete. These systems are good in that they check for active modalities before considering an interaction complete, but it is unclear whether or not additional periods of inactivity are also used. Without the use of an inactivity period, only temporally overlapped multimodal input will be correctly recognized, leaving multimodal input interactions that are provided sequentially in time to be terminated prematurely. In (Gupta, 2003), a statistical linear predictor is used to adaptively determine an expected timeframe based on criteria such as statistical averages of the time required to enter input in a given modality and whether multimodal input is currently available for processing.

As described below, the MSA/BPN approach for determining an appropriate timeframe for user interaction incorporates many of the lessons learnt from the above described systems. User interaction in the MSA/BPN is initiated by the user, timeout periods are dependent on the type (e.g. feature or object) and the order in which semantic information is written to the modality fusion blackboard (e.g. first feature then object or vice-versa), and timeout periods incorporate differences known to exist in the time required to use differing modality combinations (e.g. SGI has been shown to be a much faster modality combination than SH).

All of the above described systems, including the MSA/BPN, need to at some point in time consider the possibility that the user will not be providing any additional input to the system, even if no communication act can be successfully completed with the current information. If this occurs, a system might for example decide to classify such input as noise or initiate a system dialogue to request the missing information from the user. Dialogue management is not a topic of focus in the MSA/BPN. In the MSA/BPN, interaction nodes located on the modality fusion blackboard reduce in salience as time progresses, until they are eventually removed from the blackboard altogether. The solution to incomplete communication acts in the MSA/BPN is such that if a user believes

an interaction to have been unsuccessfully recognized, he or she can call up a dialogue box that shows the N-best list of results recorded by the system for each of the semantic constituents in the last user interaction and select values from these N-best lists, as shown in figure 5.11.

Timeframe allocation in the MSA/BPN: The processing of multimodal input in the MSA/BPN is activated when a query+feature or a command is written to the modality fusion blackboard by any of the system's recognizers. Once the modality fusion process has been activated, a time limit of 500ms is enforced on the user, after which time the modality fusion component attempts to find the appropriate communication act based on the available input (see figure 5.5). Subjective results and user observations gained during usability studies on the MSA have indicated that although 500ms is not a very long period of time to conclude an interaction, it is sufficient for users that temporally overlap modalities (e.g. speech combined with gesture), users that are experienced with the system, and users that provide input sequentially using only the fastest of modality combinations like SGI. If the modality fusion component can not successfully populate the slots of a communication act, because no (or not enough) information is present on the modality fusion blackboard, it will assign an additional time period of two seconds to the user to account for the possibility that information (e.g. objects) are provided by the user only after a query+feature or a command has been provided.

If at this stage the modality fusion component can still not successfully populate the slots of any communication act, it will assign a further period of time (e.g. one second, five seconds, or 10 seconds) in which the user is able to continue providing input to the system. This period of time is designed to account for the different modality preferences that users may have (e.g. SGI is a fast modality combination to use while SH is a slow modality combination) and can be altered manually in a user settings file. It is foreseeable that such user adaptation will be automated in the future based on statistical data that has been collected on the different modality combinations at CeBIT 2006 and also based on the temporal patterns that individual users might have.

If after these additional time allocations the modality fusion component is still unsuccessful in populating the slots of a communication act, recently mentioned elements in the history context are included to account for the possibility that the user interaction contained anaphora or ellipsis, and if this still proves unsuccessful, the system assumes the user input was noise and returns. Whereas spoken dialogue systems are often able to identify the use of anaphora based on the linguistic form of an utterance (e.g. "What is *its* price?", where 'its' denotes the use of anaphora), multimodal systems like the MSA/BPN, which also cater for modalities such as handwriting and gesture, are not always provided the information to do this (e.g. a user might simply scribble down the keyword 'price'). Ellipsis is similarly difficult to recognize through means other than the time at which the current query was issued and the referents already existing in the history context. In the MSA/BPN, the history context is defined as being all elements on the blackboard that have the TimeType 'past'.

Incrementally extending the user interaction timeframe based on the presence of certain conditions is important to avoid the appearance of a sluggish system, as each extension in time that is not utilized by a user will ultimately be seen as slowing down the system. For user interactions where all the required information is provided to the system in the current user-turn, the system will not allocate any additional time. From experimentation with the MSA/BPN, a non-utilized time period of up to two or three seconds still makes for a usable system over short periods of use, but is less appropriate for more familiar and advanced users of the system.

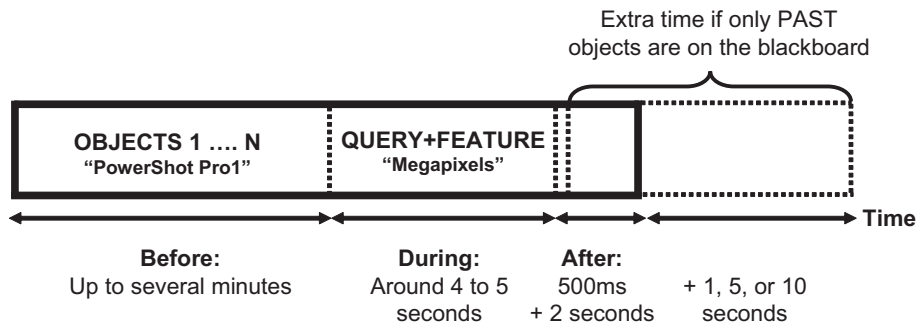


Figure 5.5: Timeframe for a typical user interaction in the MSA/BPN.

The use of statistical data to estimate the time required to complete different modality combinations: The timeframe in which a user may provide information to the system on a particular communication act can be seen to extend from the end of the previous communication act to 500ms (plus a variable amount of time defined in the user settings file) after the current communication act has been triggered. Table 4.7 in chapter 4 illustrates the total average time taken to interact in a variety of different modality combinations based on data collected at the CeBIT 2006 fair in Hannover. These times refer however only to complete interaction times. From data and experience gathered during this study and previous studies conducted at Conrad Electronic in Saarbrücken, the additional time periods outlined in table 5.4 were determined as being adequate to conclude an interaction in each of the given semantically non-overlapped modality combinations. The modality combinations depicted in this table are linked to a fixed semantic order of constituents, where the first modality always refers to a query+feature or a command and the second modality always refers to an object reference. This linking of semantics to modality type is important in the table because the timeout period for an interaction is based on a user issuing a query+feature or a command rather than an object. The alternative case where objects are provided first is far less challenging because these entries do not automatically activate the modality fusion process and thus timeout periods in the MSA/BPN.

Modality Combinations	Extra time (Secs)	Modality Combinations	Extra time (Secs)
SS	0 (5)	HGI	1
SH	10	HGE	1
SIG	0	GIS	5
SGE	1	GIH	10
HS	5	GIGI	1
HH	0 (10)	GIGE	1

Table 5.4: Additional time periods required for entering object information after a query+feature or command has been issued to the system. The parenthesized values for the unimodal combinations indicate alternate time periods based on whether query+feature/command and object input is provided together or during two separate actions.

The table shows that when both feature and object information are provided unimodally during

the same recognition event, no or only very little additional time is required for an interaction to be classified as complete. Entering object information in speech at the end of an interaction takes up to an additional five seconds of additional time and up to an additional 10 seconds of additional time for handwriting. Intra-gesture and extra-gesture require very little additional time to conclude an interaction due to gesture being a fast modality and a modality that is easily conducted in parallel, assuming that the required referents are visible to the user.

5.2.3 Saliency in the MSA/BPN

Determining the correct timeframe to use for modality fusion is not an easy task. If the timeframe is too long, some interaction nodes might be incorrectly included as part of a communication act. If the timeframe is too short, some interaction nodes might be incorrectly discarded. The occurrence of anaphora and ellipsis during multimodal dialogue discourse has the further effect that a communication act will not always be completely populated by events gathered during the current user interaction, even though an interaction timeout has completed successfully. It is for this reason that the timeframe for capturing a user-turn as described above can not alone be used to determine all relevant referents in a user interaction. The saliency of a referent is used to incorporate referents that may otherwise have been discarded, due to being outside the bounds of a valid user interaction timeframe.

In (Huls, Claassen, & Bos, 1995), *saliency* is defined to mean “notable significance” and is used to rate the likelihood that a reference is referring to one referent over another referent. A diversity of factors contributes to the saliency of a referent. In (Alshawi et al., 1987), referents are each assigned significance weights and a decay function that is used to decrease the weights over time. By using the notion of saliency, a system interpreting a referring expression can choose the most salient entity that meets the type constraints imposed by the available communication acts. Various aspects contribute to a referent’s saliency, including ‘recency of mention’, ‘markedness of expression’, and perceptual factors like ‘visibility’. In the MSA/BPN, recency of mention is based on the use of timestamps, while markedness of expression is based on the interpretation of confidence values, and visibility is based on analysing which objects are currently in focus in a visual sense, for example the elements that are currently visible on the PDA’s display and those elements that exist in the physical shelf that the user has synchronized with.

Past and Present Referents: One of the most significant contributors to the saliency of a referent is that of time, i.e. a referent’s recency of mention. In the MSA/BPN, each event on the modality fusion blackboard is classified by one of the following TimeTypes: present, past, and none. The term ‘present’ is used to denote elements written to the blackboard during the current user-turn. Elements that were used in the last successful communication act are labelled ‘past’, as too are present elements in the current user-turn that are older than 30 seconds. An element typed as past will retain this value for multiple user-turns only if it is continuously referred back to in following communication acts. Past elements no longer referred to in following communication acts are removed from the blackboard and written to the discourse history log file where they are relabelled with a TimeType of ‘none’. The advantage of basing temporal saliency on a combination of both discrete user-turns and time periods is that an element accessed even several minutes ago, will retain the TimeType ‘past’ for as long as the user has not concluded their user-turn, thus making it possible to refer back to the element with relative ease. This is particularly useful for the scenario

of shopping, where a user might first select a product (e.g. via extra-gesture by picking it up) and then analyse the product for sometime before eventually deciding to query its features.

Confidence Values and Perceptual Factors: Two other aspects modifying the salience of a referent are the confidence value and visibility of the referent. Confidence scoring was the topic of section 4.1.2, which discussed how confidence values were generated for each of the modalities used in the MSA/BPN. The more confident a recognition result is, the more salient that result should be. This is well demonstrated by the modality of speech, where a clear pronunciation of an object referent, i.e. markedness of expression, should make that referent more significant than a referent with a less clear pronunciation and thus lesser confidence value. Confidence values are used in the MSA/BPN during conflict resolution where, for example, multiple referents on the modality fusion blackboard have the same semantic type (e.g. feature or object) and occur within the same user interaction timeframe.

Perceptual factors are also used in the MSA/BPN to increase the salience of a referent. In particular, for the modality of gesture, only objects in the real-world that exist on the shelf that the user is currently synchronized with and only objects visible to the user on the PDA's display are considered viable referents. Similarly, only features that are currently visible on the device's display via the visual-WCIS scroll bar are considered viable feature referents to the system.

5.3 Modality Fusion in the MSA/BPN

In chapter 2, a range of benefits to the user were identified with regards to interacting multimodally, including naturalness, transparency, ease of use, ease of learning, flexibility, efficiency, and suitability for more challenging applications. The term 'modality fusion' was also introduced along with a range of synonyms for this term such as 'media fusion' and 'multimodal integration', and closely related terms like 'mutual disambiguation' were also briefly discussed. The goal of modality fusion was outlined to be the merging of multiple modality input streams into a single modality-free result that combines the semantic meaning from each of the individual input streams. In this section, the modality fusion strategies applicable to the MSA/BPN are discussed.

5.3.1 Previous Work on Defining Modality Fusion Strategies

A range of different modality fusion strategies have been implemented in a number of systems in the past. The strategies often differ due to a variety of factors like the domain of implementation and the modality combinations that are catered for. This section outlines several strategies, found in the literature, that relate to the fusion of multimodal input. The first strategy focuses on reference resolution and the synchronization of inputs, while the second strategy focuses on a set of unification operations used to merge old and new information arising over multiple user-turns in a dialogue discourse. The third strategy then briefly outlines an attempt to calculate the joint probabilities for redundant and complementary input. Although the approaches all differ slightly in focus from one another and to the MSA/BPN, where the main focus lies in the resolution of conflicts between semantically overlapped input arising from either same-type or different-type recognizers, the concerns that these strategies attempt to resolve are important for all multimodal systems.

Some systems place different importance on the communication modes available to their users. In (Chai et al., 2004), the primary communication mode is that of speech, while gesture is used as a lesser expressive secondary mode for the selection of map entities on a graphical display. Such multimodal systems are often adapted from spoken dialogue systems to allow for key parts of an utterance to be expressed more naturally by other means of communication like gesture. These architectures almost always cater only for the use of complementary multimodality (i.e. they do not support supplementary multimodality as defined in section 4.2.2) and are thus less able to effectively cater for mobile users in contexts that have changing environment characteristics (e.g. noise levels and crowdedness), in which certain parts of an utterance suddenly become better suited to input in other modalities due to reasons like accuracy and user privacy. The lack of complementary multimodal support also limits these systems' ability to capture semantically overlapped input in the different modalities, and this is also the case in (Chai et al., 2004) where overlapped input is only possible for the selection of graphical map entities but not, for example, for specifying attributes about these map entities.

The fusion strategy in (Chai et al., 2004) is based on a probabilistic approach to reference resolution, for which different types of reference are resolved using a graph-matching algorithm. Similar to other multimodal systems including the MSA/BPN, semantic, temporal, and contextual constraints (i.e. conversation history) are used to identify the most probable referents in an utterance. In the described system, information is gathered from the user in the form of speech input, gesture input, and conversation context, and is represented by three attributed relational graphs (ARGs). The gesture and conversation context graphs are then combined to form a referent graph, while the speech ARG is taken as the referring graph, and reference resolution then becomes a constrained probabilistic graph-matching problem that aims to find the best match between referent and referring graph. Due to timing limitations in the MSA/BPN, which was designed for mobile devices rather than a desktop computer as in the described system, the process of aligning information that occurs in the different modalities is only ever an estimate that is often based on timeframes rather than timestamps and the temporal order of input. The conversation history in the MSA/BPN is also only used as a last resort to finding any missing referents, rather than as an equal counterpart to speech and gesture input as in the described system. Furthermore, whereas the focus of the described work is on the alignment of possible references with their underlying referents, a major focus of the MSA/BPN is on the resolution of semantically overlapped references where it is already known which references belong to which referents and where the goal is thus on the use of certainty factors to disambiguate results contained in the N-best lists produced by same-type and different-type communication modes. Having said this, the ability to recognize utterances involving multiple referring expressions accompanied by multiple gestures and the ability to recognize single referring expressions accompanied by multiple gestures (both of which are a focal point in (Chai et al., 2004)) is also possible in the MSA/BPN. In the MSA/BPN, such utterances occur for example when the user says: "Compare this camera <Gesture> with this camera <Gesture>" (i.e. multiple referring expressions and multiple gestures) and "Compare these two cameras <Gesture><Gesture>" (i.e. single referring expression and multiple gestures).

Other works are designed, similar to the MSA/BPN, to cater for complementary multimodal interaction. MATCH (see section 3.1.4) is one such system, and uses a finite-state automaton for the multimodal integration of the communication modes speech and pen. In this approach, the parsing, integration, and understanding of speech and gesture inputs are captured in a single declarative multimodal context-free grammar that is compiled into a multimodal finite-state device. The finite-state device is simulated using two transducers, the first G:W to align speech

words and gesture input, and the second G_W:M to take as input a composite alphabet of speech words and gesture input and to use this to output meaning. In the MATCH system, the finite-state multimodal integration component is combined with a speech-act based multimodal dialogue manager, which allows multimodal commands to be distributed over multiple dialogue turns and also allows ambiguous multimodal input to be resolved using dialogue context (Johnston et al., 2002).

In (Johnston, 1998; Johnston et al., 1997), it is shown, as part of the QuickSet system (see section 3.1.3), how spoken and gestural input can be integrated by using a unification-based operation over typed feature structures that represent the semantic contributions of the different modes. This approach is superseded in (Alexandersson & Becker, 2003), where a second operation called overlay is defined to work alongside the unification operation. With regards to multimodal discourse processing, these two operations can be used to determine the consistency of frame-based structures, and in comparison to unification, which returns a null value each time a conflict is detected, the overlay operation uses its first argument as default during conflicts and thus always returns a result, even when conflicting information is present (Jan Alexandersson, 2006). The unification and overlay operations were conceived under the SmartKom project (see section 3.1.7) and have since been used as part of other projects like COMIC (see section 3.1.8), where it is stated that overlay is particularly useful when a user provides only slightly conflicting information, for example “Show me this bathtub in blue” (while pointing at a white bathtub) (Pfleger, 2004). In this case, it is described that the background and the covering information would combine to give an object representation of a blue bathtub with the remaining features of the white bathtub that the user pointed at.

Unification and overlay is generally needed when new information is added to existing older information during the course of a dialogue discourse spanning multiple user-turns, for example a computer-guide that helps users design their own bathroom. The goal of the MSA/BPN was, in comparison, to allow users to retrieve information over a set of objects using a very wide range of different modality combinations. Dialogue management is not the main focus of the MSA/BPN, where discourse between the user and the computer rarely spans multiple user-turns. Old and new information is however still unified in the MSA/BPN (assuming that type constraints between multiple objects are not breached), and to demonstrate, this occurs when a user selects a particular camera and then queries different attributes of the camera over multiple user-turns, for example “What is the price?” and then “How many megapixels does it have?”.

The modality fusion algorithm described in (Kaiser et al., 2003) uses a generalized chart-parser to fuse redundant and complementary information, based on a set of predefined rules, a type hierarchy, and a set of spatiotemporal constraints. In comparison to this system, which uses simple multiplication to derive the joint probability of multimodal speech and gesture input (a method that discounts joint probabilities for instances where one of the values is equal to ‘0’), the MSA delves deeper by first removing the bias that exists between the presumed confidence of competing recognizers and then deriving joint probabilities based on the use of certainty factors (see sections 5.3.3 and 5.3.6).

5.3.2 Processing Multimodal Input in the MSA

Figure 5.6 illustrates how multimodal user input in the MSA/BPN is converted from raw signals to machine-interpretable communication acts. The different processing stages are shown in the centre of the diagram with the associated input representations at each stage of processing shown on the right. The figure illustrates the representation of speech and gesture input, which is captured

as audio signals (when a user speaks into the PDA's microphone) and screen coordinates (from tapping with a stylus or finger on the PDA's display). These signals are parsed by modality-specific recognizers and are then mapped to semantic elements such as $\langle Q_{wh-yn} = \text{"What is"} \rangle$, $\langle \text{Feature} = \text{"price"} \rangle$, $\langle \text{Object} = \text{"camera"} \rangle$, and $\langle \text{Object} = \text{"PowerShot S50"} \rangle$. These elements are then fused to form a modality-free and unambiguous communication act.

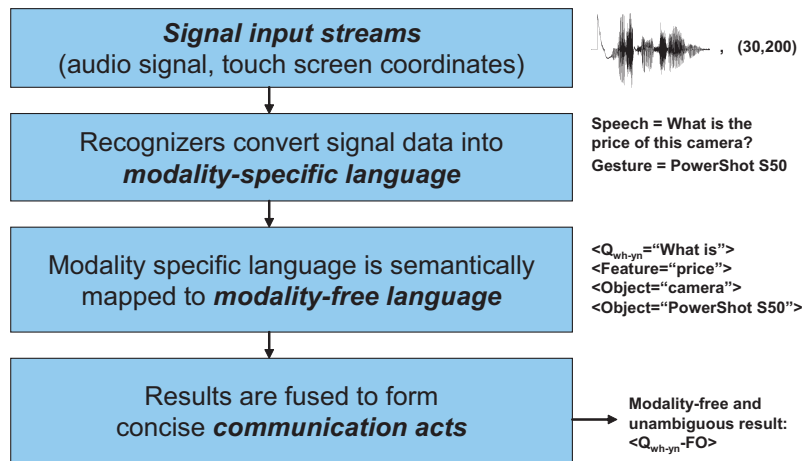


Figure 5.6: Processing multimodal input in the MSA/BPN.

5.3.2.1 Modality Fusion Architecture and Blackboard Design

The main components underlying the multimodal system architecture in the MSA/BPN are outlined in figure 5.7. In contrast to projects like QuickSet (Cohen et al., 1997) and SmartKom Mobile (Bühler et al., 2002), which have a heavy reliance on distributed and client-server architectures, the processing of interaction in the MSA/BPN, including recognition, interpretation, and fusion, is performed locally on the mobile PDA device. Only the SQL database of shopping products and map data is stored on an external server, and only the extra-gesture ‘pickup’ and ‘putdown’ events, which are based on RFID technology, are not recognized locally on the mobile PDA device. Once a user has synchronized with a data container (e.g. a shelf), the PDA device can be used entirely offline without any connection to the public server or public infrastructure. Even the extra-gesture ‘pickup’ and ‘putdown’ actions that allow for interaction with the real physical world can be replaced by other real-world interaction like the extra-gesture ‘point’ action that is supported by the PDA’s onboard CF card-slot camera.

During user interaction with the system, modality events are written to the central blackboard and stored as interaction nodes. These nodes provide the main source of information required for the modality fusion component to make informed decisions about the entities on the blackboard. Figure 5.8 shows a simplified graphical illustration of the nodes and their attributes, as previously outlined in table 5.1, including information on the type of semantic constituent (e.g. feature or object), the raw user input after recognition (e.g. the recognized screen coordinates for an intra-gesture event), the interpreted user input (represented as one or more semantic mappings to strings and/or object instantiations), the 3-best result matches including their confidence scores from 0.0 to 1.0, the parent modality group (i.e. speech, handwriting, gesture), an underlying modality type where appropriate (e.g. point, pickup, putdown), the origin of the event (e.g. PDA, server),

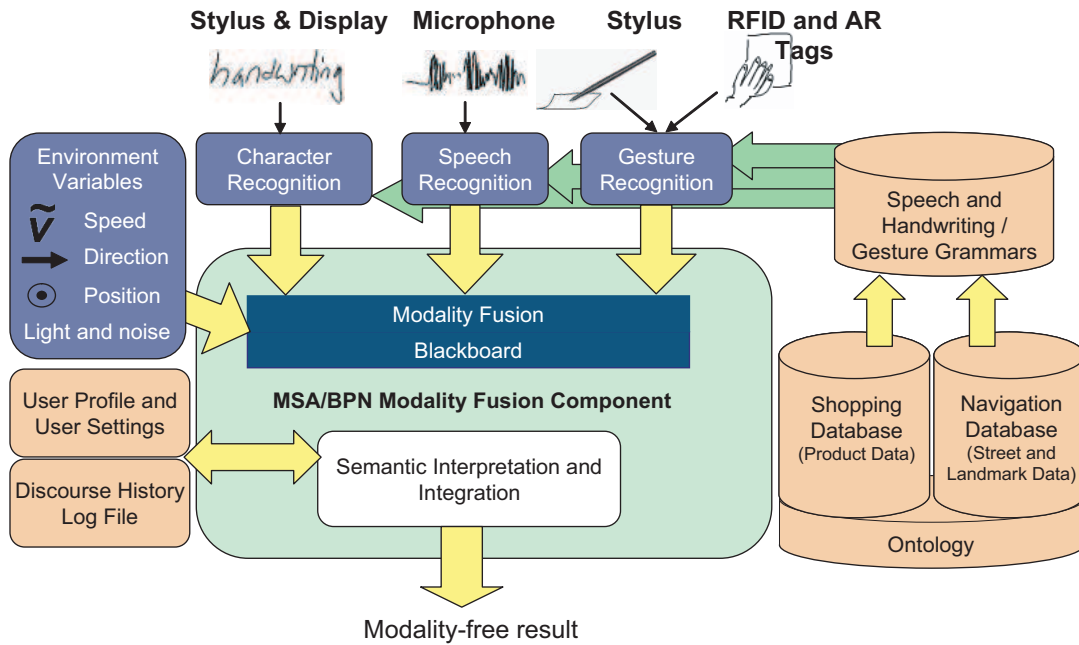


Figure 5.7: Modality fusion architecture: Data flow between the communication modes, the user settings and log files, the knowledge sources, and the modality fusion component.

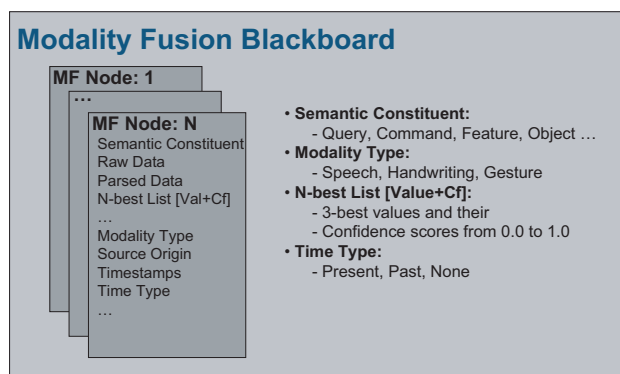


Figure 5.8: Modality fusion blackboard illustrating the main data and method attributes contained inside each interaction node on the blackboard.

timestamps for the semantic constituent (including start and finish times), and a time classification (i.e. present, past, or none).

5.3.2.2 Conflict Resolution in the MSA/BPN

This section describes the process of modality fusion and conflict resolution as it applies to the MSA/BPN. The section assumes, as described previously, that modality events have already been written to the central blackboard together with their method and data attributes, and that the modality fusion component has been triggered to commence processing.

The events that are written to the blackboard decay over time and are eventually removed entirely from the blackboard, at which stage they are recorded in the discourse history log file. The decay time is based on a combination of discrete user-turns and an elapsed time period of 30 seconds. ‘Present’ referents that are not used within 30 seconds of being referred to or have just been referred to in a current interaction are downgraded to the TimeType ‘past’. These past events are then written to the discourse history log file at the end of the following user interaction and stripped of their TimeType classification. At this point, there are several further stages of processing that the modality fusion component is responsible for:

1. Confidence values for each event are re-weighted to account for recognizer inconsistencies over the spectrum of all known recognizers, through the use of statistical probabilities collected during a field study on recognizer accuracy rates.
2. A communication act is chosen such that the events currently on the blackboard best match the slots of a communication act in the MSA/BPN’s predefined modality-free language, and where necessary conflict resolution is performed.
3. The events on the blackboard are then filtered in accordance to temporal constraints, so that only the most relevant nodes still exist for the process of modality fusion.
4. The communication act slots are then populated with the remaining events, and where necessary conflict resolution between semantic elements is performed.

This dissertation focuses in particular on points 1 and 4, i.e. methods for re-weighting confidence values so that they are comparable over a range of recognizers and the resolution of conflicts that occur between individual semantic elements written to the modality fusion blackboard.

5.3.3 Using Statistical Probabilities to Re-weight Confidence Values

Along with timeframes and timestamps, confidence values are one of the most important parameters used during modality fusion and particularly during conflict resolution. Confidence scoring was first discussed (with respect to each individual modality) in section 4.1.3 of this dissertation, but this discussion did not delve into the reliability of confidence scores generated by a single recognizer across a given range from 0.0 to 1.0, nor the comparability of confidence scores generated by two or more recognizers. Figure 4.12 illustrates the percentage of occurrences for which a given confidence value was generated during a field study and also includes the recorded accuracy for each of these confidence values. It can be seen that 25% of all speech occurrences were recorded with a confidence value of Cf=0.0. The same data is analysed in figure 5.9 and shows that 84.93% of the occurrences were in fact correct. In a unimodal spoken dialogue system, this

over-modesty in identifying correct and incorrect utterances has the negative effect that the system would need to confirm the input with the user 84.93% more often than required, perhaps with a clarification dialogue such as “Did you mean PowerShot S50 or PowerShot S60?”. A second problem is that in multimodal dialogue systems these confidence values are not by default comparable across multiple recognizers. Take for example the confidence values generated by the modalities speech and handwriting for object recognition (see OBJ_S and OBJ_H in figure 5.9). For speech, a confidence value of Cf=0.0 has an average accuracy rate of 96.10% (based on 77 occurrences) while for handwriting, a confidence value of 1.0 only has an average accuracy rate of 77.61% (based on 67 occurrences).

The inability to directly compare confidence values between recognizers, be that same-type recognizers (e.g. two speech recognizers) or different-type recognizers (e.g. a handwriting and a speech recognizer), is still greatly an unsolved problem in state-of-the-art systems. In the *Verbmobil* spoken dialogue system for example there is an entire chapter devoted to “Speech Recognition Performance Assessment” (Malenke, Bäumlner, & Paulus, 2000), in which the authors attempt to record the performance of three different speech recognizers over several years. The authors of the chapter state that “an immediate comparison of results achieved for different [speech engine] modules cannot be recommended” and rather than use the multiple engines in competition to generate more accurate results, each individual recognizer was assigned a specific purpose such that they did not compete with one another. The work in (Oviatt & Cohen, 2000; Wu, Oviatt, & Cohen, 1999) presents one solution to the problem in which probability estimates from different recognizers are created by a so-called MTC approach (Members-Teams-Committee). In this approach, the individual recognizers are termed ‘members’ and generate independent posterior estimates for recognition results. These estimates are then interpreted by modules called ‘teams’, which are trained on different datasets and can apply different weighting schemes to the estimates. Finally, a module called the ‘committee’ is used to rank the most relevant recognition results identified by the teams. Similar to the MSA/BPN, this approach leverages the fact that research on unimodal recognition techniques and the generation of unimodal posterior probabilities (i.e. confidence values) is relatively mature, and that these unimodal posterior probabilities can be used as a starting point in multimodal algorithms that re-weight recognition results for comparison over a range of different recognizers.

During a field study that was conducted on the MSA/BPN at the CeBIT 2006 fair in Hannover (see section 4.1.3), a dataset of accuracy statistics (as referenced extensively in this section) was accumulated for each of the individual recognizers. In the MSA/BPN, this dataset provides a means to re-weighting confidence values, to better reflect the accuracy of recognized input over the confidence range of a single recognizer and when comparing confidence values generated by different recognizers, be that same modality-type or different modality-type recognizers.

Figure 4.12 plots the occurrence and accuracy of feature and object input over a given range of confidence values for each of the recognizers. This figure is derived from the tables in figure 5.9, in which the number and percentage of correctly recognized occurrences are listed in the columns ‘Corr/Occur’ (i.e. Correct/Total Occurrences) and ‘% Correct’ for each of the 7 different semantic-modality categories - FTR_S, FTR_H, FTR_GI, OBJ_S, OBJ_H, OBJ_GI, and OBJ_GE - and their associated confidence value range from 0.0 to 1.0. The third column displayed for each of the semantic-modality categories in the table is titled ‘Weighted’ and refers to the re-weighted confidence values that are used in the MSA/BPN as a means to comparing results across the board of available recognizers. The re-weighted confidence values for the semantic-modality categories are also plotted with respect to confidence value in figure 5.10.

Cf	FTR_S			FTR_H			FTR_GI		
	Corr / Occur	% Correct	Weighted	Corr / Occur	% Correct	Weighted	Corr / Occur	% Correct	Weighted
0.0	124 / 146	84.93%	0.9078	0 / 0		0.3979	0 / 0		0.9404
0.1	32 / 32	100.00%	0.9175	1 / 1	100.00%	0.4690	0 / 0		0.9404
0.2	39 / 39	100.00%	0.9272	1 / 5	20.00%	0.5400	0 / 0		0.9404
0.3	40 / 42	95.24%	0.9368	0 / 3	0.00%	0.6111	0 / 0		0.9404
0.4	48 / 48	100.00%	0.9465	9 / 12	75.00%	0.6822	0 / 0		0.9404
0.5	46 / 46	100.00%	0.9562	11 / 15	73.33%	0.7533	0 / 0		0.9404
0.6	41 / 41	100.00%	0.9659	26 / 29	89.66%	0.8243	0 / 0		0.9404
0.7	28 / 28	100.00%	0.9755	46 / 54	85.19%	0.8954	0 / 0		0.9404
0.8	23 / 24	95.83%	0.9852	88 / 90	97.78%	0.9665	142 / 151	94.04%	0.9404
0.9	26 / 26	100.00%	0.9949	130 / 131	99.24%	1.0000	0 / 0		0.9404
1.0	82 / 84	97.62%	1.0000	114 / 114	100.00%	1.0000	0 / 0		0.9404

Cf	OBJ_S			OBJ_H			OBJ_GI			OBJ_GE		
	Corr / Occur	% Correct	Weighted	Corr / Occur	% Correct	Weighted	Corr / Occur	% Correct	Weighted	Corr / Occur	% Correct	Weighted
0.0	74 / 77	96.10%	0.9752	0 / 0		0.1861	0 / 0		1.0000	0 / 0		1.0000
0.1	19 / 19	100.00%	0.9783	0 / 0		0.2489	0 / 0		1.0000	0 / 0		1.0000
0.2	15 / 15	100.00%	0.9814	0 / 0		0.3117	0 / 0		1.0000	0 / 0		1.0000
0.3	19 / 19	100.00%	0.9845	0 / 0		0.3745	1 / 1	100.00%	1.0000	0 / 0		1.0000
0.4	22 / 22	100.00%	0.9876	0 / 0		0.4372	3 / 3	100.00%	1.0000	0 / 0		1.0000
0.5	24 / 24	100.00%	0.9907	1 / 2	50.00%	0.5000	4 / 4	100.00%	0.9996	0 / 0		1.0000
0.6	10 / 10	100.00%	0.9938	5 / 8	62.50%	0.5628	16 / 16	100.00%	0.9991	0 / 0		1.0000
0.7	10 / 10	100.00%	0.9969	16 / 19	84.21%	0.6255	79 / 79	100.00%	0.9987	0 / 0		1.0000
0.8	9 / 9	100.00%	1.0000	31 / 51	60.78%	0.6883	191 / 191	100.00%	0.9983	0 / 0		1.0000
0.9	7 / 7	100.00%	1.0000	55 / 86	63.95%	0.7511	247 / 248	99.60%	0.9979	0 / 0		1.0000
1.0	6 / 6	100.00%	1.0000	52 / 67	77.61%	0.8139	51 / 51	100.00%	0.9974	117 / 117	100.00%	1.0000

Figure 5.9: The tables show the statistical dataset and the re-weighted confidence values for each of the 7 different semantic-modality categories that were derived from field studies: FTR_S, FTR_H, FTR_GI, OBJ_S, OBJ_H, OBJ_GI, and OBJ_GE. The column ‘Corr / Occur’ represents the number of correct occurrences over the total number of occurrences for each Cf value.

The equations used to calculate the re-weighted confidence values are based on linear trend-lines that are calculated for each of the individual semantic-modality categories. These trend-lines are of the form $y = mx + b$ and are created based on two points (x_1y_1, x_2y_2) derived from the data in figure 5.9. The lower point x_1y_1 is obtained from the set of confidence values from 0.0 to 0.5 inclusive, and the upper point x_2y_2 is obtained from the set of confidence values from 0.6 to 1.0 inclusive. In particular, for FTR_S, the value x_1 is calculated based on the average of the confidence values in the first set, i.e. $x_1 = \frac{0.0+0.1+0.2+0.3+0.4+0.5}{6} = 0.25$, while the value y_1 is obtained based on the sum of the correct occurrences divided by the sum of the total occurrences in this set, which in the case of FTR_S would be $y_1 = \frac{124+32+39+40+48+46}{146+32+39+42+48+46} = 0.9320$. Similarly, the value x_2 is calculated based on the average of the confidence values in the second set, i.e. $x_2 = \frac{0.6+0.7+0.8+0.9+1.0}{5} = 0.80$, while the value y_2 is obtained based on the sum of the correct occurrences divided by the sum of the total occurrences in this second set. Using coordinate geometry, these two points can be used to determine the gradient of a line $m = \frac{y_2-y_1}{x_2-x_1}$ and the y-intersect $b = -mx_1 + y_1$ and thus the equation of each of the trend-lines as shown in table 5.5. The re-weighted confidence values are then generated based on these trend-lines, by substituting each confidence value for x in the equation of the line to determine y (i.e. the re-weighted value). Re-weighted values that fall outside the confidence value range from 0 to 1 during the process of scaling are rounded up/down to conform to the given range.

Feature	Trend-lines	Object	Trend-lines
Speech	$y = 0.0967x + 0.9078$	Speech	$y = 0.0310x + 0.9752$
Handwriting	$y = 0.7108x + 0.3979$	Handwriting	$y = 0.6277x + 0.1861$
Intra-Gesture	$y = 0.9404$	Intra-Gesture	$y = -0.0043x + 1.0017$
		Extra-Gesture	$y = 1.0000$

Table 5.5: Trend-lines for the semantic-modality categories: *FTR_S*, *FTR_H*, *FTR_GI*, *OBJ_S*, *OBJ_H*, *OBJ_GI*, and *OBJ_GE*.

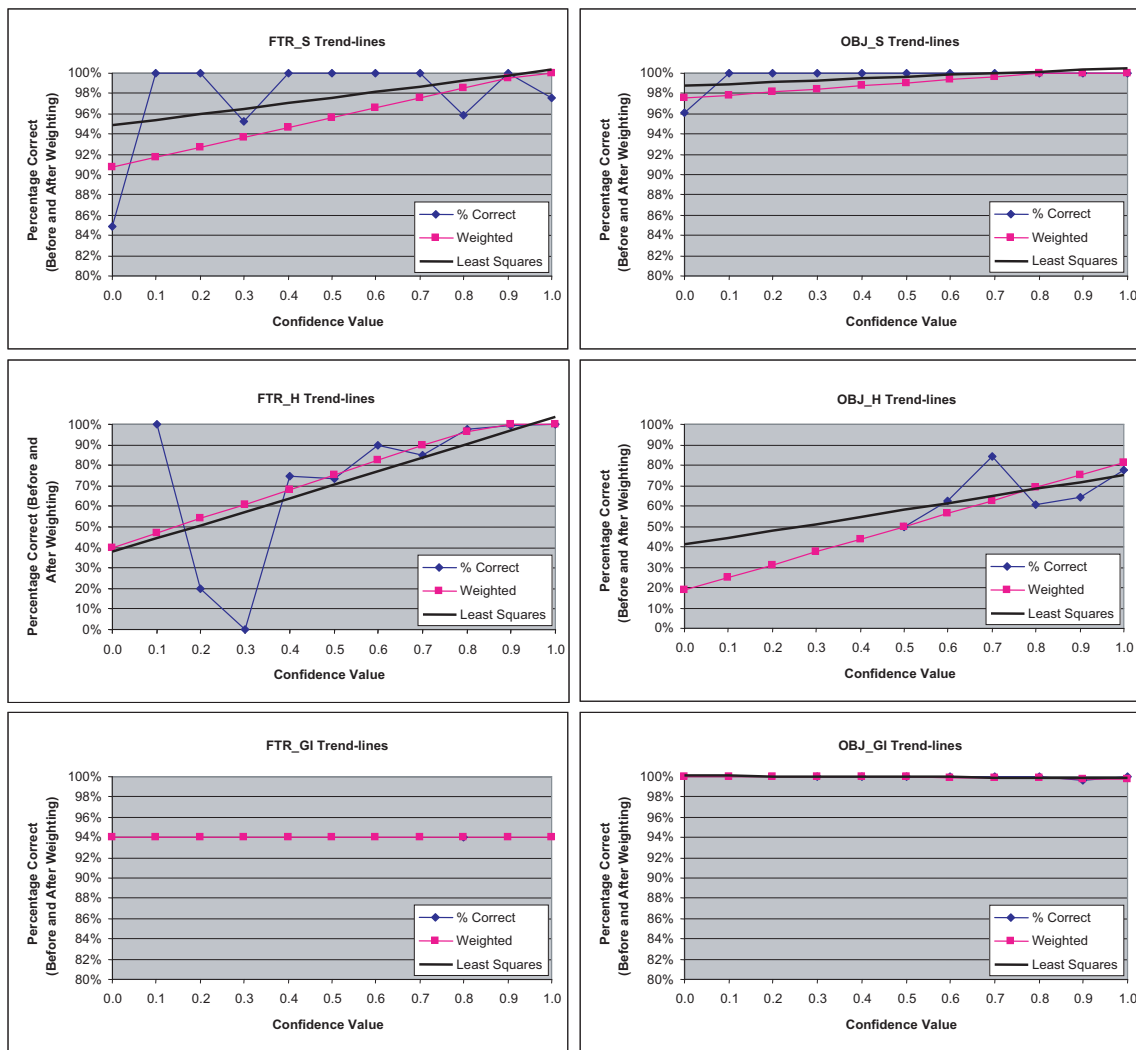


Figure 5.10: Trend-lines for the semantic-modality categories: *FTR_S*, *FTR_H*, *FTR_GI*, *OBJ_S*, *OBJ_H*, and *OBJ_GI*. Each figure plots the line representing the percentage of correct occurrences (% Correct), the weighted values as used in the MSA/BPN (Weighted), and the least squares trend-line (Least Squares).

The tables in figure 5.9 also show that some semantic-modality categories do not have a recorded number of occurrences for all confidence values. Using the terms black-box and glass-box as defined in (Tessitore & Hahn, 2000), it can be said that speech recognition in the MSA/BPN is ‘black-box’ because the methods for generating confidence values are hidden from the programmer, while the handwriting and gesture recognizers are ‘glass-box’ because the generation of confidence values by these components are self-designed. From those recognizers used in the MSA/BPN at the time of the field study, it is known that some confidence values will never be generated by certain recognizers, for example FTR_GI which was at the time set to return only the value Cf=0.8 and OBJ_GE which is set to return only the confidence value of Cf=1.0 (a value that the results show is indicative of the recognizer’s accuracy). For these communication modes, the trend-lines are taken to be a horizontal line that is derived from the single point in which occurrences were recorded in the lookup tables. Other semantic-modality categories like FTR_H, OBJ_H, and OBJ_GI (which is only capable of returning confidence values within the range from 0.25 to 1.0) have occurrence recordings for more than one confidence value but not for all confidence values. The trend-lines for these communication modes are calculated based on the average of only those values with a recorded occurrence, as this avoids divide-by-zero errors. To demonstrate, the lower point for FTR_H can be seen to result in the coordinates $x_1 = \frac{0.1+0.2+0.3+0.4+0.5}{5} = 0.30$ and $y_1 = \frac{1+1+0+9+11}{1+5+3+12+15} = 0.6111$.

A current limitation of this method for re-weighting confidence values is that if given very little data, minor inconsistencies can arise, as can be seen for the instance of OBJ_GI in which the re-weighted values actually decrease as the confidence value increases (from 1.0000 at Cf=0.0 to 0.9974 at Cf=1.0). A solution to this particular case would be to create a rule that ensures lower-valued confidences have either the same or lower re-weighted value as their neighbouring higher-valued confidences. In the longer term however, feedback on recognition accuracy that is collected from the user during normal system use is expected to improve the reliability of the dataset for all confidence values that do not currently have or have only a limited number of recorded occurrences. Section 5.3.3.1 outlines how the MSA/BPN is able to capture such feedback, which may in the future be used for machine learning purposes. The MSA/BPN is currently only able to use the static trend-line equations defined in table 5.5, although it would not require much effort for user feedback to be dynamically incorporated into the lookup tables outlined in figure 5.9.

The method used to re-weight confidence values is considered a more reliable interpretation than simply taking the percentage of correct occurrences per confidence value because the re-weighting approach accepts that neighbouring confidence values are related to one another rather than being entirely distinct from one another. This can be seen in figure 5.10 where the ‘Weighted’ line shows none of the jumps that exist in the ‘% Correct’ line. The use of only two points in creating the trend-lines is considered sufficient for most modalities, although some discrepancy for OBJ_H and FTR_S can be seen when compared to the method of least squares. The method of least squares is however considered computationally-expensive because it is, in the future, planned to automatically compute the re-weighted confidence values in the MSA/BPN each time user feedback is provided to the system. Of particular importance in the approach to re-weighting confidence values is the fact that although different recognizers might have incomparable confidence values due to any number of reasons (see the end of section 4.1.2), the accuracy values for each individual recognizer are comparable provided that such information is generated under the same constraints (e.g. the same dataset and the same environment context).

The use of accuracy values as an independent means of comparison between results returned

by different recognizers still has a number of drawbacks, particularly with regards to system portability. The accuracy values were for example generated based on a predefined set of constraints: the environment was noisy, the user was stationary, the scenario was that of shopping, the grammars entailed only products of type digital camera for which there were 12 feature attributes and only 13 objects, and the interactions were all generated by a single familiar user of the system. The accuracy values are also only representative of the recognizers used during the study, such that new recognizers added to the system or even updates to existing recognizers (e.g. improved acoustic models for the speech recognizer) would influence the accuracy of the results in the current lookup table. Such changes would then require the system to be retrained by either the end user or the program developer.

5.3.3.1 User feedback on Recognition Accuracy

To allow for the continued improvement of accuracy values in the dataset, a method for providing user feedback into the system was implemented. As shown in figure 5.11, it is possible for a user to access the N-best list of values for each semantic element in the last most recently recognized utterance. Based on the example in the figure, had the ‘optical zoom’ query or the ‘PowerShot G3’ object reference been incorrect, a user could have indicated this to the system in addition to indicating what the correct recognition result should have been, either from the associated 3-best list or via the pull-down menu titled ‘Other’. In the approach taken in the MSA/BPN, the N-best lists that the user has access to are in fact the fused results from one or more modality inputs, and as seen in the figure the lists also indicate the modality used, confidence/accuracy information, and the recognized value. In the given example, in which no semantically overlapped information is present, modality information is visible for both the object (denoted by a ‘G’ for gesture) and the query (denoted by a ‘H’ for handwriting). For semantically overlapped input, modality information would no longer be available, as the N-best lists only show the final fused result rather than the lists generated by each individual recognizer.

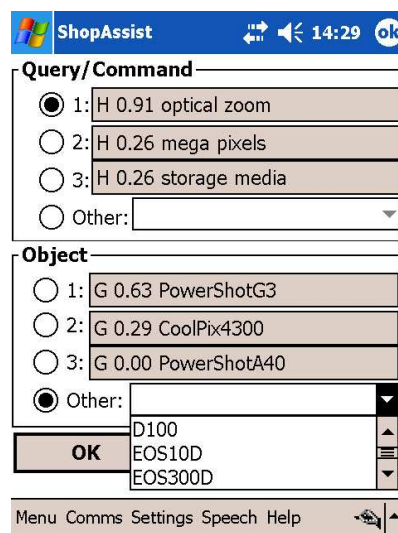


Figure 5.11: N-best list feedback for the semantically non-overlapped handwriting-gesture input: H=“Optical Zoom”, G=“PowerShot G3”.

The foreseeable benefit of this means of capturing recognizer accuracy is that future versions of the MSA/BPN will be able to dynamically adjust to different environment contexts and different users, based on runtime updates to the otherwise static lookup tables. Such adaptation will nonetheless require input from the user in signalling when input is incorrect, and careful consideration will also be needed to determine whether no feedback is sufficient in assuming that a recognition is correct.

Feedback accumulated by the user in this manner would in fact contain more information than was collected in the field study, because a user is able not just to indicate whether a result was correct or incorrect but also what the correct result should have been. The field study data in comparison, only logged correct and incorrect recognitions. This additional information might allow for correction algorithms to be based not only on recognizer or modality accuracy, or semantic type accuracy, but also on semantic value accuracy, for example the percentage of times that the keyword ‘price’ is incorrectly recognized and what it is most often incorrectly recognized for.

5.3.4 Conflict Resolution between Multiple Communication Acts

Communication acts are chosen based on a best-fit principle, in which semantic values on the blackboard are matched to communication act slots in the predefined modality-free language. This predefined language is hard-coded in the system and currently only a limited number of communication act types are implemented in the MSA/BPN shopping scenario.

Communication acts in the MSA/BPN are selected based on the recognized ‘query+feature’ or ‘command’ elements in an interaction. Triggering the modality fusion process is based on the issuing of either of these two elements, thus making them a reliable starting point in assuming that an interaction has taken (or is still taking) place. Table 5.3 in section 5.1 shows the following implemented communication acts:

$\langle Q_{wh-yn} \rangle \langle F \rangle \langle O \rangle$: E.g. “How many megapixels does the PowerShot S50 have?”
 $\langle C_f \rangle \langle O \rangle$: E.g. “Find the PowerShot S50”
 $\langle C_c \rangle \langle O \rangle \langle O \rangle$: E.g. “Compare the PowerShot S50 to the PowerShot S1 IS?”

It can be seen from the above examples that if a ‘query+feature’ is provided by the user, one object-identifying reference is still required for the interaction to be complete. Similarly, if a ‘find’ command is detected, then one object reference is also required, while for a ‘compare’ command two object references would be required for the interaction to be complete. A limitation of this approach is that the selection of a communication act is effectively based on the type of query or command that was recognized, rather than on the type and number of objects that were provided, or on a combination of query/command and object information. As a result, the number of objects located on the blackboard is not used as an indication of the likelihood of a particular communication act having been provided. Although this is sufficient for the current implementation, this approach is less scalable than one that would incorporate both the type of query/command as well as the type and number of object references. A good example illustrating the boundaries of the MSA/BPN approach to communication act resolution is seen by the communication acts $\langle Q_{wh-yn} \rangle \langle F \rangle \langle O \rangle$ and $\langle Q_{wh-yn} \rangle \langle F \rangle \langle O+ \rangle$, where the number of allowable object references is independent to the type of query/command, e.g. $\langle Q_{wh-yn} \rangle \langle F \rangle \langle O \rangle \langle O \rangle$ = “What is the price of the PowerShot S50 and the PowerShot S70?”. The three different types of communication act implemented in the MSA/BPN are however considered sufficient because they demonstrate the different types

of multimodal dialogue phenomena outlined in section 2.3.2, including mutual disambiguation, deixis and crossmodal reference resolution, and anaphora and ellipsis. The implemented acts are additionally sufficient to demonstrate the different types of multimodal interaction relevant to this dissertation as outlined in section 4.2, including temporally and semantically non-overlapped and overlapped input. Finally, a focus point of this dissertation is the resolution of conflicts arising from multiple references addressing the same referent, rather than the resolution of best-fitting communication acts.

On determining the best-fit communication act, the MSA/BPN filters the events on the multimodal blackboard to select the semantic constituents most appropriate for the required communication act slots. It is at this stage that references are unified and conflicts are resolved. Conflicting features (e.g. <S="What is the price <H="name"> of the PowerShot S50?">) and conflicting objects (e.g. <S="What is the price of the PowerShot S50?"><GI="PowerShot S45">) are resolved in the MSA/BPN based on confidence values that can be re-weighted according to each recognizer's own accuracy lookup table, as defined earlier in section 5.3.3.

5.3.5 Multimodal Blackboard Event Filtering

The modality events written to the multimodal blackboard and stored as interaction nodes are filtered based on time constraints, so that only the most salient nodes still exist for modality fusion processing. 'Query+feature' and 'command' events require minimal filtering (and subsequently also only minimal fusion), because these inputs are responsible for triggering the modality fusion component, thus making it rare for many of these events to appear at the same time on the blackboard. Multiple object references on the other hand occur quite frequently and may be separated by timestamps differing by up to several minutes. It is these object references, rather than the feature references, that are focussed on in this dissertation. Choosing the appropriate object references for a communication act is based on the analysis of timeframe and timestamp information and on the analysis of confidence values.

This section concerns itself with the temporal analysis of object references that have been written to the blackboard, while the next section focuses on the use of confidence values to resolve conflicts among multiple semantic elements. The section differentiates itself from section 5.2 in that the timing aspects that are now discussed focus not on the complete user interaction but rather on the duration of individual modal inputs and on the relative temporal location of semantic elements (e.g. before, after). The section assumes that a user interaction has completed successfully, thus populating the blackboard with at least one query/command and one or more objects of TimeType 'past' and/or 'present', and it is also assumed that a particular communication act has been selected based on the provided 'query+feature' or 'command' information.

Given a suitable communication act (e.g. 'find' or 'compare'), the MSA/BPN tries to select the required number of salient objects from the modality events written to the blackboard. The devices used to capture user input are treated independently based on their individual recognizer IDs (e.g. speech recognizer, ID=007), to help distinguish between events generated by same-type recognizers. The objects that are selected are, wherever possible, chosen from the current user interaction (i.e. with a TimeType of 'present'), but if insufficient 'present' object references exist, 'past' objects are also included in the selection.

Once the required number of objects have been identified (e.g. one for a 'find' command and two for a 'compare' command), a timeframe is calculated to further reduce the number of valid references. For semantically non-overlapped input, this timeframe is calculated to include only

the number of most recent object references that are required for the communication act. For semantically overlapped input, this timeframe is calculated to include only the required number of most recent object references generated by a single recognizer. The timeframe is then extended by one extra second to account for the occurrence of other temporally similar events and is finally extended again to cover any events that are only partially covered by the existing timeframe.

It is possible that no conflicts between interaction nodes exist at this stage, in which case the communication act slots are populated with the events and parsing is complete. In many cases, no semantically overlapped information will occur, for example $\langle H = \text{"Price"} \rangle \langle GI = \text{"PowerShot S50"} \rangle$, where each slot in the communication act [$\langle Q_{wh-yn} \rangle \langle F \rangle \langle O \rangle$] is defined once only. When semantically overlapped information is present however, conflict resolution may be necessary before unification can proceed. Conflict resolution between two semantically overlapped elements is a focal point of this dissertation and is discussed in detail in section 5.3.6.

A limitation of the MSA/BPN is that only semantic elements that are overlapped with a maximum of two input streams can be processed, for example speech overlapped with handwriting or gesture, but not speech overlapped with handwriting and gesture. This is reasonable when one considers the effort a user would have to go to in order to enter the same information in three differing modalities, but is a minimalistic approach when one considers that multiple input streams can also be generated from a single user input, for example speech being recognized by a number of different speech recognizers. The following two examples illustrate the type of semantically overlapped references that the MSA/BPN is capable of resolving. The first example illustrates a single semantically overlapped reference, while the second example illustrates multiple semantically overlapped references.

$\langle S = \text{"What is the price of the PowerShot S50?"} \rangle \langle GI = \text{"PowerShot S45"} \rangle \rangle$.
 $\langle S = \text{"Compare the PowerShot S50"} \rangle \langle GI = \text{"PowerShot S50"} \rangle$ to the $\langle S = \text{"PowerShot S70"} \rangle \langle GI = \text{"PowerShot S70"} \rangle \rangle$.

The MSA/BPN is only able to resolve semantically overlapped references in which it is known that the references point to the same referent. This is determined based on the number of references provided by each modality per user-turn, for example in the above examples one speech and one gesture reference is provided in the first example, while two speech and two gesture references are provided in the second example. Two speech references and only one gesture reference would not be resolvable in the MSA/BPN, as it would be unclear just which speech reference the gesture reference refers to without having access to detailed timestamp information that is unavailable in most embedded recognizers. Furthermore, only elements with the TimeType ‘present’ are included in the resolution of semantically overlapped input. Thus, had any of the references in the second example been marked with a TimeType ‘past’ they would have been automatically filtered out, increasing the likelihood that only semantically non-overlapped information remain.

Table 5.6 illustrates how the temporal placement of semantically overlapped references can affect the fusion of information from different modalities. In particular, four examples of the alignment of speech with different gesture references are shown for a ‘compare’ communication act. The MSA/BPN is able to fuse the speech and gesture references provided in examples 1 and 2, based on the temporal order of the speech and gesture references, but the MSA/BPN would discard the gesture references in examples 3 and 4 because their number (i.e. three and one respectively) does not match the total number required by the identified ‘compare’ communication act (i.e. two). These examples show the usefulness in having information on the temporal order of events,

while at the same time they show the limitations of the MSA/BPN implementation that arise from not having precise timing information for individually recognized words in an utterance, which is particularly important for speech, where semantic meaning is based on particular keywords found in the utterance.

Modalities	Timeframe for semantically overlapped Input	
	$t_{start} - 1 \text{ sec}$	Period of Time from t_{start} to t_{finish}
Speech		“Compare the PowerShot S50 to the PowerShot S70.”
+ Gesture	G_1	G_2
+ Gesture		G_1 G_2
+ Gesture		G_1 G_2 G_3
+ Gesture		G_1

Table 5.6: Four examples of speech combined with different gesture inputs illustrating the affect that temporal placement has on semantically overlapped references.

To further demonstrate the usefulness of temporal order, it can be seen that although the exact timestamps for the speech references ‘PowerShot S50’ and ‘PowerShot S70’ are unknown, it is still possible to connect them to the corresponding gesture events in examples 1 and 2 based on the temporal placement of the speech events, i.e. the first speech object reference occurred ‘before’ the second and is therefore associable with the event G_1 . The fusion of the references in examples 3 and 4 would still be far from guaranteed, even if timestamp information were available on a per-word basis, as users do not always temporally overlap semantically overlapped information. Reliably predicting the temporal placement of information in different modalities is still an ongoing field of research (Xiao et al., 2003).

A final interesting aspect of the MSA/BPN implementation is that constraint checks on the semantic type of different objects (e.g. ‘digital camera’, ‘language technology’) are not necessary, as the user is asked to select only a single product type during shelf synchronization, as shown in figure 2.14. This voids the possibility that conflicting object types arise and thus also voids the need to consider the entire range of different partial semantic overlap categories (unidentifiable, type identifiable, and uniquely identifiable), as defined in section 4.2.2.

5.3.6 Conflict Resolution between Multiple Semantic Elements

Semantically overlapped elements might arise due to a number of different reasons, such as a user providing the same information in multiple modalities like SGI, SGE, or SH, or due to two or more of the same-type recognizers being employed (e.g. two speech recognizers, SS, or two handwriting recognizers, HH) to recognize input in a particular modality. It is also foreseeable that overlapped information might arise through a combination of multiple modalities and multiple same-type devices like SSGI in which two speech recognizers return results for the verbal representation of a referent while a gesture recognizer returns a result for the graphical representation of the same referent. The resolution of conflicts among semantically overlapped elements can thus be seen to be an important aspect of a modality fusion component, especially in instrumented environments where it will not be uncommon for many recognition services to be available to nearby mobile devices. This section discusses the resolution of conflicts among semantic elements in the MSA/BPN, starting with an outline of uncertain reasoning and certainty theory and

then illustrating the use of certainty factors in a practical example from the MSA/BPN.

5.3.6.1 Uncertain Reasoning

Wahlster (2003) states that the key function of modality fusion is “the reduction of the overall uncertainty and the mutual disambiguation of the various analysis results”, and Kaiser et al. (2003) states that “multimodal architectures must cope first and foremost with uncertainty”. One of the earliest techniques used to manage uncertain reasoning was probability theory, which is well-founded mathematically. Probability theory does however require a statistical basis not always available in the types of problems occurring in dialogue systems. Certainty theory (Shortliffe & Buchanan, 1975) in comparison, provides a practical alternative for managing inexact reasoning. It relies on ascribing judgemental belief values to uncertain statements, and although lacking a formal foundation, the technique offers a simple approach and produces results that are acceptable in many applications (Durkin, 1994a).

Certainty factors quantify the confidence that an expert might have in a conclusion that he or she has arrived at, and they are used to obtain estimates of the certainty to be associated with conclusions drawn from uncertain rules and uncertain evidence. Certainty factors are quantified linguistically through terms like: certain, fairly certain, likely, unlikely, highly unlikely, and definitely not, and by numeric scales like: 0 to 1 and -1 to $+1$. One of the first systems to use certainty factors was the rule-based expert system MYCIN (Durkin, 1994b; Shortliffe, 1976), which diagnosed infectious diseases and recommended appropriate antibiotics. This system used rules of the form:

```
IF   The infection is primary bacteremia
AND  The site of the culture is one of the sterile sites
AND  The suspected portal of entry is the gastrointestinal tract
THEN There is suggestive evidence (0.7) that the infection is bacteriod.
```

Using the numeric scale from -1 to $+1$, it can be seen that a certainty factor approaching -1 would imply that there is strong evidence against a given hypothesis, while a certainty factor approaching $+1$ would imply that there is strong evidence for a given hypothesis. A certainty factor of 0 would correspond to little evidence either for or against a given hypothesis (i.e. neutral). Certainty Factors (CF) are used in the evaluation of rules containing one or more premise. Rules generally add to the belief or disbelief of a conclusion, and if two rules contribute to the same conclusion, they may be combined based on the following propagation equations:

$$CF_{R1}(C) + CF_{R2}(C) - CF_{R1}(C) * CF_{R2}(C), \quad (5.1)$$

where $CF_{R1}(C)$ and $CF_{R2}(C)$ both represent positive values.

$$CF_{R1}(C) + CF_{R2}(C) + CF_{R1}(C) * CF_{R2}(C), \quad (5.2)$$

where $CF_{R1}(C)$ and $CF_{R2}(C)$ both represent negative values.

$$\frac{CF_{R1}(C) + CF_{R2}(C)}{1 - \min(|CF_{R1}(C)|, |CF_{R2}(C)|)}, \quad (5.3)$$

where $CF_{R1}(C)$ and $CF_{R2}(C)$ are of opposite sign.

In the MSA/BPN, certainty factors are used to combine confidence values from the individual recognizer's N-best lists. This is particularly relevant when semantically overlapped input exists. In the MSA/BPN, only positive values are returned in the N-best lists generated by the different speech, handwriting, and gesture recognizers, such that the lowest value that may be returned by a recognizer corresponds to a neutral evidence (i.e. 0) and the highest value that may be returned corresponds to strong evidence for a given hypothesis (i.e. +1). Because only positive values are returned by the recognizers, Equation 5.1 is most relevant to the MSA/BPN. This equation, $f(x,y)=x+y-xy$, is also shown as a three-dimensional graph in figure 5.12.

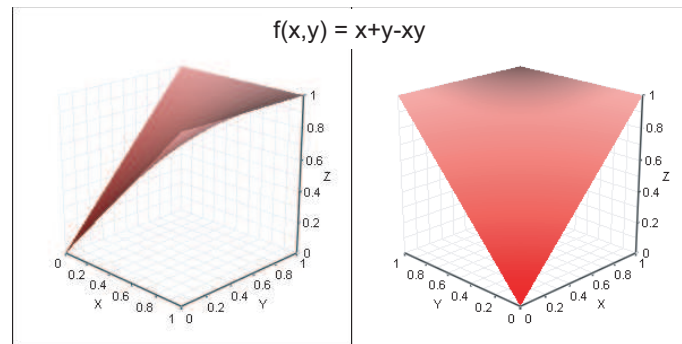


Figure 5.12: Graph of the certainty factor's equation as used in the MSA/BPN: $f(x,y)=x+y-xy$.

Some alternatives to using certainty factors would have been to use an approach based on Bayesian decision theory (Pearl, 1988), the Dempster-Shafer theory of evidence (Shafer & Pearl, 1990), or Markov decision theory (Puterman, 1994).

Decision theory represents beliefs about the world as probabilities and provides a useful formal framework for modelling problems of inference and decision. Bayesian networks, or belief networks, build on the concept of decision theory by mapping out cause-and-effect relationships among key variables and by encoding these variables with values that represent the extent to which one value is likely to affect another (Pearl, 1988). Bayesian networks are a fundamental tool used in artificial intelligence and can generate optimal predictions or decisions, even when key pieces of information are missing. The underlying data in a Bayesian model is generally constructed based on a variety of different information sources, including past data accumulated on the system or related systems, the judgement of subject matter experts, and the judgement of experienced model builders. Although Bayesian decision theory could have been implemented in the MSA/BPN using recently developed technologies for mobile devices (Brandherm & Jameson, 2004), it was decided to keep the implementation computationally simple and to instead use certainty factors for conflict resolution between semantically overlapped referents.

The Dempster-Shafer theory (Shafer & Pearl, 1990; Bauer, 1996) is a mathematical theory of evidence based on belief functions and plausible reasoning. In comparison to Bayesian theory, where probabilities are assessed directly for the answer to a question of interest, Dempster-Shafer theory assesses probabilities for related questions and then considers the implications of these probabilities for the question of interest (Shafer & Pearl, 1990). A Dempster-Shafer approach has the advantage (over Bayesian methods) that a-priori probabilities need not be specified, however the formulation of the decision process can still become very complex.

Markov decision processes constitute the mathematical framework for decision-theoretic planning (Puterman, 1994; Bohnenberger, 2005) and are commonly used to solve planning problems

based on a transition model. Markov decision processes are defined by the tuple $\langle S, A, T, R \rangle$, in which 'S' represents a set of states, 'A' a set of actions that can be taken from each state, and 'T' and 'R' transition and reward functions respectively. The uncertain outcomes of an action are represented by non-deterministic state transitions and policies are used to dictate which action to take from each state so as to maximize a reward.

Although the use of Bayesian networks, Dempster-Shafer evidences, and Markov decision processes might lead to more accurate calculations than the implemented certainty factors, exact calculations for mobile devices are not possible due to processor and memory limitations. Approximations for these processes do exist for mobile devices, but these are then only approximations. It is for this reason that certainty factors can be seen to be similarly useful in the given purpose of resolving uncertainty between semantically overlapped referents.

5.3.6.2 Walkthrough of the Evaluation of semantically overlapped and Conflicting Input

The following example illustrates how conflicting input between two semantically overlapped object references can be resolved in the MSA/BPN. Assume a user issues the combined spoken-handwriting dialogue by speaking "What is the price of the PowerShot S45?" while simultaneously scribbling the word 'PowerShot S45' on the display of the PDA, as indicated in figure 5.13A. The N-best lists in figure 5.13B show that the object information was misinterpreted by both the speech recognizer and the handwriting recognizer. Whereas the speech recognizer returned similar sounding results, i.e. 'PowerShot G5', 'PowerShot S45', and 'PowerShot A75', the handwriting recognizer returned results with a similar length and containing similar characters, i.e. 'PowerShot S50', 'PowerShot S45', and 'PowerShot G3'. In this case, the object references are semantically overlapped and conflicting because the best returned spoken utterance segment refers to the 'PowerShot G5' while the best returned handwriting result refers to the 'PowerShot S50'.

The values used in this example bear all the hallmarks of a typical real-world scenario in that the embedded speech and handwriting recognizers used in the MSA/BPN quite often return confidence values of 0.0 and 0.9 respectively (see figure 4.12 and table 5.9). Through the use of certainty factors the correct result is however still determinable and this is despite the desired referent (i.e. 'PowerShot S45') being neither the best returned speech nor the best returned handwriting result. Interesting to note is that if the N-best list of non-weighted confidence values had been used instead of the re-weighted confidence values, the final result would have been incorrectly selected as the 'PowerShot S50', regardless of whether or not the principle of certainty factors were applied. The simple multiplication of joint probabilities as applied in (Kaiser et al., 2003) would also fail this test. Furthermore, had only the result with the best re-weighted confidence value been selected, without regard for certainty factors, the result would still have been incorrectly selected as the 'PowerShot G5'. This shows the importance in re-weighting recognizer confidence values and also the benefit in using certainty factors during the resolution of semantic conflicts.

Describing in detail the operations performed to derive the values shown in figure 5.13C, it can be seen that the certainty factors equation for positive values, $f(x,y)=x+y-xy$, is used to generate a combined list of N-best results. In this equation, 'x' represents each of the speech objects and 'y' represents each of the handwriting objects. For the two sets of 3-best lists that are provided, a total of 18 different combinations exist, from which the best three combinations are outlined in the table in figure 5.13C. The 18 combinations arise in that each of the 3-best list of speech values is compared with each of the 3-best list of handwriting values and vice-versa, so that $S_{N=1}$ (and later

$S_{N=2}$ and $S_{N=3}$) is compared with $H_{N=1}$, $H_{N=2}$, and $H_{N=3}$, and then $H_{N=1}$ (and later $H_{N=2}$ and $H_{N=3}$) is compared with $S_{N=1}$, $S_{N=2}$, and $S_{N=3}$. As explained earlier, when multiple rules all contribute to the same belief (e.g. $S_{N=2}$: PowerShot S45 and $H_{N=2}$: PowerShot S45 in figure 5.13B), the resulting value is increased by the use of certainty factors (i.e. $N=1$: PowerShot S45 in figure 5.13C). The reason why 18 combinations exist rather than only nine is because the resulting confidence values from SH and HS are not equivalent, for example $S_{N=1}H_{N=1}$ (i.e. $0.9752+0-0=0.9752$) is not the same as $H_{N=1}S_{N=1}$ (i.e. $0.7511+0-0=0.7511$). Although only the resolution of two overlapped elements currently occurs in the MSA/BPN, this could easily be increased to three or more overlapped elements by using the certainty factors equation to combine the first two elements and then recursively using the equation to combine this result with each additional element.

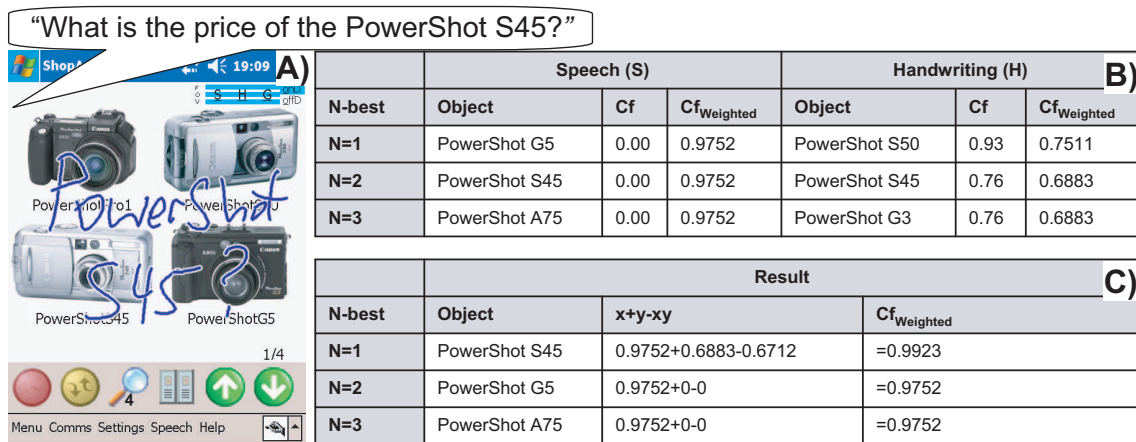


Figure 5.13: Conflict resolution between semantically overlapped object references, showing A) the user input, B) the two sets of N-best lists, and C) the resulting list of N-best values and their associated confidence values.

This chapter describes the results of two empirical usability studies that were conducted on the Mobile ShopAssist (MSA) demonstrator. The primary goals of the studies were to measure user preference for a wide range of modality input combinations, first within a private laboratory setting and then within a public real-world setting. In comparison to most other studies conducted on multimodal interaction (see section 6.1), these studies were designed to closely model a mobile scenario, i.e. that of shopping for products located on the shelves of a store. During the studies, a total of 23 different modality combinations were tested, and these were derived from the three elementary communication modes speech, handwriting, and gesture. The interaction combinations ranged from unimodal to multimodal and from non-overlapped to overlapped input (e.g. where speech and handwriting are used to provide duplicate information). In addition to modality combination preference, the studies also provide insight into aspects such as the intuitiveness of the individual modality combinations, the effect that being in a public or private environment has on modality usage, gender differences that exist between modality usage, and how accepting users are of conversing with anthropomorphized objects.

Before the results from the usability studies are described in section 6.2, some of the more prominent usability studies that have been conducted on the topic of multimodal interaction in the past are outlined.

6.1 Previous Usability Studies on Multimodal Interaction

Previous studies on multimodal interaction have focused on a range of aspects such as efficiency gains and recognition accuracy (Oviatt, 2002). These studies are however often conducted in stationary settings and are based on desktop computing hardware and Wizard of Oz (WoZ) mock-ups and simulations. Only the most recent of studies are beginning to focus on multimodal interaction in natural environments, where the user is mobile and subject to adverse or changing environment conditions like privacy issues and background noise (Kumar et al., 2004; Wasinger et al., 2005; Wasinger & Krüger, 2005, 2006; Wasinger & Wahlster, 2006). Other lines of work have focused on creating guidelines for the advantages and disadvantages of different modalities (e.g. Sun's Java Speech API guidelines¹), however although such reports outline factors that can affect modality usage (e.g. background noise can have a negative effect on speech), they only target individual modalities in isolation, rather than the combination of multiple modalities. This section takes a

¹Java Speech API, <http://java.sun.com/products/java-media/speech/>

brief look at some of the more prominent usability studies conducted in the past on multimodal interaction and in particular the work of Sharon Oviatt and her colleagues from the Center for Human Computer Communication at the Oregon Graduate Institute of Science and Technology. These studies are considered representative of the times, although this is not to say that they are the only studies on multimodal interaction. For example, during the SmartKom project, a number of subjective and objective usability studies were conducted on three different technical scenarios (SmartKom Public, Mobile, and Home) and a variety of different task domains (Schiel, 2006). The Institute of Phonetics and Speech Communication at the Ludwig-Maximilians-University in Munich also generated an annotated corpora for usability purposes during the course of the SmartKom project. This corpora consists of multimodal recordings from 224 persons, the results of which are a set of 145 DVDs covering 146 recorded sessions from 73 subjects in a mobile tourism scenario (see <http://www.bas.uni-muenchen.de/Bas/BasSmartKomMobileng.html> and also Schiel and Türk (2006)).

In (Oviatt & Olsen, 1994), 44 native English white-collar professionals (excluding computer scientists) were analysed to determine if people select to ‘write’ or ‘speak’ when performing verbal/temporal tasks (e.g. conference registration) and computational/numeric tasks (e.g. personal banking). The communication modalities included speech-only, pen-only, and combined pen/voice. On average, content was written around 15% of the time and spoken around 85% of the time. Digits and proper names were shown to have a higher likelihood of being written than other textual content, and 57% of all pen-voice combined interaction was used for the following contrastive functionalities: ‘original input and correction’, ‘data and command’, and ‘digits and text’. For example, subjects who corrected an error were significantly more likely than chance to use contrastive modalities to distinguish corrections from the original input. Subjects were also significantly more likely to use modalities contrastively to distinguish ‘data’ (i.e. digits and computational signs) from ‘commands’, and written data and spoken text had a 73% likelihood versus a 27% likelihood for spoken data and written commands.

In (Oviatt, Cohen, & Wang, 1994), 18 subjects were studied while interacting with the same system described above, to examine how input modality (speech, writing, combined pen/voice) and presentation format (structured, unconstrained) influence linguistic complexity. Highly structured form-based interfaces were seen to reduce the number of words, length of utterances, syntactic ambiguity, perplexity, and errors. The modality writing was also seen to reduce wordiness and utterance length, and led to fewer noun phrases (e.g. “James Green from the NOA”) and full sentences (e.g. “Pickup car and drop off at Oakland airport”), consisting instead primarily of ‘value’ (e.g. “Oakland”) or ‘attribute+value’ input (e.g. “Name John Smith”). Perhaps most significant was that 58% of unconstrained speech, 91% of constrained speech, 73.5% of unconstrained writing, and 100% of constrained writing consisted only of ‘value’ information.

In (Oviatt, 1996, 1997), the temporal and numeric tasks for the service transaction system defined in (Oviatt & Olsen, 1994) were complemented by a visual-spatial task, in which subjects were required to select real-estate property from a map (presented in a minimally structured and a highly structured format) using different modalities (speech, pen, and combined pen/voice). The goals of the study were to examine modality preference and performance issues (like accuracy and efficiency) when interacting multimodally with interactive maps. In contrast to the 56% of subjects who preferred to interact multimodally in the verbal/temporal domain, and the 67-89% (minimally structured, highly structured respectively) who preferred to interact multimodally in the computational/numeric domain, 95-100% of subjects preferred to interact multimodally in the visual-spatial domain. Also, in contrast to the rise in preference for multimodal interaction, speech-only

interaction decreased from 39-42% in the verbal domain (minimally structured, highly structured), to 11-33% in the computational domain, and to 0% in the visual-spatial domain. Results further showed that 36% of all content errors and 50% of spoken disfluencies could be eliminated by permitting users to interact multimodally, and that task completion times were 10% faster during multimodal interaction (taking an average of 249 seconds) than during speech-only (278 seconds) and written-only interaction (410 seconds). These efficiency gains were particularly evident for location descriptions, where for example drawing a circle to zoom in on a house was matched by the spoken utterance "Show me the house at the southwest corner of Nevada and Broad Streets".

In (Oviatt & VanGent, 1996), 20 native English speakers were studied while interacting with the service transaction system defined in (Oviatt & Olsen, 1994) to examine how users strategically adapt their use of input mode while resolving system recognition errors in a multimodal interface. For this purpose, 24 simulated errors were collected from each of the subjects during tasks which contained error rates ranging from low (6.5% of input slots) to high (20% of input slots). Results show that during normal interaction, speech accounted for 81.5% of all words and writing accounted for only 18.5% of all words. During error resolution, spoken language dropped to 70% of all words and written language increased to 30% of all words. Results also show that during non-error interactions, the base rate of spontaneously shifting modalities from speech to writing (or vice-versa) was 4.8 per 100 words and increased to 15.7 per 100 words during error resolution. For spiral-errors (with a spiral depth of 1 to 6 repeats to resolve an error), mode switching on the first repetition was shown to be significantly lower (14%) than switching on repetitions 2 to 6. Similar to above, 0.5% of all words were simultaneously spoken and written during non-error interactions, and this increased only to 0.7% of all words during error resolution, thus implying that people do not use simultaneous redundant spoken and written input as a technique for emphasis or clarification during error resolution.

In (Oviatt et al., 1997), 18 native English speakers were studied while interacting with the real-estate property task described in (Oviatt, 1996). The goal of the study was to determine whether speech and writing are used to convey specific semantic constituents (i.e. Subject (S), Verb (V), Object (O), and Locatives (LOC)), and in the expected canonical word order (for English: S-V-O and S-V-O-LOC). A second goal was to determine which of the following command types were most likely to be expressed multimodally: spatial commands (e.g. add, move, or modify objects on the map), selection commands (e.g. zooming in on an object), and general action commands (e.g. print-screen). 96% of unimodal spoken utterances reserved LOCs for sentence-final position and 98% also conformed to the standard S-V-O order for typical English. In comparison, 95% of multimodal constructs began with drawn graphics that conveyed LOC information (i.e. LOC-S-V-O), but with the exception of LOCs, 97% also conformed to the standard S-V-O order for typical English. Also of relevance is that within multimodal constructs, pen input was used 100% of the time to convey LOC information, while speech was used 100% of the time for S and V constituents, and O constituents were spoken 85% of the time and provided via pen 15% of the time. Spatial location commands were shown to account for 86% of multimodal utterances. Selection commands accounted for 11% and general task commands for only 3% of multimodal utterances, whereby it can be noted that interactions in which an object was selected in one interaction and then referred to anaphorically in subsequent interactions was no longer classified as a selection command for the subsequent interactions. Also of interest is that 41% of multimodal constructs contained a spoken deictic and 96% of these involved the terms: 'here', 'there', 'this', and 'that'.

In (Oviatt & Kuhn, 1998), linguistic constructions from 18 subjects performing a similar task to that described in (Oviatt, 1996) were analysed to determine the proportion of terms in multi-

modal language that represent referring expressions such as definite and indefinite noun phrase references, deictic references, co-reference/anaphora, and linguistic indirection/ellipsis. Results show that referring expressions are significantly more common during speech-only interaction than with multimodal interaction and pen-only interaction. Co-references/anaphora made up 63% of spoken referring expressions, which was significantly more than the 53% of multimodal referring expressions and the 54% of pen-only referring expressions. Definite noun phrases (e.g. the post office) and indefinite noun phrases (e.g. a hospital, hospital) accounted for 60% of spoken referring expressions, which was also significantly more than the 45% of multimodal and 45% of pen-only referring expressions. Deictic references in contrast accounted for 11% of multimodal input, which was significantly more than the 0% of spoken and pen-only input. 11% of speech-only constructions also contained linguistic indirection (e.g. "I'd like a house [on the map] next to the museum"), which was also significantly more than the 7% of multimodal constructions and 2% of pen-only constructions in which linguistic indirection occurred. In summary, these results show that although spoken language contains significantly higher levels of co-reference, definite/indefinite reference, and linguistic indirection when compared to multimodal and pen-only interaction, multimodal language contains significantly higher levels of deictic reference compared to spoken and pen-only language.

In (Oviatt, 1999), 8 native and 8 accented English speakers were analysed while interacting multimodally with the QuickSet system (Cohen et al., 1997) to examine how often mutual disambiguation might aid in the resolution of spoken and pen input, both at the signal processing level (based on rankings in the individual speech and gesture N-best lists) and at the parse level after natural language processing had occurred (based on rankings in the multimodal N-best list). Mutual disambiguation is defined to be the "recovery from unimodal recognition errors within a multimodal architecture" (see section 2.2.4), and the author makes note that although QuickSet can process both unimodal and multimodal input, subjects were requested to deliver only multimodal commands to increase the number of interactions consisting of redundant information that might benefit from mutual disambiguation. Results showed that although speech recognition as a stand-alone performed poorly for accented speakers, the multimodal recognition rates did not differ significantly from those of native speakers, implying that an alternate mode can act as a stabilizer in promoting overall mutual disambiguation. For native speakers, signal-level mutual disambiguation accounted for 8.5% of multimodal utterances being corrected, and parse-level mutual disambiguation accounted for 25.5%. This was much higher for non-native speakers where signal-level mutual disambiguation accounted for 15% of multimodal utterances being corrected, and parse-level mutual disambiguation accounted for 31.7%.

The above study was then complemented in (Oviatt, 2000c, 2000b) where rather than examining the benefits of mutual disambiguation for native and non-native English speakers, the benefits of mutual disambiguation for mobile and stationary system use was examined. The study was conducted on 16 subjects who each performed a stationary and a mobile task (i.e. community fire and flood simulations) based on the QuickSet system. The stationary task was conducted with subjects sitting in a quiet room that had an average noise level of 42dB, while the mobile task was conducted with subjects that were walking around a cafeteria that had an average noise level of 49dB (the noise levels ranged from 40-60dB). User parameters such as cognitive load were not considered and the hardware used was a Fujitsu Stylistic 1200 PC, which is a tablet PC with a much larger physical form factor than a PDA and also possessing technological power not available on current state-of-the-art PDAs with regards to hard disk, graphics, memory, and CPU (see figure 6.1). The results show that mutual disambiguation can in a similar way to native and

non-native speakers reduce the performance difference between stationary/quiet and mobile/noisy environments, and this is further amplified by the quality of the microphone used. In particular, for stationary users with a high-end noise-cancelling microphone (and then with a low-end in-built microphone), signal-level mutual disambiguation accounted for 7.5% (11.4%) of multimodal utterances being correct and parse-level mutual disambiguation accounted for 11.9% (15.4%), while for mobile users, signal-level mutual disambiguation accounted for a larger 11% (21.5%) of multimodal utterances being corrected and parse-level mutual disambiguation accounted for a larger 16% (23.3%).

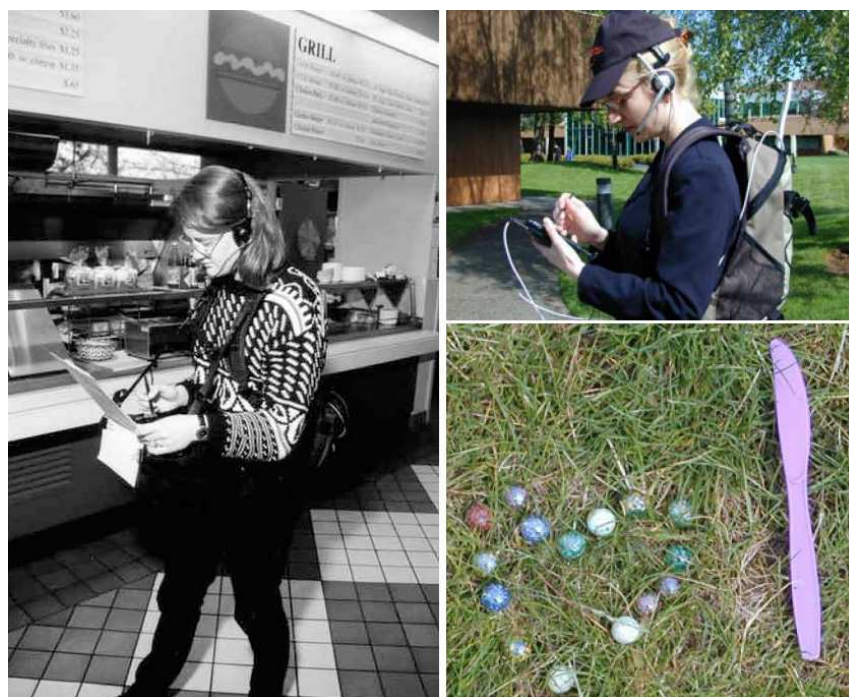


Figure 6.1: Usability demonstrators most similar to the MSA, left (Oviatt, 2000b) and right (Kumar et al., 2004).

In (Oviatt et al., 2003), the interactions from 12 native English speaking adults were examined to identify how malleable (or resistant to change) a subject's integration pattern is, particularly when a high error rate (40%) is applied to help force a switch in integration patterns. The integration patterns in this study are defined as either simultaneous (i.e. where multiple modalities are temporally overlapped) or sequential (i.e. where multiple modalities are separated by time lags). Novel in this study is that the analysis is compared to the principles of Gestalt theory and a behavioural/structuralist perspective. In particular, it is stated that Gestalt theory would predict users to fortify or entrench their existing integration pattern, while a behavioural/structuralist perspective would expect users to switch their integration pattern. The results show that the Gestalt theory was correct in that a subjects' dominant multimodal integration pattern was resistant to change, even when strong selective reinforcement through errors was delivered to encourage switching from a sequential to a simultaneous integration pattern or vice-versa.

In (Xiao et al., 2003), 15 healthy senior subjects (aged between 66 and 86 years old) had their multimodal integration patterns compared to the results from 24 children (aged between 7 and 10)

(Xiao et al., 2002) and 18 adults (Oviatt et al., 1997), during interaction with a map-based task. The goal was to compare the integration patterns of different user groups and to determine how predictable user integration patterns are. The results found that children and seniors tend to be simultaneous integrators (77% and 80% respectively), while adults tend to be sequential integrators (66%), although the value for adults was contradicted in a later study dealing with the effects of varying task difficulty on unimodal and multimodal interaction (Oviatt, Coulston, & Lunsford, 2004) in which 85.6% of interactions were delivered simultaneously rather than sequentially by a subject pool of 10 adult participants. Results also showed that all children and adults demonstrate a dominant integration pattern which was predictable for 92% of children (and 100% of adults) based on their very first multimodal construction. In comparison, only 87% of seniors demonstrated a dominant integration pattern, of which 85% were then predictable based on their very first multimodal construction.

As described in section 3.1.3, Kumar et al. (2004) evaluates the performance of a multimodal interface under exerted conditions in which 14 male subjects were subjected to high exertion (see figure 6.1). The goal of this study was to analyse the relationship of speech, pen-based gesture, and multimodal recognition as a function of user exertion (stationary, running, and running with a load of around 3kg), and to also examine the rate of mutual disambiguation at these different exertion levels. Interaction in this study provided a solid foundation for the collection of mutual disambiguation data in that subject's were required to describe unique objects via two differing modalities (speech and pen). Results from the study show that the recognition rates for multimodal interaction (with a success rate of around 81.5%) and speech interaction (around 77%) remained constant over the different exertion levels, while the recognition rate for gesture interaction decreased significantly from the stationary to the first running exertion state. The rate of mutual disambiguation during multimodal interaction increased over the exertion levels, from 8% in the stationary state to 14% and 17% in the two running states, thus showing similar to the previous studies that mutual disambiguation is capable of compensating for the degradation in accuracy of the individual recognizers.

6.2 Modality Preference in Private (Laboratory) and Public (Real-world) Environments

Designed with shopping scenarios in mind, the MSA assists a user in retrieving product information and product comparison information while shopping in an RFID technology enabled store. The system easily caters for all sorts of products ranging from grocery items such as green eggs and ham, to electronics equipment such as PDAs and digital cameras. Due to the current spate of digital camera models in recent years and the associated popularity of such devices, 'digital cameras' are the primary context in the usability studies that are discussed below.

The central theme that the demonstrator conveys is that of mobile and multimodal interaction. The system is classified as mobile because the user is free to walk around a store while interacting with data containers and objects of his or her choice (e.g. shelves and digital cameras respectively). However, in order to make the studies manageable, subjects remained in front of a single shelf for the purposes of these usability studies. The input modalities that a subject could exploit when communicating with the system included speech, handwriting, intra-gesture, and extra-gesture. Multimodal user interaction was mapped to the modality-free language: [`<Query>`]`<Feature>``<Object>`+

The grammars employed throughout the usability studies encompassed a product set of digital cameras containing 13 objects each with 12 feature attributes. ‘Objects’ refer to products such as ‘PowerShot S60’ and ‘EOS 300D’, while ‘features’ refer to product attributes such as ‘price’, ‘megapixels’, and ‘optical zoom’. To demonstrate, a typical user input was as follows: <S=“How many megapixels does this camera <G=“PowerShot S50”> have?”>.

Section 6.2.1 describes two usability studies that were conducted in a private laboratory environment (Wasinger et al., 2005) and in a public real-world environment (Wasinger & Krüger, 2006), while section 6.2.2 describes the quantitative results obtained from these two studies. Section 6.2.3 then highlights the significant differences between the results of these two studies, and section 6.2.4 discusses the results from a field study on direct interaction with anthropomorphized objects (Wasinger & Wahlster, 2006). Section 6.2.5 summarizes the qualitative results that subjects provided regarding interaction with the communication modes speech, handwriting, intra-gesture, and extra-gesture (Wasinger & Krüger, 2005).

6.2.1 Usability Study Descriptions

6.2.1.1 Study 1 - Private (Laboratory) Tests

The first usability study (Wasinger et al., 2005) was conducted at the University of Saarland² in one of the department’s computer terminal rooms. This laboratory setting differs from a real-world environment in two ways. Firstly, there were few if any other people in the terminal room during the times that testing was conducted, and secondly, background noises were kept to a minimum. The study was conducted on a total of 14 subjects who were either only slightly familiar or completely unfamiliar with the system. The study was conducted in English with subjects that could speak fluent English. 10 of the subjects were students and lecturers from the computer science department aged between 25 and 37 years, while the remaining 4 subjects were not from the computer science department and were completely unfamiliar with the system. Over the testing period for this first study, a total of 440 user interactions were logged by the system, averaging 31 interactions per subject.

The MSA demonstrator allowed subjects to mix-and-match modality input combinations when creating their feature-object dialogue inputs and also allowed them to overlap modalities when communicating with the system. A total of 23 individual modality combinations were tested, 12 of which were non-overlapped, while the remaining 11 were overlapped. Accuracy was not a focus of the study and subjects were told to ignore erroneous system output as the study was only concerned with their input into the system. The complete range of modality combinations that were available in the system (both those that were implemented and not implemented) can be seen in figure 6.4.

Task: The usability study was conducted as a within-subject design, meaning that all of the subjects performed the exact same usability test. Each test session generally required between 45 and 60 minutes to complete, of which around 10 minutes were used to brief the subjects on the underlying system. Each subject was told that the system was to provide support in retrieving product information on digital cameras, and that the study would focus on the input modalities that they used while communicating with the system. The base modalities - speech, handwriting,

²University of Saarland, <http://www.uni-saarland.de>

intra-gesture, and extra-gesture - were explained to the subjects and this was followed with an explanation of the modality combinations that could be used when building feature-object dialogue interactions, for example speech for the feature and extra-gesture for the object. After each interaction, the subjects were asked to rate the modality combination by answering the question "Would you use this modality combination?". The rating scale used was a set of preferences that were later mapped onto a scale from 0.0 to 3.0, in which: '0=prefer not', '1=maybe not', '2=maybe yes', and '3=prefer yes'. Subjects were told that the order of the input was irrelevant, i.e. feature then object, or object then feature, and that system errors were to be expected but should not bias their answer as not all modality combinations had been implemented. The task given to the subjects was to find a camera that suited them best. They were told that this task was only minimally important and that the focus of the study rested on how they communicated with the system. After explaining the different non-overlapped and overlapped modality input combinations, the base modalities were once again repeated, subjects were asked if they had any questions, and the usability study then began.

Procedure: The usability study was divided into two parts, the first being a 'practical component' in which subjects were observed while they interacted with the system, and the second being a 'written component' in which subjects were required to fill in a questionnaire. Subjects were given a PDA device and a headset connected to the PDA's audio jack so that they could speak and listen to the output. They were asked to stand in front of an instrumented shelf containing real physical camera boxes. Also situated on the shelf was a printed list of available feature keywords similar to that shown on the PDA's display.

The first part of the observation was to allow each subject to freely choose their own modality combinations and to rate them while interacting with the system. Most subjects managed around 4 to 5 different modality combinations within this part before needing to be reminded of the remaining modality combinations. At this point, subjects were specifically told the order in which they should use the remaining modality combinations, first those that had been implemented in the system, and then those that had not been implemented in the system. For the non-implemented modality combinations, audio output was turned off and the subjects were told to focus only on their input rather than the system output.

Following the practical component, the subjects were asked to complete a written questionnaire that again asked them to repeat their preference for each individual modality combination and to also state whether or not they thought the modality combinations were intuitive. For this written component, the instructor guided the subject by demonstrating each modality combination as they filled in the questionnaire. Several other questions relevant to mobile and multimodal interaction were also asked, and the survey ended with the subject stating their favourite input modality combination.

The involvement of the subjects (and thus the quality of the results) was considered high, firstly due to each subject being tested individually by the instructor and secondly because the tester was responsible for carefully guiding the subjects through the usability study, encouraging them to explain their decisions when it was thought that they had not fully considered the question.

6.2.1.2 Study 2 - Public (Real-world) Tests

The primary goal of this second study was to test modality preference and modality intuition in a public environment (Wasinger & Krüger, 2006) and to identify how accepting people are of

conversing with anthropomorphized products such as digital cameras, within a shopping context (Wasinger & Wahlster, 2006). The study complements the first study but also extends on it in several key aspects. Firstly, the study was conducted in a real-world setting rather than in a laboratory, and secondly, it encompassed nearly twice as many subjects. In addition, the study incorporated new themes, in particular that of anthropomorphized objects. The description of this study and the accompanying testing procedure is limited to include only the differences that exist between the previously described laboratory study and this real-world study.

The real-world usability study took place inside an electronics store of the 'Conrad' Chain. Conrad Electronic³ is an electronics store that sells a range of more than 50,000 technology and electronics products. Their marketing catalogue encompasses the following range of products: computer and office, multimedia, telephone and radio, sound and lighting, batteries and accessories, home systems, tools and soldering, car Hi-fi and accessories, electronics and metering, modelling, and materials. One user group that this store targets is the home do-it-yourself hobbyist. This mindset of both the shoppers and the Conrad team was a primary reason why the experiment setup was so well accepted by Conrad. In fact, the staff were so accommodating that the shelf got placed in a central part of the store, as shown in figure 6.2.



Figure 6.2: MSA demonstrator installation situated at Conrad Electronic in Saarbrücken, Germany.

The experiment setup consisted of a single instrumented shelf that was situated at the intersection of two aisles combining the telephone and answering machine section with that of the digital camera and computer sections. The aisle dividers used at Conrad are a little below shoulder height, which permits customers to see a large area of the store from any given location.

In comparison to the first study, subjects in this second study were subjected to a real-world environment setting, which contained the following forms of background noise and disturbances:

- **Loudspeaker:** A loudspeaker was positioned above the location of the shelf used in the

³Conrad Electronic, Trierer Straße 16-20, 66111 Saarbrücken, <http://www.conrad.de>

usability study. This speaker was used for calling cashiers to the shop front and played music when not otherwise in use.

- **Goods trolleys and shelf packing:** Goods trolleys were wheeled past the shelf from time to time. These were used for the restocking of products, which due to the lead up of Christmas sales preparation occurred more often than normal.
- **Bystanders:** The test coordinator was instructed to note the number of people that could be seen from the shelf's location at the start of each test session. An average of 13.8 people were recorded for these test sessions. The people included those that were directly watching the experiment, as well as those that were looking at products in the store, and others that were walking among the aisles. The bystanders that were directly watching the experiment were among the minority and generally displayed discretion.
- **Sales assistants:** Each section in the store had its own specialist sales assistants to help people in choosing a product. The usability study took place in a section of the store specializing in answering machines.

The testing was conducted over a two week period in November, during the pre-Christmas sales. During this period, a total of 1489 interactions were logged from 28 different subjects, averaging 55 interactions per subject. Each test session generally took between 45 and 60 minutes to complete.

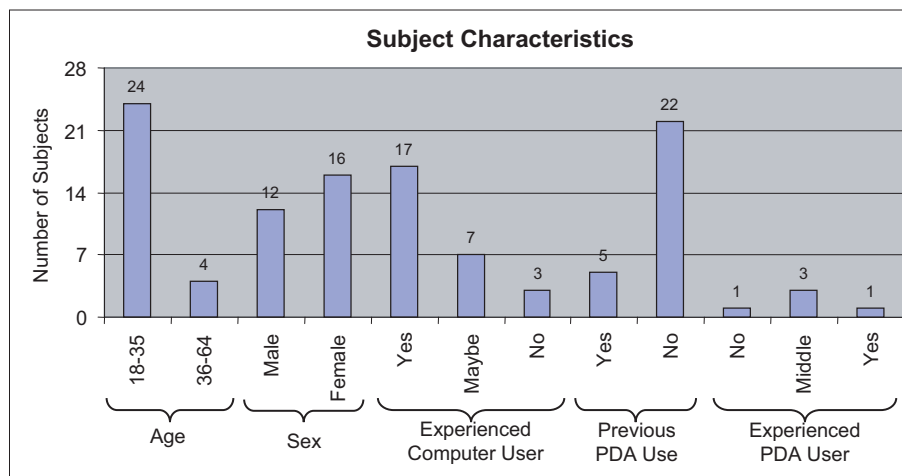


Figure 6.3: Summary of the demographics for subjects that partook in the real-world usability study.

The sample of test persons consisted of 28 people, 16 female and 12 male, and ranging in age from 19 to 55 (mean: 28.3 years). 24 subjects were in the age group 18 to 35, while another 4 were in the age group 36 to 64. The results from one female subject within the 36 to 64 age bracket were discounted, bringing the sample down to 27 subjects (15 female, 12 male). This discounting was attributed to a lack of training time provided on how to use the speech recognizer and in issues concerning the coordination of overlapped input segments. 26 subjects conducted the usability study in German, while 2 did the study in English. 21 subjects were students, and in comparison to the first study, only one was a computer science student. Two people from the first

study also took part in the second study conducted two months later, to provide a small insight into the differences between laboratory and real-world testing. Aside from these two subjects, all were unfamiliar with the system. 17 of the 28 subjects stated that they were experienced computer users. 5 subjects also mentioned that they had previously used a PDA, including one very experienced user, 3 middle experienced users, and one not at all experienced with PDAs. A summary of the subject demographics can be seen in figure 6.3, however due to the sample size, gender is the only category for which trends were obtained.

6.2.2 Quantitative Analysis and Results

Many of the findings outlined in this section are based on methods of statistical data analysis, and in particular the non-parametric tests: Chi-square (used for testing goodness of fit and independence), Mann-Whitney U (for testing two independent samples), and Wilcoxon (for testing two dependent samples). The findings are, where relevant, also accompanied by their measure of significance. ‘Statistical significance’ - also known as ‘p’ (p-value) or ‘ σ ’ - refers to the probability that an observed relationship found in a data set occurred by pure chance and is thus not actually representative of the data set. Generally speaking, significance is a measure on the degree to which a result can be seen as being true. In this dissertation, $p=\sigma=0.05$ (i.e. 1 in 20) is used as the cutoff value and implies that statistical significance occurs only if there is a less than 5% chance that a relationship is due to chance. For more detail on the background of statistical data analysis and the Chi-square, Mann-Whitney U, and Wilcoxon tests, see Hill and Lewicki (2006).

The 23 modality combinations provided to the subjects all stem from the unimodal language [*<Query>*]*<Feature>**<Object>*, where a ‘feature’ (FTR) may be entered via speech, handwriting, and intra-gesture, and an ‘object’ (OBJ) may be entered via speech, handwriting, intra-gesture, and extra-gesture. As described in section 4.1.1, the individual modality combinations will be referred to via their abbreviations, where for example ‘SS’ is analogous to *<Feature modality=“speech”>**<Object modality=“speech”>*.

6.2.2.1 Effects of a Consolidated View

The practical results obtained from the observation component and the written results obtained from the written questionnaire show a consolidation of user preference between the time that subjects trialled each individual modality combination in practice, and the time that they rated each modality combination when filling in the written questionnaire. The practical results are useful in that they depict a subject’s initial feelings on a modality combination immediately after having interacted with the system, and it is for this reason that these practical values are taken as the basis for evaluating all remaining aspects of the studies.

Analysing the differences between practical and written components in more detail (see figure 6.4), it can be seen that only a few modality combinations exhibit large swings between their practical and written values. In the laboratory study, the largest swings were positively inclined and encompassed the modality combinations SGI (Overlapped FTR), GIGI and GIGE (+0.86, +0.64, and +0.5 preference points respectively), while the modality combination HH experienced the largest negative swing (-0.5). The largest swings within the real-world study occurred for the modality combinations HGE and GIGE, both of which dropped in value between the practical and written components (-0.88 and -0.66 preference points respectively).

A notable trend that can be seen from the results, and in particular from the laboratory study

results shown in figure 6.4A, is that those modality combinations rated better within the practical component (i.e. ‘maybe yes’ and ‘prefer yes’) almost always rose in value when rated in the written component, while those rated worse (‘maybe no’ and ‘prefer no’) were almost always rated lower in the written component. This (de)amplification of the preference ratings further iterates the consolidation of subject opinions that took place during the practical and written components of the study.

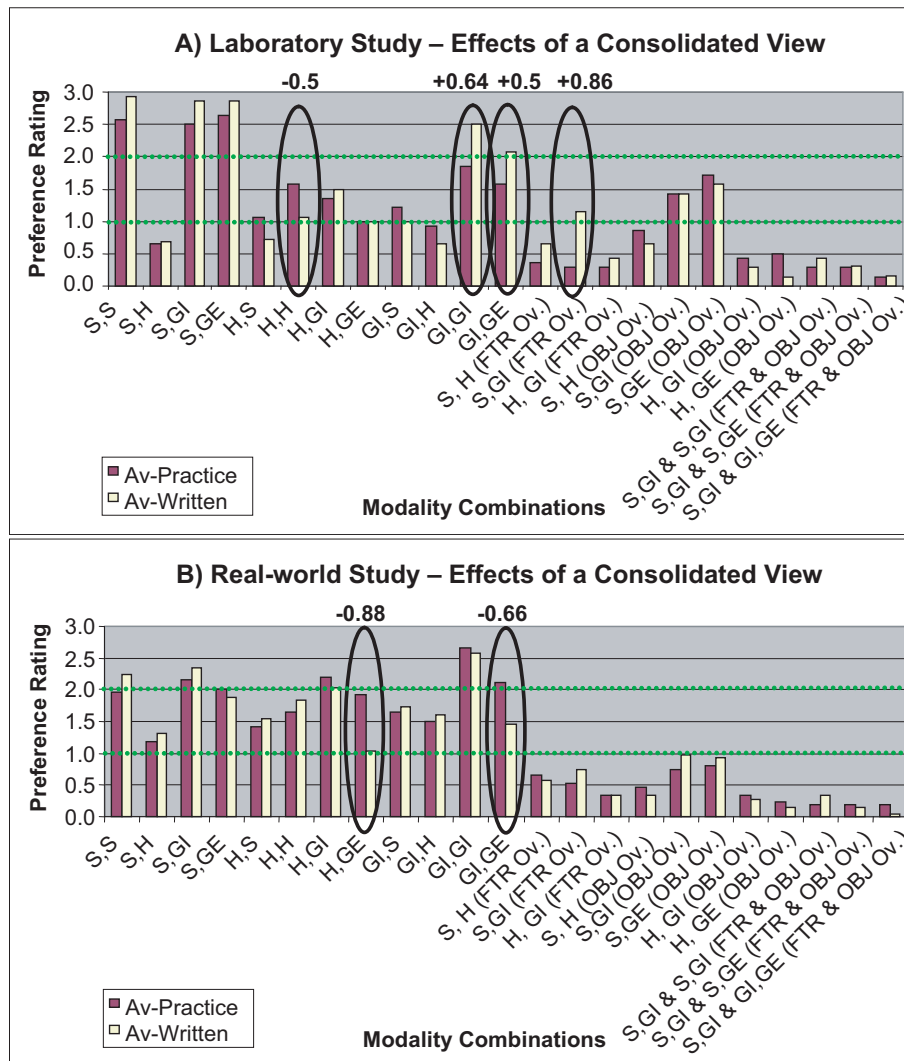


Figure 6.4: The effects of user preference consolidation as shown in A) the laboratory study and B) the real-world study.

Another aspect to note with regards to the results in these usability studies is that the preference values that subjects allocated to the individual modality combinations are by far and large conservative. This is assumed to be due to subjects being allowed to implicitly select the most intuitive - and thus better - modality combinations to start with, thus raising the bar for all subsequent modality combinations that by default were less intuitive than the first modality combinations that were chosen.

6.2.2.2 Preferred Modality Combinations Ranked by Feature Group

Laboratory Study: Figure 6.5 shows the modality combinations categorized into the groups semantically non-overlapped and overlapped. From the averages (A_v) shown in figure 6.5A, it can be seen that subjects prefer non-overlapped modality combinations ($A_v=1.58$) over overlapped modality combinations ($A_v=0.60$). Using a Mann-Whitney U-test, this was also shown to be statistically significant in 8 out of 14 subjects: $U(12,11)<35$, $p<0.05$, and only 3 subjects had a $p>0.12$.

The non-overlapped combinations have been further grouped according to their feature modality (i.e. the modality used to input feature attributes like 'price'). Analysing each of these non-overlapped subgroups - speech, handwriting, and intra-gesture - it can be seen that the use of speech for the feature ($A_v=2.09$) is notably preferred to intra-gesture ($A_v=1.39$) and handwriting ($A_v=1.25$). Unimodal modality combinations also received the highest score for the subgroups intra-gesture (GIGI, preference=1.86) and handwriting (HH, 1.57), while speech (SS, 2.57) was only marginally below the top rated modality combinations within its subgroup.

The overlapped combinations can also be categorized by their overlapping segment types: feature, object, or both feature and object. From figure 6.5A it can be seen that subjects preferred overlapped object information ($A_v=0.99$) to overlapped feature information ($A_v=0.31$), and both overlapped feature and object information ($A_v=0.24$). Indeed when compared to the rating scale, subjects would prefer not to use the latter two sets of combinations. The rating for overlapped object information increases to $A_v=1.33$ when speech is set as one of the overlapped modalities and increases further to $A_v=1.57$ when handwriting is excluded from the possibilities (i.e. overlapped speech and gesture for the object). This can be taken to imply rather logically that speech (in comparison to handwriting) is a modality that subjects prefer to use when providing duplicate information.

Real-world Study: As in the laboratory study, subjects preferred semantically non-overlapped combinations ($A_v=1.87$) over semantically overlapped modality combinations ($A_v=0.43$), and a Mann-Whitney U-test showed this to be significant in 23 out of 26 subjects: $U(12,11)<35$, $p<0.05$. For the remaining 3 subjects: $U(12,11)<38$, $p<0.091$.

In comparison to the laboratory study in which the use of speech for providing the feature was the preferred modality group ($S_{A_v}=2.09$, $GI_{A_v}=1.39$, $H_{A_v}=1.25$), intra-gesture was the preferred modality group in the real-world study ($GI_{A_v}=1.98$, $S_{A_v}=1.83$, $H_{A_v}=1.80$). Another difference between the studies is that the laboratory study exhibits a much greater difference between preference ratings for the modality groups than the real-world study, which is expected to be due to the modality of speech excelling in preference when under private laboratory conditions.

Similar to the real-world study, the unimodal combinations GIGI (preference=2.65), SS (1.96), and HH (1.65) all rated well within their individual modality groups, although HH was exceeded by the mixed combinations HGI and HGE.

Figure 6.5B also shows the overlapped combinations grouped by their overlapping segment types: feature ($A_v=0.51$), object ($A_v=0.52$), and feature and object combined ($A_v=0.19$). Similar to the laboratory study, these groups would not be used by subjects in the given scenario. One reason for this, as stated by the subjects was that the system worked fine without needing to provide duplicate information. Scenarios exhibiting a harsh environment (e.g. very noisy), or where the user is vulnerable to making mistakes (e.g. running, see Kumar et al. (2004)), or in applications where high levels of accuracy are required (e.g. flight controllers) may provide differing results.

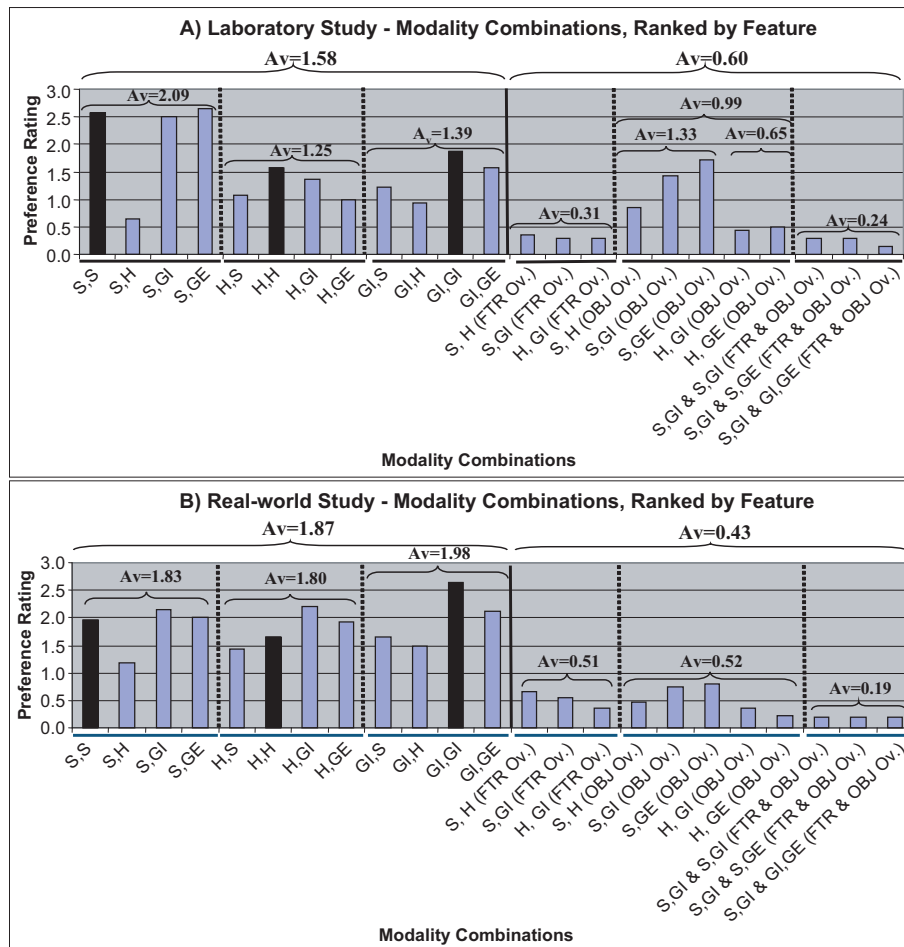


Figure 6.5: Preferred modality combinations ranked by feature group as shown in A) the laboratory study and B) the real-world study. The unimodal modality combinations are represented in a darker colour.

6.2.2.3 Preferred Modality Combinations Ranked by Preference

Laboratory Study: Figure 6.6 shows the modality combinations ranked in order of user preference. It can be seen that SGE (2.64) is the most preferred modality combination and that this is very closely followed by SS (2.57) and SGI (2.50). Using the Mann-Whitney U-test, the preference for these three modality combinations when compared to all other modality combinations was significant in 12 out of 14 subjects: $U(3,20) < 9, p < 0.05$. The success of these three modality combinations was further iterated by the comments that subjects made in the written component, and it is interesting to note that these modality combinations are directly representative of how people interact with other people and in particular with sales assistants.

The benefit of allowing for deictic references can also be seen in that 2 of the top 3 modality combinations, and 7 of the top 9 modality combinations used gesture to identify the object. The successful pairing of the modalities speech and gesture is further exemplified within the group of overlapped object combinations, where the preference of SGE (1.71) and SGI (1.43) was shown to be significant in 6 from 14 subjects (U-test, $p < 0.36$) when compared to the other overlapped

modality combinations.

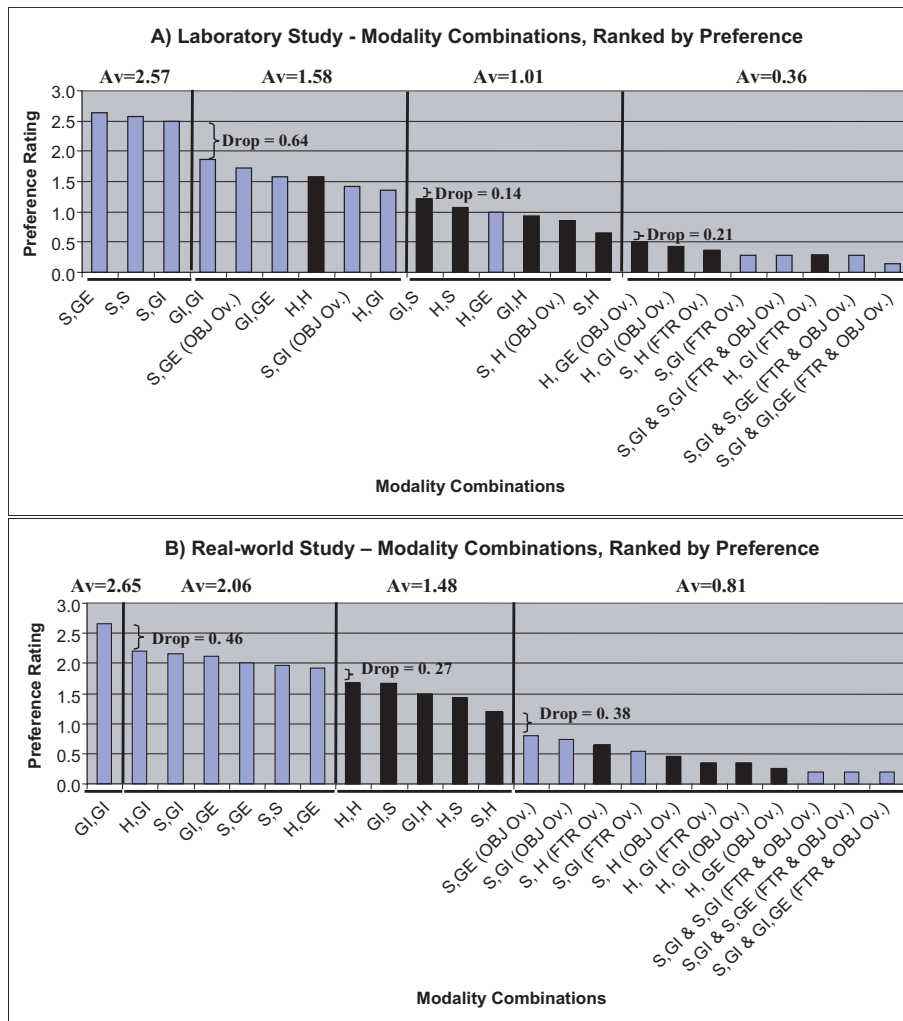


Figure 6.6: Preferred modality combinations ranked by preference as shown in A) the laboratory study and B) the real-world study. The modality combinations that had not been implemented at the time of testing are represented in a darker colour.

As shown in figure 6.6A, the 23 modality combinations have been grouped according to rating point falls between the individual modality combinations, where the first drop of 0.64 borders on significant (Wilcoxon, $z=-1.807$, $p=0.071$). The first set of modality combinations is preferred by the subjects ($Av=2.57$), while the second set of modality combinations lie midway between the categories ‘maybe no’ and ‘maybe yes’ ($Av=1.58$), and the third set of modality combinations has a ranking value equivalent to ‘maybe no’ ($Av=1.01$). These results on modality combination groupings are expected to be applicable to many other areas in which mobile multimodal interaction is a requirement. The results in effect form a guideline for implementing modality combinations that users will ‘want’ to use.

The darker columns in figure 6.6 represent those modality combinations that were not implemented in the system, and although most of these modality combinations exist on the lower side

of the ranking scale (implying a well thought-out system design), the value of the modality combination HH was clearly underestimated and should thus also be a provided interaction means in any good multimodal design.

Real-world Study: As shown in figure 6.6B, GIGI (2.65) can be seen to be the most preferred modality combination tested in the real-world study, followed by HGI (2.19) and SGI (2.15). The modality combinations SGE, SS, and SGI, which were rated highest in the laboratory setting can now be seen to rank 5th (2.00), 6th (1.96), and 3rd (2.15) in the real-world usability study respectively, showing the clear preference that subjects had for the non-observable modalities such as handwriting and intra-gesture, over speech and extra-gesture.

Similar to the laboratory study, the use of deictic references in identifying an object occurred in 6 out of the 7 top modality combinations, implying that the use of deictic references when interacting multimodally are a much liked tool. SGE (overlapped object) and SGI (overlapped object) were similarly the most preferred overlapped modality combinations, indicating that these combinations although not rated very high, still stand a good chance of being accepted by users.

A further grouping of the modality combinations, based on the large rating point falls occurring between the individual modality combinations, shows that the first drop of 0.46 is very close to significant (Wilcoxon, $z=-1.930$, $p=0.054$) and that the third drop of 0.38 is significant (Wilcoxon, $z=-1.978$, $p=0.048$). The first set of modality combinations (or to be precise, the modality combination GIGI) is preferred by subjects, while the second set ($Av=2.06$) equates to ‘maybe yes’, and the 3rd set ($Av=1.48$) lies midway between the category ‘maybe no’ and ‘maybe yes’. The fourth set of modality combinations (marked by the significant drop of 0.38 preference points) is made up entirely of the 11 overlapped modality combinations, which reiterates the subjects’ lack of preference for this type of interaction when in a real-world setting.

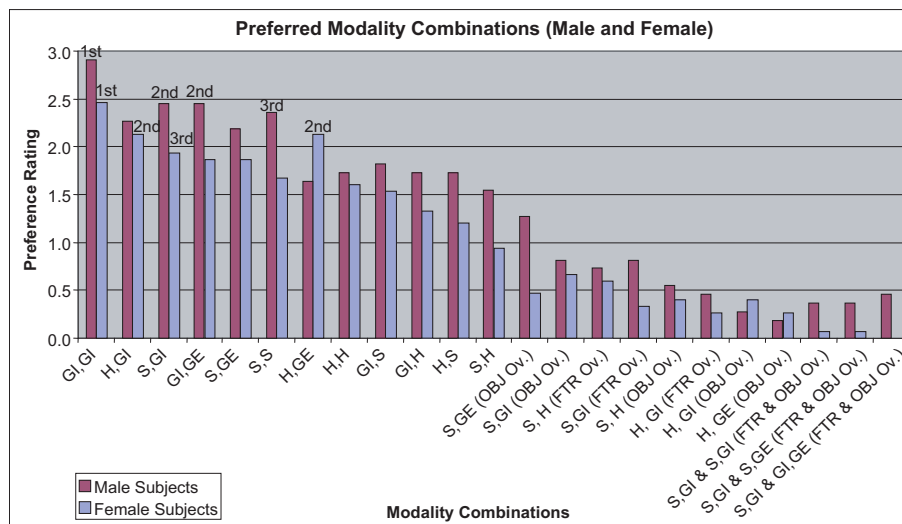


Figure 6.7: Preferred male and female modality combinations.

Due to the increased subject population size in the real-world study, it was possible to analyse modality combination preferences based on gender, and to identify some general trends (see figure 6.7). In particular, one can see that the most preferred combinations for men were GIGI (2.91), SGI (2.45), GIGE (2.45), and SS (2.36), while those for women were GIGI (2.47), HGI (2.13),

HGE (2.13), and SGI (1.93). The largest differences between men and women were seen in the modality combinations SGE (Object overlapped, difference of -0.81), SS (-0.70), and SH (-0.61), perhaps attributable to the observable nature of speech. Also of relevance is that the preference ratings provided by women were more conservative than those provided by men (i.e. lower for all but 3 modality combinations), as can be seen in table 6.1.

	Male	Female	Difference (Female-Male)
Av (Non-overlapped)	2.07	1.72	-0.35
Av (Overlapped)	0.57	0.32	-0.25
Av (All)	1.35	1.05	-0.30

Table 6.1: Male and female preferences for non-overlapped and overlapped modality combinations.

6.2.2.4 Modality Intuition

One of the conducted tests dealt with how intuitive the different modality combinations were to use. Modality intuition was measured in two separate tests, the first conducted during the written component where subjects were asked to answer the question: “Do you feel that this modality combination was intuitive to use?” (‘no’, ‘yes’), and the second conducted during the practical component where the first four modality combinations used by subjects were recorded. These first four modality combinations were weighted exponentially, such that a modality combination chosen 1st received a weighting of 1000, while modality combinations chosen 2nd, 3rd, or 4th received the values 100, 10, and 1 respectively. The resulting weights for the individual modality combinations are shown in the bottom right of figures 6.8B and 6.9B. The reason that only the first four modality combinations were used as the second measure of intuition was that all subjects managed to complete four different modality combinations when initially interacting with the system, but many required help after this in remembering what other combinations still existed.

Laboratory Study: The written component (see figure 6.8A) showed that 5 of the 12 non-overlapped modality combinations (SS, SGI, SGE, GIGI, and HH) were rated significantly intuitive by the subjects: $\text{Chi}^2(1, N=14) > 10.286$, $p < 0.001$, while 2 out of these 12 non-overlapped modality combinations were rated significantly non-intuitive (SH and HS): $\text{Chi}^2(1, N=14) > 4.571$, $p < 0.033$. In comparison, 6 of the 11 overlapped modality combinations were rated significantly non-intuitive by the subjects. These included HGI (feature overlapped), HGI (object overlapped), HGE (object overlapped), and all 3 combinations with both the feature and object overlapped: ($\text{Chi}^2(1, N=14) > 4.57$, $p < 0.033$). SGI (overlapped object) and SGE (overlapped object) were however rated intuitive by the majority of subjects.

When correlated with the lower graph in figure 6.8 one can see that the modality combinations SGI, SGE, SS, and GIGI were mirrored as being intuitive. The modality combination HH was however never selected for use by any of the subjects within their 1st four interactions, despite 13 out of 14 subjects rating the modality combination as being intuitive during the written component. Many people commented that handwriting was too slow to use, and perhaps this was a reason why the subjects never selected HH within the practical component.

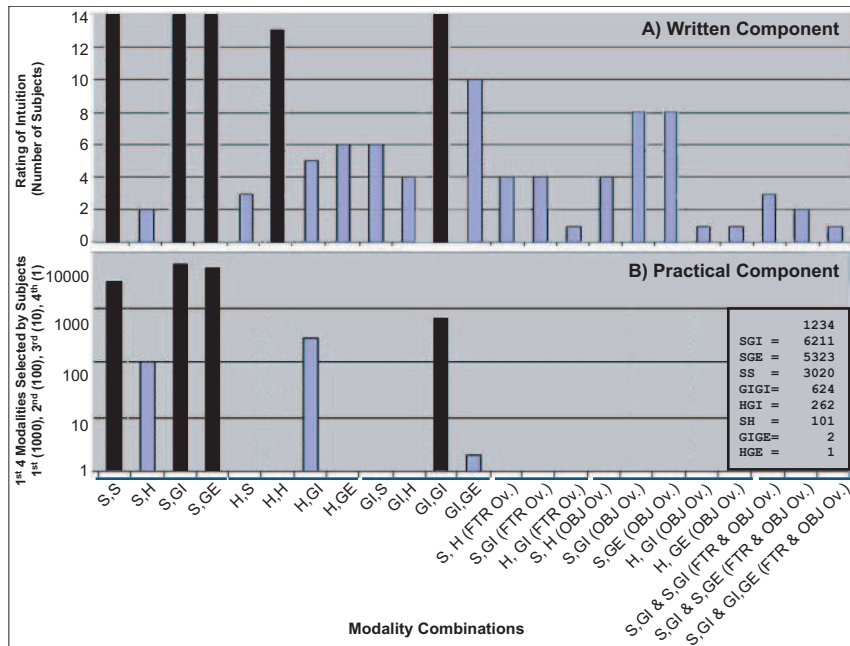


Figure 6.8: Intuitiveness of the 23 different modality combinations as rated during A) the written and B) the practical components of the laboratory study. The modality combinations that were rated significantly intuitive in the written component are represented in a darker colour.

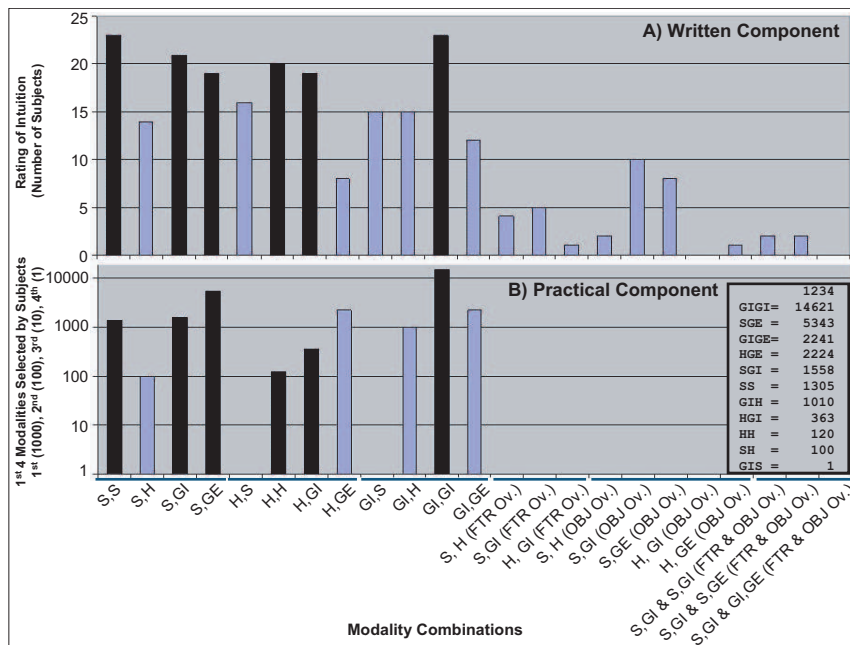


Figure 6.9: Intuitiveness of the 23 different modality combinations as rated during A) the written and B) the practical components of the real-world study. The modality combinations that were rated significantly intuitive in the written component are represented in a darker colour.

The overlapped modality combinations, SGI (object overlapped) and SGE (object overlapped) were also never used within the 1st four modality combinations, which may have been due to the combinations simply being overlooked by subjects due to the already wide range of non-overlapped modality combinations to choose from.

Figure 6.10A shows the number of times and the order in which subjects selected their first four unique modality combinations. It shows that SGI and SGE were selected 1st most often, while GIGI was the 2nd choice, and HGI the 3rd choice for most subjects. These results further emphasize the intuitiveness of these select modality combinations.

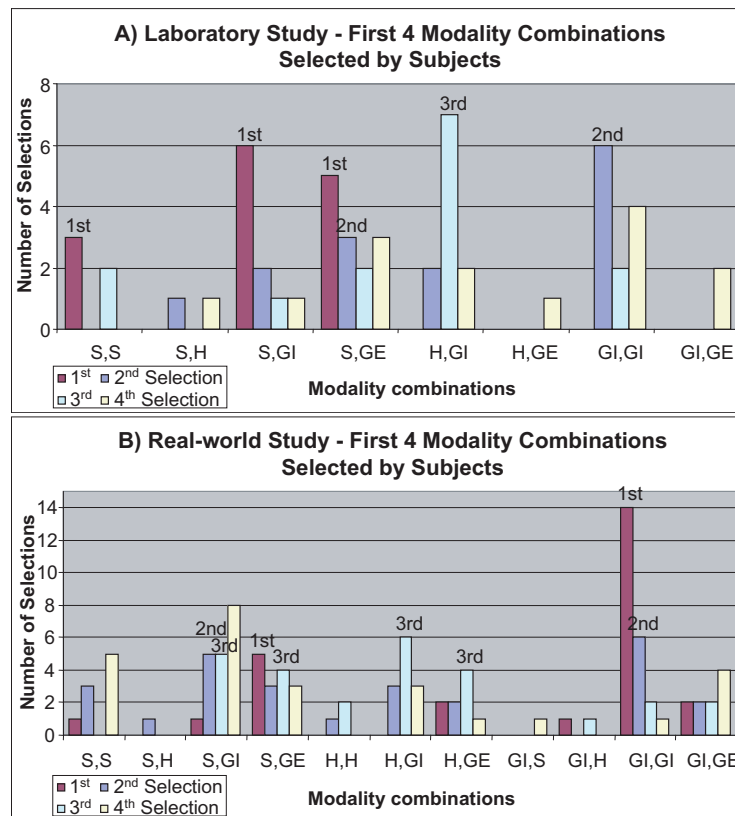


Figure 6.10: The first 4 modality combinations selected by subjects in A) the laboratory study and B) the real-world study. The classifications 1st, 2nd, and 3rd denote the order in which modality combinations were most often selected, for example in B), SGI was selected 1st most often.

Real-world Study: For the real-world study, the written component showed that 6 out of the 12 non-overlapped modality combinations (SS, GIGI, SGI, HH, SGE, and HGI) were rated significantly intuitive by the subjects: $\chi^2(1, N=25) > 6.760$, $p < 0.009$, while the modality combinations HS, GIS, GIH, SH, and GIGE were rated intuitive by more than half of the subjects. In comparison, 9 out of the 11 overlapped modality combinations - all except for SGI (object overlapped) and SGE (object overlapped) - were rated significantly non-intuitive by the subjects: $\chi^2(1, N=25) > 9$, $p < 0.003$.

When correlated with the lower graph in figure 6.9, one can see that the six modality combinations rated as being significantly intuitive in the written component do in fact correlate with

the practical component. Furthermore, these combinations also encompass the set classified as significantly intuitive in the laboratory study.

Similar to the laboratory study, figure 6.10B shows the number of times and the order in which subjects selected their first four unique modality combinations during the practical component. GIGI, SGI, and HGI, all rate highly in this section, further emphasizing the intuitiveness of these modality combinations, not just under laboratory conditions, but also under real-world settings.

6.2.2.5 Modality Usage in Public and Private Environments

A topic that was covered inside the scope of the written questionnaire had to do with how the subjects would feel using the modalities speech, handwriting, intra-gesture, and extra-gesture, firstly within a public environment (e.g. in a shopping mall), and secondly within a private environment (e.g. at home). These public and private settings were taken to be analogous to the real-world and laboratory settings in which the studies took place. A twist to the task was that subjects that took part in the laboratory study had to hypothesize about how they would feel when using the base modalities in a real-world setting, and those subjects that took part in the real-world study had to conversely hypothesize about modality usage in a laboratory setting. The scale provided to the subjects for this task was: 'comfortable', 'hesitant', and 'embarrassed'.

Chi-square tests showed that the subjects who took part in the laboratory study would (hypothetically) feel comfortable using intra-gesture, extra-gesture, and handwriting (but not speech) within a public environment: $\text{Chi}^2(2, N=14) > 8.714$, $p < 0.013$, and would feel comfortable using all base modalities in a private environment: $\text{Chi}^2(2, N=14) > 8.714$, $p < 0.013$. These results were identical with those generated from the real-world study, in which subjects confirmed that they would feel comfortable using intra-gesture, handwriting, and extra-gesture (but not speech) within a public environment: $\text{Chi}^2(2, N=27) > 12.667$, $p < 0.002$, and would (hypothetically) feel comfortable using all base modalities in a private environment: $\text{Chi}^2(2, N=27) > 10.889$, $p < 0.004$.

Interesting to note is that as shown in figure 6.11, for both the laboratory and the real-world studies, the values provided for intra-gesture and handwriting remained the same when used in public and private environments, while the values provided for extra-gesture and speech changed. This affirms the notion that modalities which may be witnessed by surrounding people, although often rated as preferred, have a trade-off with respect to user privacy for both public and private environments.

The modality of handwriting provided mixed results with regards to comfortability, and the reasons for this are thought to be two-fold. Firstly, many subjects said that the modality was impractical to use because it took too long to write on the small display. Secondly, subjects often stated that they were very self-conscious about the legibility of their handwriting, regardless of whether or not their input was accurately recognized.

With regards to gender, the non-observable modalities were rated similarly by both men and women, while the observable modalities showed notable differences, particularly when used in a public environment. Gender differences for the observable modalities are shown in table 6.2. In a public environment, only 33% of women said they would feel comfortable using speech in comparison to 67% of men, and only 40% of women would feel comfortable using extra-gesture in comparison to 83% of men. In a private environment, the main gender difference for the observable modalities was with regards to extra-gesture, where only 53% of women would feel comfortable using extra-gesture in comparison to 100% of men.

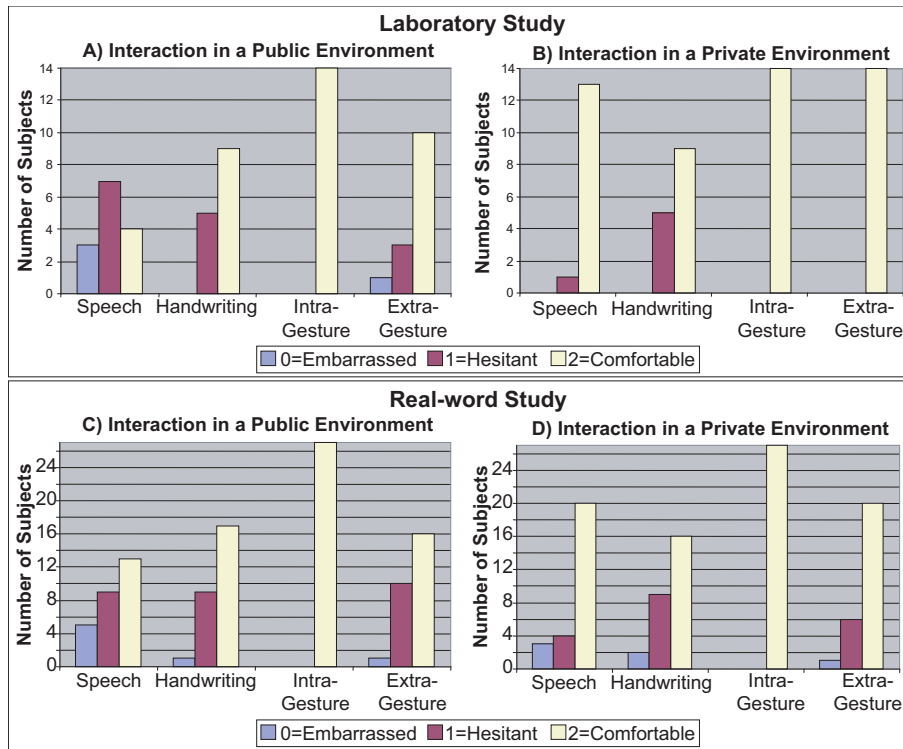


Figure 6.11: Modality usage in public and private environments as shown in the laboratory study (A and B) and the real-world study (C and D).

	Public Environment				Private Environment			
	Speech		Extra-Gesture		Speech		Extra-Gesture	
	Female	Male	Female	Male	Female	Male	Female	Male
Comfortable	33%	67%	40%	83%	73%	75%	53%	100%
Hesitant	40%	25%	53%	17%	20%	8%	40%	0%
Embarrassed	27%	8%	7%	0%	7%	17%	7%	0%

Table 6.2: Modality usage in public and private environments illustrating gender differences.

6.2.2.6 The Effects of Observability on Modality Combination Preference

Modality combinations that are (non)observable by surrounding people are used as an index for measuring concerns that a user might have when interacting with certain modalities in public or in private. In comparison to the above section, which analysed the effects that public and private environments might have on the base modalities speech, handwriting, intra-gesture, and extra-gesture, this section focuses on a subset of individual modality combinations and draws its results from the preference ratings that subjects provided throughout the practical components of each study. For the purpose of this task, the two most preferred ‘entirely observable’ and the two most preferred ‘entirely non-observable’ modality combinations are compared. Entirely observable modality combinations refer to those comprised only of the modalities speech and extra-gesture (SGE, SS), while the entirely non-observable modality combinations refer to those comprised only of the modalities handwriting and intra-gesture (GIGI, HH, HGI, GIH).

Within the laboratory study, the two most preferred entirely observable modality combinations (SGE=2.64, SS=2.57, Av=2.61) rated higher than the two most preferred entirely non-observable modality combinations (GIGI=1.86, HH=1.57, Av=1.72). This implies that at least for the product type ‘digital cameras’, the feelings of embarrassment and hesitation described in the above section had little effect on modality combination preference under a laboratory setting. For the real-world study however, where an average of 13.8 people could be seen from the shelf’s location during each of the tests, a notable shift in subject preference was recorded. In particular, for this real-world setting the two most preferred non-observable modality combinations (GIGI=2.65, HGI=2.19, Av=2.42) were rated higher than the two most preferred observable modality combinations (SGE=2.00, SS=1.96, Av=1.98). As outlined in table 6.3, these results imply a significant modality preference shift towards non-observable modalities when in a public environment and a significant modality preference shift towards observable modalities when in a private environment.

	Laboratory Study	Real-world Study	Difference (Real-Lab)
Observable MCs (SGE, SS)	Av=2.61	Av=1.98	-0.63 U(14,26)=96, p=0.011
Non-observable MCs (GIGI, HH)	Av=1.72	Av=2.42	+0.70 U(14,26)=116, p=0.056
Difference	-0.89	+0.44	

Table 6.3: Effects of observability on Modality Combination (MC) preferences during interaction in private (laboratory) and public (real-world) environments, and supported by Mann-Whitney U-test significance values.

This trend was also well supported by subject comments made throughout the studies. For example, subjects stated that they did not like the observable nature of speech, and extra-gesture was also seen to exhibit an observable nature because third parties could easily identify the objects that one physically interacted with. For these two modalities, subjects commented that their use of the modality would be situation-dependent (e.g. dependent on whether the modality might disturb surrounding people or arouse undesired attention from other people). The use of these modalities was also said to be product-dependent, as some subjects commented that using speech interaction on a can of baked beans would be excessive, but would be suited to products with complex functionality such as a TV. Furthering this, subjects stated that they would not (or could

imagine not being permitted to) use extra-gesture on fragile products (e.g. porcelain) or high-value products (e.g. jewellery). Socially sensitive products, such as clothing undergarments, was another area for which subjects stated they would be weary in using speech and extra-gesture. Finally, the comments from subjects indicate that liking a modality is also user-dependent, and this has significant effect on the observable modalities speech and extra-gesture. For example, some subjects stated that they like to talk, and yet other subjects said that they would not mind using extra-gesture in a public environment because touching a product is part of the buying process.

6.2.2.7 General Results Regarding the MSA Demonstrator

One set of questions that the subjects in the real-world usability study were asked to answer dealt with general aspects on interaction with the system, such as what their favourite modality combination was, whether or not they would ever consider using overlapped modality combinations ('yes', 'maybe', 'no'), how they would rate the learnability of the system ('easy', 'ok', 'not easy'), and whether or not they would consider using the presented form of shopping as an alternative to their current practices ('yes', 'maybe', 'no'). The results are as follows:

- GIGI was noticeably rated the favourite modality combination in the real-world study (13 out of 27 subjects). This was then followed by the modality combinations SGE (4 subjects) and SGI (3 subjects).
- 21 subjects would not consider using overlapped modality combinations, which is significant: $\text{Chi}^2(2, N=27) > 24.889$, $p < 0.000$. However, as stated earlier, this result is not representative of all mobile scenarios as the recognition accuracy of the MSA system was fairly high and thus did not warrant the subject supplying additional (redundant) information.
- 23 subjects thought that the system was easy to learn, which is significant: $\text{Chi}^2(2, N=27) > 32.889$, $p < 0.000$.
- 18 subjects would consider using the presented form of shopping as an alternative to their current shopping practices: $\text{Chi}^2(2, N=27) > 14$, $p < 0.001$, while another 6 might, and only 3 would not. Aside from emphasizing that applications supporting multimodal interaction are beneficial to users, these results show that the pool of subjects was open to new interaction techniques and thus change in general.

6.2.3 Quantitative Comparisons Between the Studies

The goal of this section is to highlight the significant differences that occurred between the laboratory and the real-world studies through direct comparison of the results (Wasinger & Krüger, 2006).

As seen in table 6.4, a first distinction is that when compared to the laboratory study, subjects in the real-world study rated non-overlapped modality combinations better (with a preference of $A_v=1.87$ versus 1.58) and overlapped modality combinations worse (with a preference of $A_v=0.43$ versus 0.60). This preference (de)amplification for modality combination types is thought to be due to the differences between public and private environment settings (such as privacy concerns).

Another interesting result arising from the comparison of the laboratory and the real-world studies is the difference in individual modality preference ratings provided by the subjects, as shown in figure 6.12. Using the Mann-Whitney U-test, this difference was shown to be significant

in 9 of the modality combinations, whose asymptotic 2-tailed significance values are shown in table 6.5. Note that these significance values are based on the combined set of 40 samples (14 subjects from the laboratory study and 26 subjects from the real-world study) and that no subject took part in both studies. These results further support the difference that exists between the use of observable and non-observable modality combinations in a private and public environment.

	All Combinations			Non-overlapped			Overlapped		
	Av	Max	Min	Av	Max	Min	Av	Max	Min
Laboratory	1.11	2.64	0.14	1.58	2.64	0.64	0.60	1.71	0.14
Real-world	1.18	2.65	0.19	1.87	2.65	1.19	0.43	0.81	0.19
Difference				+0.29	+0.01	+0.55	-0.17	-0.90	+0.05

Table 6.4: Summary of the differences in modality combination preference between the laboratory and the real-world studies, where ‘0=prefer not’, ‘1=maybe not’, ‘2=maybe yes’, and ‘3=prefer yes’.

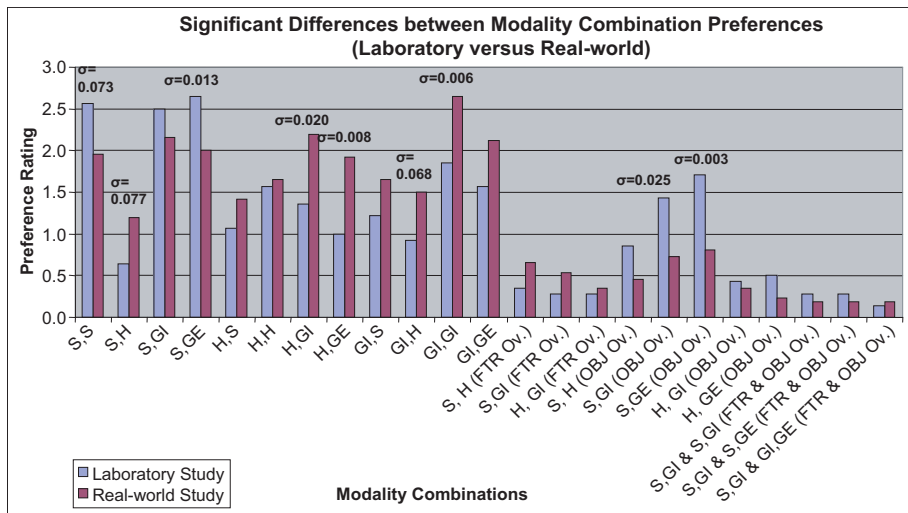


Figure 6.12: Comparison of modality combinations for the laboratory and the real-world studies showing significant differences (σ) in user preference.

Figure 6.13 illustrates the similarities in non-overlapped modality intuition that exist between studies. It can be seen that the modality combinations rated most intuitive (GIGI, SS, HH, SGI, and to a lesser extent SGE) correlate well between studies, implying that the intuitiveness of modality combinations is not affected by a change in environment setting. Although not significant, this trend was confirmed by a Mann-Whitney U-test ($U(13, 25)$), which tested the null-hypothesis to show that the differences ‘were’ due to chance (i.e. $p \geq 0.75$).

Another difference worthy of mention is with regards to the hypothesized values that subjects provided for using a modality in public, based only on their experience with the system in a private environment, and vice-versa. As illustrated in figure 6.14, testing the null-hypothesis shows that handwriting and intra-gesture correlate well for both public and private environments ($p \geq 0.75$), while speech and extra-gesture do not. The lack of correlation for speech and extra-gesture might

Modality Combination	Laboratory Values	Real-world Values	Difference Real-Lab	Asymp. Significance (2-tailed)
SS	2.57	1.96	-0.61	U(14,26)=123.5, p=0.073
SH	0.64	1.19	+0.55	U(14,26)=123.0, p=0.077
SGE	2.64	2.00	-0.64	U(14,26)=101.0, p=0.013
HGI	1.36	2.19	+0.83	U(14,26)=104.0, p=0.020
HGE	1.00	1.92	+0.92	U(14,26)=92.0, p=0.008
GIH	0.93	1.50	+0.57	U(14,26)=121.5, p=0.068
GIGI	1.86	2.65	+0.79	U(14,26)=99.5, p=0.006
SIGI (Obj Ov.)	1.43	0.73	-0.70	U(14,26)=107.5, p=0.025
SGE (Obj Ov.)	1.71	0.81	-0.90	U(14,26)=83.5, p=0.003

Table 6.5: Modality combinations exhibiting a significant difference in preference between the laboratory and real-world studies.

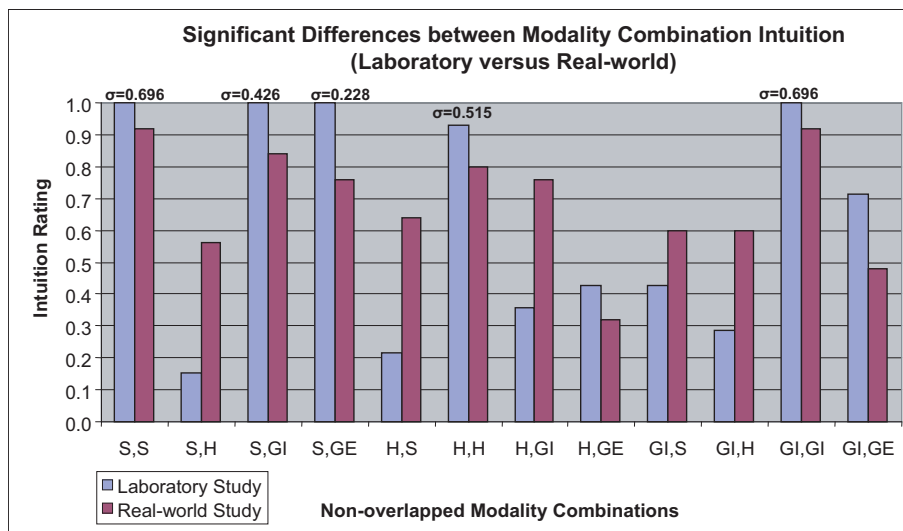


Figure 6.13: Significant differences between modality combination intuition arising from the analysis of results across both studies.

imply that subjects cannot easily estimate the impact of using observable modalities in public and private environments without first obtaining hands-on experience with these modalities in such settings.

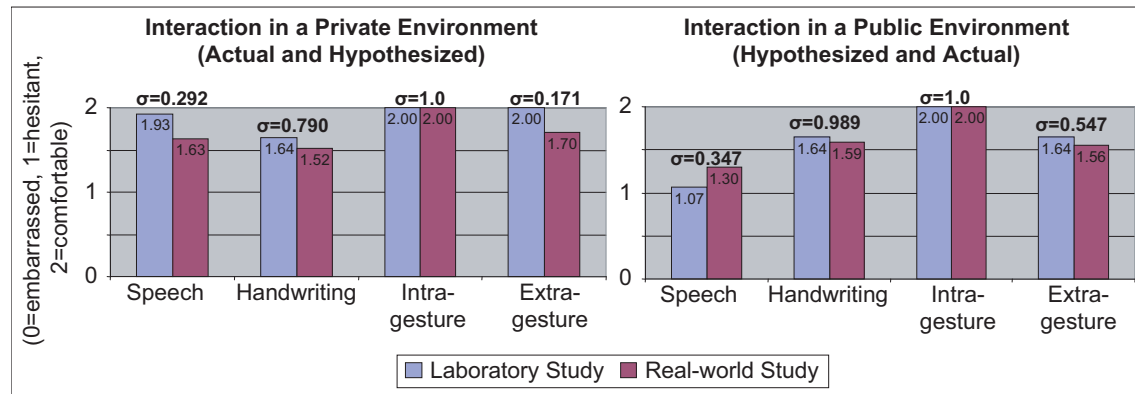


Figure 6.14: Differences between hypothesized and actual modality preference values in the real-world and laboratory studies.

6.2.4 Anthropomorphized Objects

The usability studies that were conducted also analysed the concept of ‘anthropomorphized objects’, and ‘indirect’ versus ‘direct’ interaction. As outlined in section 4.3, anthropomorphization is the tendency for people to think of inanimate objects as having human-like characteristics. To realize this, the MSA had two different operating modes, one in which the objects did not take on human-like characteristics (the default mode used for the majority of the study) and one in which the objects spoke directly to the subjects. During the laboratory usability study, subjects were asked to complete a written component that asked them whether or not they would (hypothetically) prefer to interact with the digital camera objects indirectly (e.g. “What is the price of this camera?”) or directly (e.g. “What is your price?”). These results from the laboratory study were then complemented with a larger set of results from the real-world study that dealt with direct and indirect interaction and anthropomorphization. The testing of these concepts took the form of a practical and written component, which was conducted after subjects had completed the tasks on modality combination preferences as described in the previous section. For this second set of tests, the MSA was set to support product anthropomorphization and subjects were instructed to select an object from the shelf and to ask it about several of its features using the 2nd person tense, for example “What is your price?” and “How many megapixels do you have?”. Objects that were picked up from the shelf automatically initiated dialogues with the subjects by introducing themselves, for example “Hi there, I’m the new camera from Canon with 5 megapixels”. All further output by the objects was also conducted in the 1st person tense, for example “My price is €599” and “My focal length is 35 to 105mm”. User interaction with the system in this mode generally lasted around five minutes, which was sufficient time for the subjects to understand the underlying concepts of anthropomorphization and (in)direct interaction. During this time, subjects were given a series of smaller sub tasks to complete such as to find the cheapest camera on the shelf and to find the camera with the largest number of megapixels.

6.2.4.1 Direct and Indirect Interaction

Laboratory Study: From the 14 subjects studied in the laboratory setting, 11 (or 79%) said that given a choice between direct and indirect interaction they could imagine preferring to interact indirectly with the products. From these 11 subjects, 8 would however convert to direct interaction if the system was designed to only support output of a direct nature. The ability to coerce a user into direct interaction with products was also supported by comments made by the subjects. For example, subjects stated that their choice in interacting directly with the system might be affected by their instinct to follow the objects' lead, or might arise out of courtesy and fairness to the object. Several subjects stated that efficiency might be an advantage of direct interaction because nouns such as 'PowerShot S70' and 'camera' can be replaced by simple pronouns such as the personal pronoun 'you'. One condition that was stated to be a necessity for direct interaction is that it would need to appear natural to talk to the object. Those that preferred indirect interaction were generally of the opinion that it was not natural to talk to objects and that they would feel silly talking to a product because it was not human.

These results from the laboratory usability study provide only a brief insight into user acceptance for anthropomorphization and (in)direct interaction. The remaining results discussed in this section summarize the findings on anthropomorphization obtained from the real-world usability study.

Real-world Study: The results from the real-world usability study show a much larger preference for direct interaction than that which was depicted in the laboratory study. This is perhaps attributable to the subjects gaining a firsthand experience in interacting directly with the anthropomorphized shopping products, and it is perhaps also due to the novelty of the concepts that were tested.

The first question that subjects were asked was which of the two interaction modes they preferred most. The proportion of subjects that preferred direct interaction over indirect interaction (18 from 27 subjects, 66%) signified a trend for direct interaction and thus anthropomorphization: $\text{Chi}^2(1, N=27)=3.00$, $p=0.083$. This tendency was also seen much more clearly in men than in women (see table 6.6), where 10 from 12 men (83%) stated that they preferred direct interaction: $\text{Chi}^2(1, N=12)=5.22$, $p=0.021$, which is significant.

	Sex		
	Male	Female	Total
Direct Interaction	10	8	18
Indirect Interaction	2	7	9
Total	12	15	27

Table 6.6: Male and female preferences for direct and indirect interaction.

Similar to the laboratory study, subjects were again asked if they would convert to direct interaction if the system was only able to support output of a direct nature. This increased the overall number of subjects that would communicate directly with the objects from 18 to 22 out of a total of 27 subjects (an increase from 67% to 81%), resulting in the overall number of subjects (22 from 27) being significant: $\text{Chi}^2(1, N=27)=10.70$, $p=0.001$. From the 9 people who originally stated that they would prefer indirect interaction, 7 were female and 2 were male. Important to

note is that a subject not willing to convert would end up in a dialogue with the system containing incoherent language similar to the following:

- U: “What is the price of this <Gesture> camera?”
O: “My price is €599”
U: “How many megapixels does this camera have? <Gesture>”
O: “I have 5 megapixels”

6.2.4.2 User-Product Relationships

One question that subjects were asked, was whether they would be more inclined to interact directly with an object if the relationship between them and the object were stronger. 13 out of 26 people (60% of all women, and 36% of all men) agreed that a stronger relationship with a particular object (e.g. something that they were particularly fond of) would make them more likely to interact directly with the object. When asked what types of objects they would consider talking directly too, the subjects mentioned: plants, soft toys, strategic computer games, electronic devices such as TVs and refrigerators, and houses. Several subjects added that the objects would need to be intelligent, technical, or complex. Food was split both ways, and three people commented that they would be happy to talk to anything and everything.

The notion that relationship has an effect in encouraging some users to interact directly with products is further supported in that the likeliness of direct interaction, for a range of different products, was always higher when the subject was classified as an ‘owner’ rather than as a ‘buyer’.

6.2.4.3 Direct Interaction with a Range of Products as a Buyer and as an Owner

A final question asked of the subjects was whether they would interact directly with a range of different product types (soap, digital camera, personal computer, and a car), first as a buyer (B) and then as the owner (O) of the product (see figure 6.15). For brevity, only the resulting significance values that were obtained from non-parametric chi-square tests are reported, where $df=1$, $N_{Buyer}=27$, and $N_{Owner}=22$. While only around 30% of subjects would interact directly with a bar of soap (as B: $p=0.034$, as O: $p=0.201$), around 70% of subjects said that they would interact directly with digital cameras (as B: $p=0.034$, as O: $p=0.033$), personal computers (as B: $p=0.012$, as O: $p=0.003$) and cars (as B: $p=0.336$, as O: $p=0.003$). It can be seen that most of these values are significant. Another visible trend is that subjects prefer interacting directly with the products as the owner rather than as a buyer, and this difference is best seen for the product ‘car’ (59% as a buyer, 82% as an owner), in which a Wilcoxon signed rank test showed this difference to be near-significant ($z=-1.890$, $p=0.059$).

Analysing differences between gender, men were more inclined to interact directly with the products ‘personal computer’ and ‘car’ when classified as the owner (100% of men vs. 69% of women, for both product types). A Mann-Whitney U-test showed this trend in gender difference to be: $U(16,12)=40.5$, which equates to $p=0.072$ for both product types. Men were in general more willing to interact directly with the products, and non-parametric chi-square tests show (with $df=1$, $N_{Buyer}=12$, and $N_{Owner}=9$) that although men would ‘not’ for example talk to soap (B: $p=0.021$, O: $p=0.317$), they would talk to digital cameras (B: $p=0.021$, O: $p=0.096$), personal computers (B: $p=0.004$, O: $p=0.003$), and cars (B: $p=0.248$, O: $p=0.003$). Subjects that said they would interact directly with a car often stated that a car was a kind of family member, while several subjects said

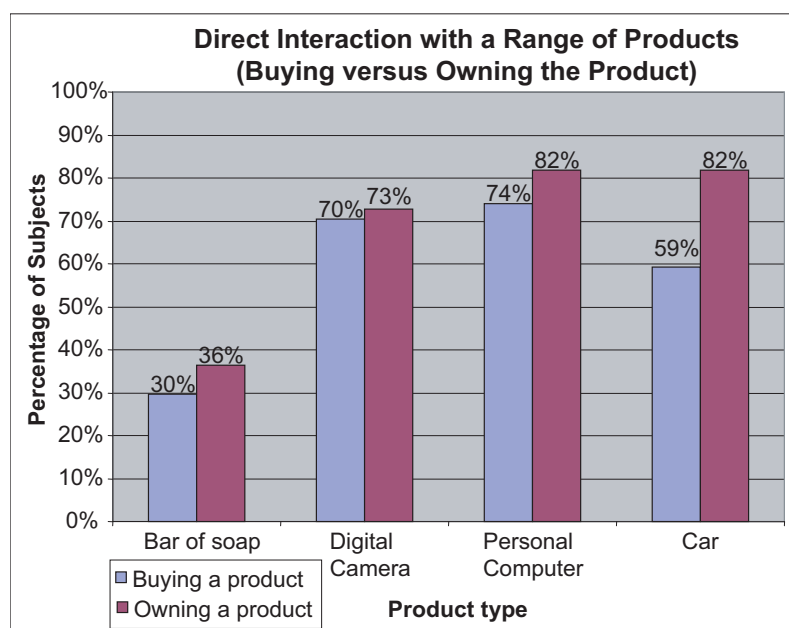


Figure 6.15: The effect that being the owner or a buyer can have on direct interaction with a range of different products.

that they would not interact directly with a car because of the similarity of this to the 1980's TV series 'Knight Rider'.

6.2.5 Qualitative Observations from the Studies

In this section, the qualitative results that were obtained from the above described studies are outlined. As discussed in (Wasinger & Krüger, 2005), these results form a general guideline in determining which base modalities and modality combinations to use when designing for mobile users, and the results may also be taken as a starting point for future empirical studies. The section starts with a discussion of the comments that subjects provided for the individual modalities and then summarizes the criterion that subjects saw to be important for multimodal interaction to be accepted by the public.

6.2.5.1 Modalities and their Combinations

Speech: Some subjects found the camera names (i.e. the object referents) such as 'PowerShot S11S' and 'FinePixA202' non-intuitive to pronounce via speech. Other subjects expressed concern about their spoken dialects, despite the system correctly understanding them and despite being told to disregard system failures. Subjects mentioned that they would find the modality of speech better if no buttons were required to start and stop the speech engine, which was observed to require some training. A single press to start (and an automatic stop via silence detection) was also stated to be a possible improvement to the modality. Subjects stated that allowing for semantically similar dialogue structures when providing speech input was also important (e.g. "What is the price of this?" and "What does this cost?"), as too was allowing for shorter and longer sentences (e.g. "Price?", "What is the price?", and "What is the price of this camera?"). One subject was left-handed, and

commented that the start/stop button was not in an intuitive location (the button to activate the speech recognizer was located on the lower left side of the PDA). In general, subjects stated that speech was an “excellent no-fuss modality”. They said it was fast, comfortable, interactive, and one subject made it very clear that she liked the modality because she enjoys talking.

Handwriting: Subjects commonly stated that it took too long to write, but that the use of abbreviations might improve the modality’s efficiency, for example ‘S70’ instead of ‘PowerShot S70’. When writing both feature and object information, the problems of handwriting were amplified. Subjects stated that the display space was too small and that the interaction took too long in comparison to the other modalities. One subject said that they preferred not to write on top of the product images on the PDA display. More so than in the other modalities, subjects were very self-conscious of their handwriting and feared that their input would be falsely recognized by the system. Some subjects stated that there were more comfortable modalities than handwriting. For the number of objects present in the product database (13), gesture was seen to be a better modality to use when selecting objects. It was however noted that similar to intra-gesture, handwriting was not easily observable and would thus be useful if privacy were required. To this extent, handwriting was also said to be good in that it would not disturb other people.

Intra-gesture: Regarding the use of intra-gesture for the selection of features (i.e. the visual-WCIS scroll bar), subjects commonly stated that the system would be better if one could see all of the options at the same time. It was stated that the modality was very fast, but that one had to wait at times, depending on whether the relevant keyword was currently visible. Some subjects stated that they would prefer the text not to scroll, while other subjects mentioned that it was important to be able to change the speed of the scrolling text (especially for large data sets), and still other subjects would have preferred a larger font size to be used for the scrolling text. Catering for such comments would however require more of the limited display space to be allocated to the visual-WCIS scroll bar. The general consensus was that searching and waiting for a feature to become visible was not good and that the modality would suffer if the user was under time pressure. Although the implemented visual-WCIS scroll bar was shown to have some weaknesses, it requires less mouse clicks when compared to a typical menu structure and does not remove a user from their current context by concealing large proportions of the display. Intra-gesture for object selection was seen by most subjects as an excellent modality, and even likened to a natural reflex.

Extra-gesture: Several subjects mentioned that touching a product was a fundamental requirement of shopping and also part of the buying process. It was suggested that just pointing at an object (or using the PDA as a pointing device) might be an appealing alternative if one’s hands were also required by other modalities like handwriting⁴. Although the weight of the empty camera boxes was insignificant, large or heavy objects were seen to pose a limitation for extra-gesture interaction, as did product placement (e.g. high-up or low-down products). Some subjects stated that extra-gestures would be even better if the scenario entailed only the user and the objects (i.e. no PDA), or a headset instead of the PDA so that the hands were still free. One extension to the

⁴When using the demonstrator, one hand is required to hold the PDA as well as to start and stop the speech recognizer, while the other hand is required to differing degrees by the modalities intra-gesture, extra-gesture, and handwriting

scenario that is being considered for the future, is to port the MSA to a tablet PC, which could then be mounted onto a shopping trolley to free up the user's hands as well as to provide a larger display for the user to write on. Subjects also stated that the boxes on the shelf should be replaced with the actual cameras, and it was also stated that it would be important for the products on the PDA's display to be sorted in a similar order to the products on the shelf to ease finding objects when navigating in a mixed-reality world. In general, extra-gesture was seen to be fun and interactive, but also one of the slower modalities.

Overlapped modality combinations: Some subjects formed fixed ideas early on during the usability study in that they categorized all overlapped modality combinations as being terrible. Subjects stated that these modality combinations were too complicated, took a lot of understanding and coordination, were time-consuming, and that the system should be able to understand a user without needing redundant information. Despite most of these modality combinations receiving a poor rating, subjects were generally aware that duplicate information would benefit the system, for example in ensuring that user input was correctly understood. Some of the overlapped modality combinations drew similarities to the process of thinking aloud, such as handwriting overlapped with speech, but for this combination subjects also stated that they felt silly because they would slur the spoken words due to handwriting being a much slower modality than speech.

6.2.5.2 Characteristics Considered Important for Multimodal Interaction

During the studies, subjects identified several aspects as being important for multimodal interaction. These included: comfort, enjoyment, familiarity, speed, accuracy, scale, accessibility, privacy, intuition, and the complexity of a modality combination.

Speech was for example seen as being 'comfortable' to use, as too was intra-gesture for object selection on the PDA display. Extra-gesture (i.e. picking up and putting down real-world camera boxes) was considered comfortable for the given scenario, but would have been rated uncomfortable had the subjects already have been carrying shopping bags or had the objects have been large and/or heavy. Handwriting (which required the use of both hands) was considered less comfortable to use. Extra-gesture was described as being 'enjoyable' by many subjects in comparison to intra-gesture, where subjects said that "clicking is boring". Subjects were however very 'familiar' with the modality of intra-gesture, which closely resembles mouse interaction. The 'speed' of the modalities also had an effect on modality preference, for example handwriting was seen to be a slow modality when compared to speech and intra-gesture. The perceived 'accuracy' of handwriting was also low despite the recognition being quite good.

Speech and handwriting were said to 'scale' better than gesture for large feature and object databases, in which it would be easier to speak out a product name than to first visually find a product and then point to it via intra-gesture. Speech and handwriting were also said to be better if an object were not 'accessible' (e.g. behind glass or out of stock). The observable modalities speech and extra-gesture were seen to disregard 'privacy', especially when used in a public environment. In comparison, the non-observable modalities handwriting and intra-gesture were noted by subjects to perhaps be beneficial when dealing with sensitive objects. Privacy was stated to be a greater concern for object information than for feature information. Some multimodal combinations (e.g. HS and GIS) and most overlapped modality combinations were seen to be less 'intuitive' than their non-overlapped and unimodal counterparts. Several modality combinations also incurred 'complexity' costs arising through modality switching, which was particularly

evident for combinations consisting of both on- and off-device interactions, such as HGE, in comparison to just on- or just off-device interaction (e.g. HGI) and unimodal interaction (e.g. SS). The benefits (but more often the disadvantages) of each modality combination were frequently amplified based on the individual characteristics of the encompassed modalities, which was particularly visible from the subject ratings for overlapped modality combinations.

6.2.6 Usability Study Conclusions

The results from these usability studies have highlighted many important characteristics about mobile multimodal interaction. Most importantly, the studies have shown that from the 23 different modality combinations offered to subjects within a mobile shopping scenario, intra-gesture and intra-gesture (GIGI), handwriting and intra-gesture (HGI), and speech and intra-gesture (SGI) were the most preferred combinations in a real-world or public environment, while speech and extra-gesture (SGE), speech and speech (SS), and speech and intra-gesture (SGI) were the most preferred combinations in a laboratory or private environment. For private environments, the combinations are also representative of how people would interact with other people. Unimodal interaction was well liked by the subjects perhaps due to the simplicity in not needing to switch between modalities. The results also suggest that certain modalities might be better suited to specific types of information, for example in the laboratory study gesture was preferred for entering object information and speech was preferred for entering feature information.

The observable nature of the modality combinations was also shown to have an effect on modality use in private and public environments, with a significant shift in preference towards the non-observable modality combinations (GIGI, HGI) in a public environment and a near-significant shift in preference towards the observable modality combinations (SGE, SS) in a private environment. Related to this, it was shown that the modalities intra-gesture, extra-gesture, and handwriting (but not speech) are comfortable to use in a public environment, and all modalities are comfortable to use in a private environment. A strong correlation between the modality combinations rated intuitive and non-intuitive in the laboratory study and the real-world study was also shown to exist, with SS, GIGI, SGI, HH, and SGE all being rated significantly intuitive in both studies. These results on modality intuition are expected to be useful for designers of interfaces that need to be learnt quickly by users (e.g. exhibitions in a museum), while the differences between observable and non-observable combinations highlight that interfaces may need to cater for users differently when situated in public and private environments. The qualitative concerns that subjects had regarding modality use can be taken to form a general guideline in determining which base modalities and modality combinations are best to use when designing for mobile users.

Regarding the overlapped modality combinations, it was seen that non-overlapped modality combinations were significantly preferred to overlapped combinations, and one reason for this is almost certainly that subjects did not see the need for these more complex overlapped interactions when the same functionality was offered by easier and similarly robust modality combinations. Scenarios exhibiting a harsher environment (e.g. noisy), or where the user is more vulnerable to making mistakes (e.g. running), or in applications where high levels of accuracy are vital (e.g. entering in credit card details), may however provide differing results, as too would a study focused on system accuracy rather than user preference for different modality combinations. Nonetheless, it was shown that overlapped object information was more preferred than overlapped feature information and that the modalities speech and gesture were better to combine than combinations comprising handwriting.

Finally, it was shown that direct interaction with anthropomorphized objects is accepted and indeed preferred by a majority of subjects, the results of which are significant for men. The product type (e.g. cosmetics, electronics, automotive), relationship to a product (e.g. buyer, owner), and gender (male, female) were also shown to have an effect on a subject's preference for direct interaction with anthropomorphized objects. These findings have already been exploited in two other projects in which interactive installations for museums and theme parks are being developed (Ndiaye et al., 2005).

This chapter outlines the scientific and practical contributions as well as the commercial significance of this dissertation on multimodal interaction for mobile users. The chapter concludes with a discussion of the opportunities for future research that have come to exist based on the findings presented in this work.

7.1 Scientific Contributions

Previous approaches in dealing with multimodal interaction often dealt with a very limited number of modalities and/or modality combinations, and in fact many focused on extending existing unimodal spoken dialogue systems to account only for complementary spoken-deictic gesture interactions. Previous approaches rarely dealt with semantically overlapped input, nor did these approaches appreciate the importance in re-weighting confidence values to remove recognizer biases. Past work has also been unable to incorporate modern interaction techniques well-suited to mobile users such as tangible interaction. Mobile contexts were also very rarely addressed, despite these contexts having an abundance to gain from flexible and natural multimodal interaction and modality fusion. In addition, previous approaches most often opted for distributed architectures based on powerful but stationary computers, rather than embedded mobile computing devices like PDAs, which the user can carry in his or her pocket and access at will and in any environment. Past approaches were as a result rarely able to conduct applicable usability studies for multimodal interaction suited to mobile contexts like shopping and navigation.

The following scientific achievements resulting from the work conducted in this dissertation are worth highlighting:

- **A formal classification for multimodal input:** Past work has tried to categorize the many different types of multimodal input, but these definitions are only partial and mostly used only to express the possibilities in one particular system. In section 4.2, this dissertation provides a formal classification for multimodal input based on time, semantics, and source origin. The classification covers unimodal and multimodal input, overlapped and non-overlapped input, and the individual modality combinations that can arise from fusing base modalities.
- **A flexible modality fusion architecture:** Designed with the goal to demonstrate modality fusion for a wide range of modalities and modality combinations, the MSA/BPN supports

recognizers currently in use as well as new recognizers that may be added in the future. This is achieved by the mobile device's modality fusion blackboard, to which recognizers can write interaction events.

- **An algorithm to re-weight confidence values:** To support both existing and future interaction devices, an approach was designed to account for biases in individual recognizer confidence values, based on the analysis of accumulated user input data. Accounting for biases in recognizer accuracy is a fundamental requirement for multimodal systems that are capable of fusing semantically overlapped information. The number of such systems is expected to increase, particularly for instrumented environments where multiple same-type recognizers will provide many benefits for minimum extra investment.
- **A method for resolving conflicts in semantically overlapped input:** This work demonstrates how certainty factors and N-best lists can be effectively used to resolve conflicts between semantically overlapped elements within a communication act. Similar to the above point on re-weighting confidence values, such a method is vital for multimodal systems that expect to be able to interpret semantically overlapped input from same-type and different-type recognition devices.

In addition to the contributions to the theory of multimodal interaction and modality fusion, a primary highlight of this work has also been the creation and analysis of new interaction metaphors, which is further supported by real-world usability field studies.

- **Multimodal interface design to support the limited resources of mobile-devices:** Limitations inherent in mobile devices can affect the design and implementation of effective multimodal systems. Such limitations include the lack of commercially available software packages, constraints relating to the form-factor of the device (e.g. limited display space and no keyboard), and computational constraints (e.g. processing power and working memory). Despite the restrictions that are accompanied with stand-alone applications designed for embedded devices like PDAs, user feedback collected during field-studies confirms that the multimodal systems designed as part of this work provide a powerfully expressive, robust, efficient, and intuitive interface for mobile navigation and shopping.
- **Analysis of supplementary and complementary input:** This work has surpassed past work with regards to the wide range of modality input combinations that have been intensively analysed. These combinations, totalling 23 in all, encompassed supplementary and complementary input as well as semantically overlapped input captured from multiple recognizers. Such analysis was made possible by carefully designing all communication modes to be equally powerful; a task that itself required the implementation of novel interaction metaphors such as the visual What-Can-I-Say scrolling text bar. This wide variety of input combinations can be seen to help cater for individual user preferences, and dynamically changing environment characteristics including privacy and background noise.
- **Analysis of tangible real-world interaction:** In addition to the flexible and natural communication modes speech and handwriting, this work has focussed on the inclusion of tangible interaction with the real physical world, through the design of pickup, putdown, and point actions that are well suited for proximal and distal interaction with real-world objects.

- **Analysis of anthropomorphized objects:** This work has created a reusable platform for analysing the concept of anthropomorphism. Anthropomorphized objects were analysed as a means to modify multimodal interaction from both an input and output perspective. Aside from modifying the person tense to create the notion of anthropomorphism, objects were given individual voices and thus personalities with which to communicate. The use of anthropomorphized objects was analysed in real-world environments for a range of different product types, and results from usability field studies can be seen to strengthen the case for anthropomorphism, which has been shown to contrast to previous beliefs of several leading researchers.
- **A strategy for selecting modalities for presentation output:** In addition to the research that has been conducted on multimodal input, this work also contributes to multimodal output presentation, by providing a strategy in which output modalities can be easily and flexibly selected, and in which user- and system-defined templates can be specified for the presentation of output, e.g. based on context or mimic. The presentation of output also addressed synchronization issues between modalities and the formatting of output to suit the inclusion of different semantic constituents (e.g. feature, object, value). Furthermore, output presentation was designed to account not just for on-device interaction but also for off-device interaction. This is supported through the use of public devices typically situated in an instrumented environment such as projected displays, plasma displays, and public audio.
- **An extension to the concept of symmetric multimodality:** This work demonstrated that the modalities used to present output can be selected not just based on the modalities used for providing input, but rather also on the specific semantic constituents that they are to present. Such an extension provides enhanced solutions in the case where privacy may be a requirement for certain types of information.
- **Usability field studies conducted under realistic environment conditions:** The real-world analysis of different interaction metaphors, as described above, was achieved through a series of usability field studies. The main study focused on user preference and modality intuition for a wide range of modality input combinations, both unimodal and multimodal, and semantically overlapped. To the author's best knowledge, with 23 different modality combinations, this study encompasses one of the largest sets of input combinations analysed. Additional usability field studies were also conducted on modality accuracy and efficiency as well as user acceptance for anthropomorphized objects.
- **Implementation of a multimodal demonstrator:** The mobile pedestrian navigation and shopping applications that were designed and implemented as part of this dissertation form a multimodal demonstrator that is capable of presenting rich interaction possibilities and the concepts of anthropomorphization and symmetric multimodality. This demonstrator has since become a critical foundation for continuing research into intelligent user interface design as outlined in the next section.

7.2 Practical Contributions

The multimodal demonstrator described in this dissertation has become a solid working platform for continuing research on Human-Computer Interaction (HCI) and Intelligent User Interfaces (IUI). The MSA/BPN code-base has been reused in the project COMPASS (Aslan et al., 2005), an initiative that will see the creation of a mobile, multimodal, and multilingual tourist guide for visitors of the Beijing 2008 Olympic Games. The user interactions that can be recorded by the MSA/BPN have also become a knowledge source for the SPECTER-Light project (Schneider et al., 2006), where an open personal memory (Kröner et al., 2006) is used to record a history of a user's experiences in order to deliver ad-hoc and subsequent context dependent user support. The MSA/BPN demonstrator is also being used for ongoing research into interaction in instrumented environments, where digital projectors (Spasova et al., 2005), spatial audio (Schmitz & Butz, 2006), migrating characters (Kruppa, 2006), instrumented shopping trolleys (Schneider, 2003), and ubiquitous user models (Heckmann, 2005) are being used to further enhance communication with users. The MSA/BPN has furthermore provided the foundations for several bachelor and master theses dealing with topics such as usability issues with mobile systems using a mobile eye tracker (Norlien, 2002), the generation of environment descriptions based on auditive perception (Maier, 2005), anthropomorphized objects (Schmidt, 2005), modality output planning¹, and embedded Augmented Reality (AR) tag recognition for mobile devices².

7.3 Commercial Significance and Contributions to Public Awareness

The multimodal demonstrator has provided an extensive code-base for the reuse of communication modes and sensor input, modality fusion strategies, the blackboard architecture, and supporting databases. The contributions of this work do not however end here. The demonstrator has, for example, been shown to the public at a variety of exhibitions (see below) and to a number of prominent people including the ambassador of Australia to the Federal Republic of Germany and the ambassador of the United States of America to the Federal Republic of Germany. The demonstrator has at differing stages of development also been shown to commercial companies with the aim of helping to shorten the path that language technology requires to move from the laboratory into realistic applications³. Some of the companies that the multimodal demonstrator has been shown to include: Fuji Research, ScanSoft, Siemens, BMW, T-Systems, Fraunhofer ISST, NTT DoCoMo, and most recently the METRO Group Future Store Initiative. In addition, the system has been broadcast on television⁴ and illustrated in the newspaper⁵. Finally, aspects of the work have been published in books, journals, and leading international conferences and workshops, as also outlined below:

- **Exhibitions:** The following is a list of exhibitions at which the MSA/BPN demonstrator has been shown to the public (sorted by date).
 - Saarland University Open Day: 01.07.2006.

¹A current bachelor thesis: 'Planung der Ausgabe des Mobile ShopAssist'

²A current bachelor thesis: 'Kamera und Marker basierte Interaktion auf Mobilien Geräten'

³This is the primary goal of the COLLATE project as outlined at <http://collate.dfki.de>

⁴SR-1 TV Reportage on Zentrum für Sprachforschung, Aktueller Bericht, 17.06.2005. 'Das sprechende Regal'

⁵Computer Zeitung, Brennpunkt: Cebit - Leben und Arbeiten in der digitalen Welt, pg. 12, 07.03.2005

- CeBIT Fair, Hannover: 09.03.2006 - 15.03.2006.
- Voice Day, Bonn: 20.10.2005 - 21.10.2005.
- CeBIT Fair, Hannover: 10.03.2005 -16.03.2005.
- Empower Germany, Saarbrücken: 30.06.2004.
- DFKI Language Technology Summit, Saarbrücken: 11.05.2004.
- **Journal and book chapters:** Contributions have been made to the book “True Visions” by Aarts and Encarnação (2006) and to the special journal issue on “Conversational User Interfaces” by Wahlster (2004).
- **International conferences:** Parts of this work have been presented at international conferences on intelligent environments (IE 2006 in Athens), intelligent user interfaces (IUI 2006 in Sydney and IUI 2004 in Madeira), mixed and augmented reality (ISMAR 2005 in Vienna), pervasive computing (Pervasive 2005 in Munich), human computer interaction with mobile devices (Mobile HCI 2003 in Udine), and speech communication and technology (Eurospeech 2003 in Geneva).
- **International workshops:** Parts of this work have also been presented at international workshops on user-experience design for pervasive computing, artificial intelligence in mobile systems, invisible and transparent interfaces, multi-user and ubiquitous user interfaces, adaptivity and user modelling in interactive software systems, physical interaction, personalization for the mobile world, and intelligent situation-aware media and presentations.

7.4 Opportunities for Further Research

The previous sections have outlined the scientific, practical, and commercial gains that have already been realized through this dissertation. In the current section, future research directions that have become discernable through this work are highlighted, and this is followed with possible future extensions to the work outlined in this dissertation.

- **Multimodal communication:** This work has highlighted that mobile users require new interaction paradigms. Tangible interaction is one such paradigm that shows great potential and could be extended to include more than the selection-gestures that were outlined for the MSA/BPN. In addition to the linguistic communication modes speech and handwriting, non-linguistic communicative modes like facial expression, eye, hand, body movements, and emotion are another largely untapped source of communication that should suit well to multimodal systems, particularly as a form of additional passive input. Supporting such communication is the need for new types of device and new methods for interpreting sensor data (e.g. biosensors). The MSA/BPN is capable of interpreting input from multiple same-type recognizers, and this would be another interesting field of study, particularly with respect to natural language processing, which is not readily available for embedded devices but would, as part of a distributed architecture, complement the finite-state grammars used in the MSA/BPN.
- **Multi-user and multiparty interaction:** One area that this work has briefly described is that of multi-user interaction, in which multiple users can interact with the same set of devices at the same time (e.g. real-world shelves and shopping products in the MSA scenario).

Future research might like to consider multiparty interaction, in which parties of people can collaborate in everyday tasks like shopping and pedestrian navigation. An example group that would benefit from multiparty designs are families. Families generally consist of parents and children, both of which can have different goals and objectives when partaking in the same task. Designing for multiple parties will require interaction paradigms to be adapted to the specific individuals, and additional functionality such as the ability to communicate and locate other people in the party will also need to be addressed.

- **Multimodal interaction with mobile referents:** One natural path for future research following on from this dissertation on mobile multimodal interaction is that of multimodal interaction where not the user is mobile but rather the referents. From a commercial standpoint, such interaction would suit well to ubiquitous advertising. Interaction could as a first step take place with stationary real-world billboards and later on with moving objects visible within digital video streams (e.g. movies and soap operas). Interaction could even be extended to encompass moving real-world objects such as nearby people and the clothes and accessories that they are currently wearing.
- **Usability studies:** A variety of additional usability studies would also benefit the design of future state-of-the-art multimodal systems. User preference for individual modality combinations could be conducted for mobile applications in different environment contexts and for different user groups (e.g. children, adults, and the elderly). These studies could also be extended to research how modality characteristics like accuracy and efficiency are affected in such contexts. Another topic requiring further research is that of anthropomorphization, for which studies could focus on a broader set of product types and user groups than that discussed in this work. Such studies could also focus on user acceptance for cross-selling (and up-selling) and user acceptance for anthropomorphized object personalities that are based on a particular product (or product type) and user (or user group).

Several topics relating to the specific implementations described in this work, namely application scalability and modality fusion, also provide ground for interesting future research.

- **Application scalability:** Domains like shopping and navigation can contain very large numbers of objects. Although a realistic implementation for interaction with the product type ‘digital cameras’ was achieved in this work, this product type represents only one from a possible many thousands of products and product types that a store might contain. As a result, access to a realistic database of objects would be a crucial next step to testing aspects like the accuracy of the system. One factor influencing accuracy is the size of the associated grammars. This issue has largely been avoided in the current work through the use of a dialog management strategy that requests the user to select one type of object from all of those that are otherwise available. A more flexible solution might however be to incorporate user-positioning technology (Brandherm & Schwartz, 2005) to identify relevant object types nearby to the user’s current location. Incorporating the ability to interact with multiple product types would then also permit the use of information on partial semantic overlap, as described in section 5.3.5.

The grammars that are currently used in the MSA/BPN have been handcrafted for each product type. This is acceptable when many products all have the same attributes (e.g. digital cameras), but is less acceptable when many different product types exist, as is the

case when modelling the entire product range of a store. One solution to this would be to automatically generate the grammars based on keywords available in the database and based on the type of questions that users may ask (e.g. wh-questions and yn-questions).

- **Modality fusion:** The modality fusion component outlined in chapter 5 contained several simplifications that could be extended upon in future work. As described in section 5.3.3.1, only a means to collect user feedback on recognition accuracy has been implemented. Machine learning principles could however be applied such that user feedback collected during runtime is automatically integrated into future confidence weightings. Accuracy data has also only been accumulated based on each recognizer, each modality, and each semantic constituent (e.g. feature and object). An extension to this would be to define accuracy in terms of each semantic constituent's value (e.g. 'price' and 'PowerShot S50'). As described in section 5.3.6, the MSA/BPN also only accounts for the fusion of two semantically overlapped elements. Certainty factors can however account for more than two input streams and this would then better permit for the use of even more communication modes, including the use of multiple same-type recognizers. Finally, now that usability field studies have been conducted to identify accuracy rates based on the individual recognizers, it would also be useful to conduct studies that determine by how much the use of this information has improved the recognition of multimodal interaction in the MSA/BPN.

- Aarts, E., & Encarnação, J. L. (Eds.). (2006). *True Visions: The Emergence of Ambient Intelligence*. Berlin, Heidelberg, New York: Springer.
- Abowd, G. D., Atkeson, C. G., Hong, J., Long, S., Kooper, R., & Pinkerton, M. (1997). Cyberguide: A Mobile Context-Aware Tour Guide. *Wireless Networks*, 3(5), 421–433.
- Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, 16, 3–9.
- Adams, D. (Ed.). (1979). *The Hitchhiker's Guide to the Galaxy*. London, UK: MacMillan.
- Alexandersson, J., & Becker, T. (2003). The Formal Foundations Underlying Overlay. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS)* (pp. 22–36). Tilburg, The Netherlands.
- Alexandersson, J., Becker, T., Engel, R., Löckelt, M., Pecourt, E., Poller, P., Pflieger, N., & Reithinger, N. (2004). Ends-based Dialogue Processing. In R. Porzel (Ed.), *Proceedings of the 2nd Workshop on Scalable Natural Language Understanding* (pp. 25–32). Boston, MA, USA: Association for Computational Linguistics.
- Allen, J. (1995). *Natural Language Understanding* (2nd ed.). Redwood City, CA: Benjamin-Cummings.
- Almeida, L., Amdal, I., Beires, N., Boualem, M., Boves, L., Os, E. den, Filoche, P., Gomes, R., Knudsen, J. E., Kvale, K., Rugelbak, J., Tallec, C., & Warakagoda, N. (2002). Implementing and evaluating a multimodal and multilingual tourist guide. In *Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*. (7 pp.). Copenhagen, Denmark.
- Alshawi, H., Boguraev, B., Bird, S., Hindle, D., Kay, M., McDonald, D., Uszkoreit, H., & Wilks, Y. (1987). *Memory and Context for Language Interpretation (Studies in Natural Language Processing)*. New York, NY, USA: Cambridge University Press.
- André, E. (2003). Natural language in multimedia/multimodal systems. In R. Mitkov (Ed.), *Handbook of computational linguistics* (pp. 650–669). Oxford University Press.
- André, E., Finkler, W., Graf, W., Rist, T., Shauder, A., & Wahlster, W. (1993). WIP: The Automatic Synthesis of Multimodal Presentations. 75–93.

- André, E., & Rist, T. (2001). Controlling the Behavior of Animated Presentation Agents in the Interface: Scripting versus Instructing. *AI Magazine*, 22(4), 53–66.
- Archambault, D., & Burger, D. (2001). From Multimodality to Multimodalities: The Need for Independent Models. In C. Stephanidis (Ed.), *Proceedings of the UAHCI Conference on Universal Access in HCI - Towards an Information Society for All* (pp. 227–231). Mahwah, NJ, USA: Lawrence Erlbaum Associates.
- Aslan, I., Xu, F., Uszkoreit, H., Krüger, A., & Steffen, J. (2005). COMPASS2008: Multimodal, Multilingual and Crosslingual Interaction for Mobile Tourist Guide Applications. In *Proceedings of Intelligent Technologies for Interactive Entertainment (INTETAIN)* (pp. 3–12). Madonna di Campiglio, Italy.
- Asthana, A., Cravatts, M., & Krzyzanowski, P. (1994). An Indoor Wireless System for Personalized Shopping Assistance. In *IEEE Workshop on Mobile Computing Systems and Applications* (pp. 69–74). Santa Cruz, CA, US: IEEE CS Press.
- Azuma, R., Bimber, O., & Sato, K. (Eds.). (2005). *ISMAR 2005: Mixed and Augmented Reality: 4th IEEE/ACM International Symposium, Vienna, Austria, 5-8 October, 2005, Proceedings*. IEEE Computer Society.
- Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., & Fischer, K. (2000). The Recognition of Emotion. In W. Wahlster (Ed.), *VerbMobil: Foundations of Speech-to-Speech Translation* (pp. 122–130). Berlin, Heidelberg, New York: Springer.
- Baudel, T., & Beaudouin-Lafon, M. (1993). CHARADE: Remote Control of Objects Using Free-Hand Gestures. *Communications of the ACM*, 36(7), 28–35.
- Bauer, M. (1996). *Ein evidenztheoretischer Ansatz zur Planerkennung*. Unpublished doctoral dissertation, Department of Computer Science, University of Saarland.
- Baus, J., Cheverst, K., & Kray, C. (2005). A Survey of Map-based Mobile Guides. In L. Meng & A. Zipf (Eds.), *Map-based mobile services: Theories, Methods and Implementations* (pp. 197–216). Berlin, Heidelberg, New York: Springer.
- Baus, J., Krüger, A., & Wahlster, W. (2002). A Resource-Adaptive Mobile Navigation System. In *Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI)* (pp. 15–22). New York, NY, USA: ACM.
- Becker, T., Blaylock, N., Gerstenberger, C., Kruijff-Korbayová, I., Korthauer, A., Pinkal, M., Pitz, M., Poller, P., & Schehl, J. (2006). Natural and Intuitive Multimodal Dialogue for In-Car Applications: The SAMMIE System. In G. Brewka, S. Coradeschi, A. Perini, & P. Traverso (Eds.), *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI)* (pp. 612–616). Amsterdam, The Netherlands: IOS Press.
- Beigl, M., Intille, S. S., Rekimoto, J., & Tokuda, H. (Eds.). (2005). *UbiComp 2005: Ubiquitous Computing: 7th International Conference, Tokyo, Japan, September 11-14, 2005, Proceedings*. Berlin, Heidelberg, New York: Springer.

- Bernstein, L. E., & Benoit, C. (1996). For Speech Perception by Humans or Machines, Three Senses are Better than One. In T. Bunnell & W. Idsardi (Eds.), *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)* (Vol. 3, pp. 1477–1480). Philadelphia, PA, USA.
- Bühler, D., Minker, W., Häußler, J., & Krüger, S. (2002). Flexible Multimodal Human-Machine Interaction in Mobile Environments. In *Proceedings of the International Conference on Speech and Language Processing (ICSLP)* (pp. 169–172). Denver, CO, USA.
- Bohnenberger, T. (Ed.). (2005). *Decision-Theoretic Planning for User-Adaptive Systems: Dealing With Multiple Goals and Resource Limitations (DISKI 289)*. Berlin, Germany: Akademische Verlagsgesellschaft.
- Bohnenberger, T., Jameson, A., Krüger, A., & Butz, A. (2002). Location-Aware Shopping Assistance: Evaluation of a Decision-Theoretic Approach. In *Proceedings of the 4th International Symposium on Mobile Human-Computer Interaction (Mobile HCI)* (pp. 155–169). Pisa, Italy.
- Bolt, R. A. (1979). *Spatial Data Management*. (DARPA Report. MIT Architecture Machine Group, Cambridge, MA, USA)
- Bolt, R. A. (1980). Put-that-there: Voice and Gesture at the Graphics Interface. *SIGGRAPH Computer Graphics*, 14(3), 262–270.
- Boves, L., Neumann, A., Vuurpijl, L., Bosch, L. ten, Rossignol, S., Engel, R., & Pfleger, N. (2004). Multimodal Interaction in Architectural Design Applications. In *Proceedings of the 8th ERCIM Workshop on User Interfaces for All (UI4All)* (pp. 384–390). Vienna, Austria.
- Boves, L., & Os, E. den. (2002). *Multimodal Services- a MUST for UMTS*. (EURESCOM Project Results: Multimodal Multilingual Information Services for Small Mobile Terminals (MUST))
- Brandherm, B., & Jameson, A. (2004). An Extension of the Differential Approach for Bayesian Network Inference to Dynamic Bayesian Networks. *International Journal of Intelligent Systems*, 19(8), 727–748.
- Brandherm, B., & Schwartz, T. (2005). Geo Referenced Dynamic Bayesian Networks for User Positioning on Mobile Systems. In T. Strang & C. Linnhoff-Popien (Eds.), *Proceedings of the International Workshop on Location- and Context-Awareness (LoCA)* (pp. 223–234). Berlin, Heidelberg, New York: Springer.
- Bregler, C., Manke, S., Hild, H., & Waibel, A. (1993). Improving Connected Letter Recognition by Lipreading. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vol. 1, pp. 557–560). Minneapolis, MN, USA: IEEE Press.
- Butz, A., & Schmitz, M. (2005). Design and Applications of a Beer Mat for Pub Interaction. In *Poster at the 7th International Conference on Ubiquitous Computing (Ubicomp)*. (3 pp.). Tokyo, Japan.

- Butz, A., Schneider, M., & Spassova, M. (2004). SearchLight: A Lightweight Search Function for Pervasive Environments. In *Proceedings of the 2nd International Conference on Pervasive Computing (Pervasive)* (pp. 351–356). Berlin, Heidelberg, New York: Springer.
- Caelen, J. (1994). Multimodal Human-Computer Interface. In E. Keller (Ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State-of-the-Art and Future Challenges* (pp. 339–373). Chichester, UK: John Wiley and Sons.
- Carpenter, B. (1992). *The Logic of Typed Feature Structures*. Cambridge: Cambridge University Press.
- Catizone, R., Setzer, A., & Wilks, Y. (2003). Multimodal Dialogue Management in the COMIC Project. In *Proceedings of the EACL 2003 Workshop on Dialogue Systems: Interaction, Adaptation, and Styles of Management*. (10 pp.). Budapest, Hungary.
- Chai, J., Horvath, V., Kambhatla, N., Nicolov, N., & Stys-Budzikowska, M. (2001). A Conversational Interface for Online Shopping. In *Proceedings of the 1st International Conference on Human Language Technology Research (HLT)* (pp. 1–4). Morristown, NJ, USA: Association for Computational Linguistics.
- Chai, J. Y., Hong, P., & Zhou, M. X. (2004). A Probabilistic Approach to Reference Resolution in Multimodal User Interfaces. In *Proceedings of the 9th International Conference on Intelligent User Interface (IUI)* (pp. 70–77). New York, NY, USA: ACM.
- Charwat, H. J. (Ed.). (1992). *Lexikon der Mensch-Maschine-Kommunikation*. München: Oldenbourg.
- Chellapilla, K., Larson, K., Simard, P. Y., & Czerwinski, M. (2005). Computers beat Humans at Single Character Recognition in Reading based Human Interaction Proofs (HIPs). In *Second Conference on Email and Anti-Spam (CEAS)*. (8 pp.). California, CA, USA.
- Cheverst, K., Davies, N., Mitchell, K., Friday, A., & Efstratiou, C. (2000). Developing a Context-aware Electronic Tourist Guide: Some Issues and Experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 17–24). New York, NY, USA: ACM.
- Chittaro, L. (Ed.). (2003). *Mobile HCI 2003: Human-Computer Interaction with Mobile Devices and Services: 5th International Symposium, Udine, Italy, September 8-11, 2003, Proceedings*. Berlin, Heidelberg, New York: Springer.
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3), 113–124.
- Clarke, A. C., & Kubrick, S. (Eds.). (1993). *2001: A Space Odyssey*. London, UK: Penguin.
- Cohen, A. A. (1977). The Communicative Functions of Hand Illustrators. *Journal of Communication*, 27(4), 54–63.
- Cohen, M. H., Giangola, J. P., & Balogh, J. (Eds.). (2004). *Voice User Interface Design*. Addison-Wesley. (Chapter 5: High-Level Design Elements)

- Cohen, P. R., Cheyer, A., Wang, M., & Baeg, S. C. (1998). An Open Agent Architecture. In M. N. Huhns & M. P. Singh (Eds.), *Readings in Agents* (pp. 197–204). San Francisco, CA, USA: Morgan Kaufmann.
- Cohen, P. R., Coulston, R., & Krout, K. (2002). Multimodal Interaction During Multiparty Dialogues: Initial Results. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI)* (pp. 448–453). Pittsburgh, PA, USA: IEEE Computer Society.
- Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., & Clow, J. (1997). Quickset: Multimodal interaction for distributed applications. In *Proceedings of the 5th ACM International Conference on Multimedia* (pp. 31–40). New York, NY, USA: ACM.
- Cohen, P. R., & McGee, D. R. (2004). Tangible Multimodal Interfaces for Safety-Critical Applications. *Communications of the ACM*, 47(1), 41–46.
- Cole, R., Mariani, J., Uszkoreit, H., Varile, G. B., Zaenen, A., Zampolli, A., Boguraev, B., Bird, S., Hindle, D., Kay, M., McDonald, D., & Wilks, Y. (Eds.). (1998). *Survey of the State of the Art in Human Language Technology*. Cambridge, UK: Cambridge University Press.
- Corradini, A., & Cohen, P. R. (2002). Multimodal Speech-Gesture Interface for Handfree Painting on a Virtual Paper Using Partial Recurrent Neural Networks as Gesture Recognizer. In *Proceedings of the International Joint Conference on Artificial Neural Networks (IJCNN)* (Vol. 3, pp. 2293–2298). Honolulu, HI, USA.
- Corradini, A., Wesson, R. M., & Cohen, P. R. (2002). A Map-Based System Using Speech and 3D Gestures for Pervasive Computing. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI)* (pp. 191–196). Washington, DC, USA: IEEE Computer Society.
- Crampton-Smith, G. (1995). The Hand That Rocks the Cradle. *The International Design Magazine (ID)*, 60–65.
- Dance, F. E. X. (1970). The ‘Concept’ of Communication. *Journal of Communication*, 20(2), 201–210.
- Dance, F. E. X. (Ed.). (1982). *Human Communication Theory: Comparative Essays*. New York, NY, USA: Harper and Row.
- Dengel, A., Hoch, R., Hönes, F., Jäger, T., Malburg, M., & Weigel, A. (1997). Techniques for Improving OCR Results. In P. Wang & H. Bunke (Eds.), *Handbook on Optical Character Recognition and Document Analysis* (pp. 227–258). Singapore: World Scientific.
- Dey, A. K., Ljungstrand, P., & Schmidt, A. (2001). Distributed and Disappearing User Interfaces in Ubiquitous Computing. In *Conference on Human Factors in Computing Systems (CHI)* (pp. 487–488). New York, NY, USA: ACM.
- Don, A., Brennan, S., Laurel, B., & Shneiderman, B. (1992). Anthropomorphism: From Eliza to Terminator 2. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 67–70). New York, NY, USA: ACM.

- Durkin, J. (1994a). *Expert Systems: Design and Development*. Macmillan. (Chapter 12: Certainty Theory)
- Durkin, J. (1994b). *Expert Systems: Design and Development*. Macmillan. (Chapter 5: MYCIN)
- Ebaugh, A., & Chatterjee, S. (2004). *SAVi: Shopping Assistant for the Visually Impaired*. Unpublished master's thesis, Department of Computer Science and Engineering, University of Washington. (Available at: <http://www.cs.washington.edu/homes/sauravc/cs477/SAVi.pdf>)
- Edwards, A. (1988). The Design of Auditory Interfaces for Visually Disabled Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 83–88). New York, NY, USA: ACM.
- Ekman, P., & Friesen, W. V. (1969). The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica, 1*, 49–98.
- Elting, C. (2002). What are Multimodalities made of? Modeling Output in a Multimodal Dialogue System. In *Workshop on Intelligent Situation-Aware Media and Presentations (ISAMP)*. (8 pp.). Edmonton, Alberta, Canada.
- Elting, C., & Michelitsch, G. (2001). A Multimodal Presentation Planner for a Home Entertainment Environment. In *Proceedings of the 2001 Workshop on Perceptive User Interfaces (PUI)* (pp. 1–5). New York, NY, USA: ACM.
- Elting, C., Rapp, S., Möhler, G., & Strube, M. (2003). Architecture and Implementation of Multimodal Plug and Play. In *Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI)* (pp. 93–100). New York, NY, USA: ACM.
- Falk, P., & Campbell, C. (Eds.). (1997). *The Shopping Experience*. London, UK: SAGE Publications.
- Feld, M. (2006). *Erzeugung von Sprecherklassifikationsmodulen für Multiple Plattformen*. Unpublished master's thesis, Department of Informatics, University of Saarland.
- Fensel, D., Hendler, J. A., Lieberman, H., & Wahlster, W. (Eds.). (2003). *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. Cambridge, MA, USA: MIT Press.
- Fensel, D., Horrocks, I., Van Harmelen, F., Decker, S., Erdmann, M., & Klein, M. (2000). OIL in a nutshell. In R. Dieng & O. Corby (Eds.), *Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management (EKAW)* (pp. 1–16). Berlin, Heidelberg, New York: Springer.
- Fitzmaurice, G. W., Ishii, H., & Buxton, W. A. S. (1995). Bricks: Laying the Foundations for Graspable User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 442–449). New York, NY, USA: ACM/Addison-Wesley.
- Geiser, G. (Ed.). (1990). *Mensch-Maschine Kommunikation*. München: Oldenbourg.
- Gellersen, H.-W., Want, R., & Schmidt, A. (Eds.). (2005). *Pervasive 2005: Pervasive Computing: 3rd International Conference, Munich, Germany, May 8-13, 2005, Proceedings*. Berlin, Heidelberg, New York: Springer.

- Goddeau, D., Brill, E., Glass, J., Pao, C., Phillips, M., Polifroni, J., Sene, S., & Zue, V. (1994). Galaxy: A Human-language Interface to On-line Travel Information. In *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP)* (pp. 707–710). Yokohama, Japan.
- Gorrell, G. (2004). *Language Modelling and Error Handling in Spoken Dialogue Systems*. Unpublished master's thesis, Department of Science and Technology, Linköping University, Sweden.
- Grice, H. P. (1975). Logic and Conversation. *Syntax and Semantics, 3:Speech Acts*, 41–58.
- Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition, 5*(2), 199–220.
- Gundel, J. K. (2003). Information Structure and Referential Givenness/Newness: How Much Belongs in the Grammar? In *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar* (pp. 122–142). Stanford, CA, USA: CSLI Publications.
- Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive Status and the Form of Referring Expressions in Discourse. *Language, 69*(2), 274–307.
- Gupta, A. (2003). A Reference Model for Multimodal Input Interpretation. In *CHI 2003 Extended Abstracts on Human Factors in Computing Systems (CHI)* (pp. 936–937). New York, NY, USA: ACM.
- Gurevych, I., Merten, S., & Porzel, R. (2003). Automatic Creation of Interface Specifications from Ontologies. In *Proceedings of the HLT-NAACL Workshop on the Software Engineering and Architecture of Language Technology Systems (SEALTS)* (pp. 60–67). Edmonton, Canada.
- Hagras, H., & Callaghan, V. (Eds.). (2005). *IE 2005: Intelligent Environments: IEE International Workshop, Colchester, UK, June 28-29, 2005, Proceedings*. The University of Essex.
- Halliday, T. (Ed.). (1998). *The Senses and Communication*. Berlin, Heidelberg, New York: Springer.
- Heckmann, D. (2005). *Ubiquitous User Modeling*. Unpublished doctoral dissertation, Department of Computer Science, University of Saarland. (URL: <http://w5.cs.uni-sb.de/publication/file/178/Heckmann05Diss.pdf>)
- Hendler, J., & McGuinness, D. (2000). The DARPA Agent Markup Language. *IEEE Intelligent Systems Trends and Controversies, 15*(6), 67–73.
- Herzog, G., Kirchmann, H., Merten, S., Ndiaye, A., & Poller, P. (2003). MULTIPLATFORM Testbed: An Integration Platform for Multimodal Dialog Systems. In *HLT-NAACL Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS)* (pp. 75–82). Edmonton, Canada.
- Herzog, G., Ndiaye, A., Merten, S., Kirchmann, H., Becker, T., & Poller, P. (2004). Large-scale software integration for spoken language and multimodal dialog systems. *Natural Language Engineering, 10*(3-4), 283–305.

- Hill, T., & Lewicki, P. (2006). *Statistics: Methods and Applications. A Comprehensive Reference for Science, Industry, and Data Mining* (1st ed.). Tulsa, OK, USA: StatSoft.
- Holtkamp, B., Gartmann, R., & Han, Y. (2003). FLAME 2008: Personalized Web Services for the Olympic Games 2008 in Beijing. In P. Cunningham, M. Cunningham, & P. Fatelnig (Eds.), *Building the Knowledge Economy: Issues, Applications, Case Studies* (pp. 93–99). Amsterdam, The Netherlands: IOS Press.
- Huls, C., Claassen, W., & Bos, E. (1995). Automatic Referent Resolution of Deictic and Anaphoric Expressions. *Computational Linguistics*, 21(1), 59–79.
- Hura, S. L., & Owens, R. (2003). *The Truth about Multimodal Interaction*. (White paper, Inter-voice, URL: <http://www.microsoft.com/speech/executive/speechtech/whitepapers/>)
- Ishii, H., & Ullmer, B. (1997). Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 234–241). New York, NY, USA: ACM.
- Jameson, A. (2002). Usability Issues and Methods for Mobile Multimodal Systems. In *Proceedings of the ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments* (p. 1). Kloster Irsee, Germany.
- Jan Alexandersson, N. P. (2006). Discourse Modelling. In W. Wahlster (Ed.), *SmartKom. Foundations of Multimodal Dialogue Systems* (pp. 237–254). Berlin, Heidelberg, New York: Springer.
- Johanson, B., & Fox, A. (2002). The Event Heap: A Coordination Infrastructure for Interactive Workspaces. In *Proceedings of the 4th IEEE Workshop on Mobile Computing Systems and Applications (WMCSA)* (pp. 83–93). Washington, DC, USA: IEEE Computer Society.
- Johnston, M. (1998). Unification-based Multimodal Parsing. In *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics* (pp. 624–630). Morristown, NJ, USA: Association for Computational Linguistics.
- Johnston, M., Bangalore, S., Stent, A., Vasireddy, G., & Ehlen, P. (2002). Multimodal Language Processing for Mobile Information Access. In *Proceedings of the International Conference on Speech and Language Processing (ICSLP)* (pp. 2237–2240). Denver, CO, USA.
- Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., & Maloor, P. (2002). MATCH: An Architecture for Multimodal Dialogue Systems. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 376–383). Morristown, NJ, USA: Association for Computational Linguistics.
- Johnston, M., Cohen, P. R., McGee, D., Oviatt, S. L., Pittman, J. A., & Smith, I. (1997). Unification-based Multimodal Integration. In *Proceedings of the 8th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 281–288). Morristown, NJ, USA: Association for Computational Linguistics.
- Kaiser, E., Olwal, A., McGee, D., Benko, H., Corradini, A., Li, X., Cohen, P., & Feiner, S. (2003). Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality.

- In *Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI)* (pp. 12–19). New York, NY, USA: ACM.
- Kehler, A. (2000). Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction. In *Proceedings of the 17th National Conference on Artificial Intelligence and the 12th Conference on Innovative Applications of Artificial Intelligence* (pp. 685–690). AAAI Press / The MIT Press.
- Kendon, A., Drew, P., Goodwin, M. H., Gumperz, J. J., & Schiffrin, D. (Eds.). (1990). *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge, UK: Cambridge University Press.
- Kipp, M. (2003). *Gesture Generation by Imitation: From Human Behavior to Computer Character Animation*. Boca Raton, FL, USA: Dissertation.com.
- Kirste, T., Herfet, T., & Schnaider, M. (2001). EMBASSI: Multimodal Assistance for Universal Access to Infotainment and Service Infrastructures. In *Proceedings of the 2001 EC/NSF Workshop on Universal Accessibility of Ubiquitous Computing (WUAUC)* (pp. 41–50). New York, NY, USA: ACM.
- Knapp, K. (2000). Metaphorical and Interactional Uses of Silence. *Erfurt Electronic Studies in English*.
- Kobsa, A., Allgayer, J., Reddig, C., Reithinger, N., Schmauks, D., Harbusch, K., & Wahlster, W. (1986). Combining deictic gestures and natural language for referent identification. In *Proceedings of the 11th Conference on Computational Linguistics* (pp. 356–361). Morristown, NJ, USA: Association for Computational Linguistics.
- Kort, T., Cozza, R., Maita, K., & Tay, L. (2006). *Record 14.9 Million PDAs Shipped in 2005, Up 19 Percent Over 2004*. (Gartner Report, 10 February 2006, http://www.gartner.com/DisplayDocument?ref=g_search&id=488746, last accessed: 22.10.2006)
- Kourouthanassis, P., Koukara, L., Lazaris, C., & Thiveos, K. (2001). Last-mile supply chain management: MyGrocer innovative business and technology framework. In *Proceedings of the 17th International Logistics Congress on Logistics from A to U: Strategies and Applications* (pp. 264–273). Thessaloniki, Greece.
- Kray, C. (Ed.). (2003). *Situated Interaction on Spatial Topics (DISKI 274)*. Berlin, Germany: Akademische Verlagsgesellschaft.
- Kray, C., Elting, C., Laakso, K., & Coors, V. (2003). Presenting Route Instructions on Mobile Devices. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI)* (pp. 117–124). New York, NY, USA: ACM.
- Kray, C., Wasinger, R., & Kortuem, G. (2004). Concepts and Issues in Interfaces for Multiple Users and Multiple Devices. In *Workshop on Multi-User and Ubiquitous User Interfaces (MU3I) at IUI/CADUI* (pp. 7–12). Madeira, Portugal.

- Kröner, A., Heckmann, D., & Wahlster, W. (2006). SPECTER: Building, Exploiting, and Sharing Augmented Memories. In *Proceedings of the Workshop on Knowledge Sharing for Everyday Life (KSEL)* (pp. 9–16). Kyoto, Japan.
- Krüger, A., Butz, A., Müller, C., Stahl, C., Wasinger, R., Steinberg, K.-E., & Dirschl, A. (2004). The Connected User Interface: Realizing a Personal Situated Navigation Service. In *Proceedings of the 9th International Conference on Intelligent User Interface (IUI)* (pp. 161–168). New York, NY, USA: ACM.
- Kruppa, M. (2006). *Migrating Characters: Effective User Guidance in Instrumented Environments*. Unpublished doctoral dissertation, Department of Computer Science, University of Saarland. (URL: http://w5.cs.uni-sb.de/mkruppa/Dissertation_Michael_Kruppa.pdf)
- Kumar, A., Pecourt, E., & Romary, L. (2002). *Dialogue Module Technical Specification*. (Technical report, LORIA, Nancy, France. Project MIAMM: Multidimensional Information Access using Multiple Modalities, EU project. IST-20000-29487, Deliverable D5.1.)
- Kumar, A., & Romary, L. (2003). A Comprehensive Framework for Multimodal Meaning Representation. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS)* (pp. 225–251). Tilburg, Netherlands.
- Kumar, S., Cohen, P. R., & Coulston, R. (2004). Multimodal Interaction under Exerted Conditions in a Natural Field Setting. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (pp. 227–234). New York, NY, USA: ACM.
- Landragin, F., & Romary, L. (2003). Referring to Objects Through Sub-Contexts in Multimodal Human-Computer Interaction. In *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 67–74). AAAI Press / The MIT Press.
- Leathers, D. G. (Ed.). (1997). *Successful Nonverbal Communication: Principles and Applications*. Boston, MA, USA: Allyn and Bacon.
- Lehtonen, T.-K., & Mäenpää, P. (1997). Shopping in the East Centre Mall. In *The Shopping Experience* (pp. 136–165). London, UK.
- Loos, E. E. (2003). *Glossary of Linguistic Terms*. (Online; accessed 22.10.2006, <http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/>. Based on the LinguaLinks Library CD-ROM Version 5.0 published by SIL International in 2003)
- Lyons, K. (2003). Everyday Wearable Computer Use: A Case Study of an Expert User. In *Proceedings of the 5th International Symposium on Human-Computer Interaction with Mobile Devices and Services (Mobile HCI)* (pp. 61–75). Berlin, Heidelberg, New York: Springer.
- Macdonald, L., & Vince, J. (Eds.). (1994). *Interacting with Virtual Environments*. John Wiley and Sons.
- Maier, A. (2005). *Umgebungsbeschreibung anhand auditiver Perzeption zur Unterstützung mobiler Navigation*. Unpublished master's thesis, Department of Computational Linguistics and Phonetics, University of Saarland.

- Malaka, R., & Zipf, A. (2000). Deep map: Challenging IT Research in the Framework of a Tourist Information System. In *Proceedings of the 7th International Congress on Tourism and Communications Technologies in Tourism (ENTER)* (pp. 15–27). Wien, New York: Springer.
- Malenke, M., Bäumlner, M., & Paulus, E. (2000). Speech Recognition Performance Assessment. In W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation* (pp. 583–591). Berlin, Heidelberg, New York: Springer.
- Masui, T., Tsukada, K., & Siio, I. (2004). MouseField: A Simple and Versatile Input Device for Ubiquitous Computing. In *Proceedings of the 6th International Conference on Ubiquitous Computing (UbiComp)* (pp. 319–328). Berlin, Heidelberg, New York: Springer.
- Maybury, M. T., & Wahlster, W. (Eds.). (1998). *Readings in Intelligent User Interfaces*. San Francisco, CA, USA: Morgan Kaufmann.
- McGee, D. R., & Cohen, P. R. (2001). Creating Tangible Interfaces by Augmenting Physical Objects with Multimodal Language. In *Proceedings of the 6th International Conference on Intelligent User Interfaces (IUI)* (pp. 113–119). New York, NY, USA: ACM.
- McGuinness, D. L. (2002). Ontologies come of age. In D. Fensel, J. Hendler, H. Lieberman, & W. Wahlster (Eds.), *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. Cambridge, MA, USA: MIT Press.
- McLeod, A., & Summerfield, Q. (1985). Intermodal Timing Relations and Audio-Visual Speech Recognition by Normal-hearing Adults. *Journal of the Acoustical Society of America*, 77(2), 678–685.
- McLeod, A., & Summerfield, Q. (1987). Quantifying the Contribution of Vision to Speech Perception in Noise. *British Journal of Audiology*, 21(2), 131–141.
- Menczer, F., Street, W. N., Vishwakarma, N., Monge, A. E., & Jakobsson, M. (2002). IntelliShopper: A Proactive, Personal, Private Shopping Assistant. In *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (pp. 1001–1008). New York, NY, USA: ACM.
- Merriam-Webster (Ed.). (1998). *Merriam-Webster's Collegiate Dictionary* (10th ed.). Springfield, MA, USA: Merriam-Webster.
- MIAMM. (2004). *MIAMM Homepage. Multidimensional Information Access using Multiple Modalities*. (IST-2000-29487 European Project. <http://www.miamm.org>. Last accessed 22.10.2006)
- Milanesi, C., Vergne, H. J. D. L., Liang, A., Desai, K., Mitsuyama, N., Nguyen, T. H., Shen, S., & Song, S.-H. (2006). *Market Share: Mobile Terminals by Region, 4Q05 and 2005*. (Gartner Report, 10 March 2006, http://www.gartner.com/DisplayDocument?ref=g_search&id=489846, last accessed: 22.10.2006)
- Minsky, M. (1975). A Framework for Representing Knowledge. In P. H. Winston (Ed.), *The Psychology of Computer Vision* (pp. 211–277). New York: McGraw-Hill.

- Müller-Prove, M. (2002). *Vision and Reality of Hypertext and Graphical User Interfaces*. Unpublished master's thesis, Department of Informatics, University of Hamburg. (Available at: <http://www.mprove.de/diplom/>)
- Myers, B. A., Malkin, R., Bett, M., Waibel, A., Bostwick, B., Miller, R. C., Yang, J., Denecke, M., Seemann, E., Zhu, J., Peck, C. H., Kong, D., Nichols, J., & Scherlis, W. L. (2002). Flexi-modal and Multi-Machine User Interfaces. In *Proceedings of 4th IEEE International Conference on Multimodal Interfaces (ICMI)* (pp. 343–348). Washington, DC, USA: IEEE Computer Society.
- Naur, P. (Ed.). (1963). Revised Report on the Algorithmic Language ALGOL 60. *The Computer Journal*, 9, 349.
- Ndiaye, A., Gebhard, P., Kipp, M., Klesen, M., Schneider, M., & Wahlster, W. (2005). Ambient Intelligence in Edutainment: Tangible Interaction with Life-Like Exhibit Guides. In M. Maybury, O. Stock, & W. Wahlster (Eds.), *Proceedings of the 1st International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN)* (pp. 104–113). Berlin, Heidelberg, New York: Springer.
- Newcomb, E., Pashley, T., & Stasko, J. (2003). Mobile Computing in the Retail Arena. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 337–344). New York, NY, USA: ACM.
- Nigay, L., & Coutaz, J. (1993). A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 172–178). New York, NY, USA: ACM.
- Nijholt, A., Rist, T., & Tuinenbreijer, K. (2004). Lost in Ambient Intelligence? In *Extended Abstracts on Human Factors in Computing Systems (CHI)* (pp. 1725–1726). New York, NY, USA: ACM.
- Norlien, M. (2002). *An Investigation of Usability Issues with Mobile Systems Using a Mobile Eye Tracker*. Unpublished master's thesis, Department of Information Technology, International University in Germany.
- Norman, D. A. (Ed.). (1998). *The Invisible Computer*. Cambridge, Massachusetts: MIT Press.
- Ogden, C. K., & Richards, I. A. (Eds.). (1923). *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism* (8th ed.). New York, NY, USA: Harcourt.
- OIL. (2000). *OIL: Ontology Inference Layer*. (<http://www.ontoknowledge.org/oil/>). Last accessed: 22.10.2006)
- Oppermann, R., & Specht, M. (2000). A Context-Sensitive Nomadic Exhibition Guide. In P. Thomas & H. Gellersen (Eds.), *Second International Symposium on Handheld and Ubiquitous Computing (HUC)* (pp. 127–142). Berlin, Heidelberg, New York: Springer.
- Oviatt, S. (1999). Ten myths of Multimodal Interaction. *Communications of the ACM*, 42(11), 74–81.

- Oviatt, S. (2002). Advances in the Robust Processing of Multimodal Speech and Pen Systems. *Multimodal Interface for Human-Machine Communication*, 203–218. (Series on Machine Perception and Artificial Intelligence)
- Oviatt, S., Coulston, R., & Lunsford, R. (2004). When Do We Interact Multimodally?: Cognitive Load and Multimodal Communication Patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI)* (pp. 129–136). New York, NY, USA: ACM.
- Oviatt, S., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M., & Carmichael, L. (2003). Toward a Theory of Organized Multimodal Integration Patterns during Human-Computer Interaction. In *Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI)* (pp. 44–51). New York, NY, USA: ACM.
- Oviatt, S., & Wahlster, W. (1997). Introduction to This Special Issue on Multimodal Interfaces Multimodal Interfaces. *Human-Computer Interaction*, 12(1 and 2), 1–5.
- Oviatt, S. L. (1996). Multimodal Interfaces for Dynamic Interactive Maps. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 95–102). New York, NY, USA: ACM.
- Oviatt, S. L. (1997). Multimodal Interactive Maps: Designing for Human Performance. *Human-Computer Interaction*, 12(1-2:Multimodal Interfaces), 93–129.
- Oviatt, S. L. (1999). Mutual Disambiguation of Recognition Errors in a Multimodal Architecture. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 576–583). New York, NY, USA: ACM.
- Oviatt, S. L. (2000a). Multimodal Interface Research: A Science Without Borders. In B. Yuan, T. Huang, & X. Tang (Eds.), *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)* (Vol. 3, pp. 1–6). Beijing, China: Chinese Friendship Publishers.
- Oviatt, S. L. (2000b). Multimodal Signal Processing in Naturalistic Noisy Environments. In B. Yuan, T. Huang, & X. Tang (Eds.), *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)* (Vol. 2, pp. 696–699). Beijing, China: Chinese Friendship Publishers.
- Oviatt, S. L. (2000c). Multimodal System Processing in Mobile Environments. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology (UIST)* (pp. 21–30). New York, NY, USA: ACM.
- Oviatt, S. L. (2002). Breaking the Robustness Barrier: Recent Progress on the Design of Robust Multimodal Systems. *Advances in Computers*, 56, 305–341.
- Oviatt, S. L. (2003). Multimodal Interfaces. In J. A. Jacko & A. Sears (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (pp. 286–304). Mahwah, NJ, USA: Lawrence Erlbaum Associates.
- Oviatt, S. L., & Cohen, P. R. (2000). Multimodal Interfaces That Process What Comes Naturally. *Communications of the ACM*, 43(3), 45–53.

- Oviatt, S. L., Cohen, P. R., & Wang, M. (1994). Toward Interface Design for Human Language Technology: Modality and Structure as Determinants of Linguistic Complexity. *Speech Communication*, 15(3-4:Spoken Dialogue), 283–300.
- Oviatt, S. L., DeAngeli, A., & Kuhn, K. (1997). Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (pp. 415–422). New York, NY, USA: ACM.
- Oviatt, S. L., & Kuhn, K. (1998). Referential Features and Linguistic Indirection in Multimodal Language. In R. Mannell & J. Robert-Ribes (Eds.), *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)* (Vol. 6, pp. 2339–2342). Sydney, Australia.
- Oviatt, S. L., & Olsen, E. (1994). Integration Themes in Multimodal Human-Computer Interaction. In K. Shirai, S. Furui, & K. Kakehi (Eds.), *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP)* (Vol. 2, pp. 551–554). Yokohoma, Japan: Acoustical Society of Japan.
- Oviatt, S. L., & VanGent, R. (1996). Error Resolution during Multimodal Human-Computer Interaction. In T. Bunnell & W. Idsardi (Eds.), *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)* (Vol. 1, pp. 204–207). Philadelphia, PA, USA.
- OZONE. (2004). *OZONE O3 Homepage, Offering an Open and Optimal roadmap towards consumer oriented ambient intelligence.* (IST 2000-30026 European Project. <http://www.hitech-projects.com/euprojects/ozone/>, Last accessed: 22.10.2006)
- Paliwal, K. K., & So, S. (2004). Scalable Distributed Speech Recognition using Multi-frame GMM-based Block Quantization. In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)* (pp. 617–620). Jeju, South Korea.
- Pastel, R., & Skalsky, N. (2004). Demonstrating Information in Simple Gestures. In *Proceedings of the 9th International Conference on Intelligent User Interfaces (IUI)* (pp. 360–361). Funchal, Madeira, Portugal: ACM.
- Pavlovic, V., Sharma, R., & Huang, T. (1997). Visual Interpretation of Hand Gestures for Human-Computer-Interaction: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), 677–695.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann.
- Pecourt, E., & Reithinger, N. (2004). Multimodal Database Access on Handheld Devices. In *the Companion Volume to the Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 206–209). Barcelona, Spain.
- Peters, B. (2006). *Multimodale Dialogverarbeitung auf mobilen Geräten.* Unpublished master's thesis, Department of Informatics, University of Saarland.
- Pfleger, N. (2004). Context Based Multimodal Fusion. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI)* (pp. 265–272). New York, NY, USA: ACM.

- Puterman, M. L. (Ed.). (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: John Wiley and Sons.
- Reeves, B., & Nass, C. (Eds.). (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Stanford, CA, USA: CSLI Publications.
- Reithinger, N., Alexandersson, J., Becker, T., Blocher, A., Engel, R., Löckelt, M., Müller, J., Pflieger, N., Poller, P., Streit, M., & Tschernomas, V. (2003). SmartKom: Adaptive and Flexible Multimodal Access to Multiple Applications. In *Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI)* (pp. 101–108). Vancouver, Canada.
- Reithinger, N., Bergweiler, S., Engel, R., Herzog, G., Pflieger, N., Romanelli, M., & Sonntag, D. (2005). A Look Under the Hood: Design and Development of the First SmartWeb System Demonstrator. In *Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI)* (pp. 159–166). New York, NY, USA: ACM.
- Reithinger, N., Fedeler, D., Kumar, A., Lauer, C., Pecourt, E., & Romary, L. (2005). MIAMM - A Multimodal Dialogue System Using Haptics. In J. van Kuppevelt, L. Dybkjaer, & N. O. Bersen (Eds.), *Advances in Natural Multimodal Dialogue Systems* (pp. 307–332). Berlin, Heidelberg, New York: Springer.
- Reithinger, N., Lauer, C., & Romary, L. (2002). MIAMM: Multidimensional Information Access using multiple modalities. In *Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*. (4 pp.). Copenhagen, Denmark.
- Rhodes, B., & Mase, K. (Eds.). (2005). *ISWC 2005: Wearable Computers: 9th IEEE International Symposium, Osaka, Japan, 18-21 October, 2005, Proceedings*. IEEE Computer Society.
- Rist, T., Baldes, S., Gebhard, P., Kipp, M., Klesen, M., Rist, P., & Schmitt, M. (2002). CrossTalk: An Interactive Installation with Animated Presentation Agents. In *Proceedings of the 2nd Conference on Computational Semiotics for Games and New Media* (pp. 61–67). Augsburg, Germany.
- Robert-Ribes, J., Schwartz, J.-L., Lallouache, T., & Escudier, P. (1998). Complementarity and Synergy in Bimodal Speech: Auditory, Visual, and Audio-visual Identification of French Oral Vowels in Noise. *Journal of the Acoustical Society of America*, 103(6), 3677–3689.
- Rocchi, C., Stock, O., Zancanaro, M., Kruppa, M., & Krüger, A. (2004). The Museum Visit: Generating Seamless Personalized Presentations on Multiple Devices. In *Proceedings of the 9th International Conference on Intelligent User Interfaces (IUI)* (pp. 316–318). New York, NY, USA: ACM.
- Rogers, W. T. (1978). The Contribution of Kinesic Illustrators toward the Comprehension of Verbal Behavior within Utterance. *Communication Research*, 5, 54–62.
- Rubin, P., Vatikiotis-Bateson, E., & Benoit, C. (Eds.). (1998). Special Issue on Audio-Visual Speech Processing. *Speech Communication*, 26(1-2).

- Rubine, D. (1991). Specifying Gestures by Example. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (pp. 329–337). New York, NY, USA: ACM.
- Schiel, F. (2006). Evaluation of Multimodal Dialogue Systems. In W. Wahlster (Ed.), *SmartKom: Foundations of Multimodal Dialogue Systems* (pp. 617–643). Berlin, Heidelberg, New York: Springer.
- Schiel, F., & Türk, U. (2006). Wizard-of-Oz Recordings. In W. Wahlster (Ed.), *SmartKom: Foundations of Multimodal Dialogue Systems* (pp. 541–570). Berlin, Heidelberg, New York: Springer.
- Schmidt, S. (2005). *Persönlichkeitsaspekte in Stimmen von anthropomorphen Produkten*. Unpublished master's thesis, Department of Informatics, University of Saarland.
- Schmitz, M., Baus, J., & Schmidt, S. (2006). Towards Anthropomorphized Objects: A Novel Interaction Metaphor for Instrumented Spaces. In T. Strang, V. Cahill, & A. Quigley (Eds.), *Pervasive 2006 Workshop Proceedings* (pp. 487–492). Dublin, Ireland.
- Schmitz, M., & Butz, A. (2006). Safir: Low-cost spatial audio for instrumented environments. In *Proceedings of the 2nd international conference on intelligent environments* (pp. 427–430). Athens, Greece.
- Schneider, M. (2003). A Smart Shopping Assistant Utilizing Adaptive Plan Recognition. In *ABIS Workshop on Adaptivity and User Modelling in Interactive Software Systems* (pp. 331–334). Karlsruhe, Germany.
- Schneider, M. (2004). Towards a Transparent Proactive User Interface for a Shopping Assistant. In *Proceedings of the Workshop on Multi-User and Ubiquitous User Interfaces (MU3I) at IUI* (pp. 10–15). Madeira, Portugal.
- Schneider, M., Kröner, A., & Wasinger, R. (2006). Augmenting Interaction in Intelligent Environments through Open Personal Memories. In *Proceedings of the 2nd International Conference on Intelligent Environments (IE)* (pp. 407–416). Athens, Greece.
- Schomaker, L., Nijtmans, J., Camurri, A., Lavagetto, F., Morasso, P., Benoît, C., Guiard-Marigny, T., Goff, B. L., Robert-Ribes, J., Adjoudani, A., Defée, I., Münch, S., Hartung, K., & Blauert, J. (1995). *A Taxonomy of Multimodal Interaction in the Human Information Processing System*. (Report from the Esprit Project 8579/MIAMI, Available at: <http://hwr.nici.kun.nl/%7Emiami/taxonomy/taxonomy.html>, last accessed: 22.10.2006)
- Shafer, G., & Pearl, J. (Eds.). (1990). *Readings in uncertain reasoning*. San Francisco, CA, USA: Morgan Kaufmann.
- Shneiderman, B. (Ed.). (1992). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (2nd ed.). New York, NY, USA: Addison-Wesley.
- Shneiderman, B., & Maes, P. (1997). Direct Manipulation vs. Interface Agents. *Interactions*, 4(6), 42–61.
- Shortliffe, E. H. (1976). *Computer-based Medical Consultations: MYCIN*. Elsevier.

- Shortliffe, E. H., & Buchanan, B. G. (1975). A Model of Inexact Reasoning in Medicine. *Mathematical Biosciences*, 23, 351–379.
- Silbernagl, S., & Despopoulos, A. (Eds.). (2003). *Taschenatlas der Physiologie* (6th ed.). Stuttgart, Germany: Thieme.
- Singh, M., Jain, A. K., & Singh, M. P. (1999). E-commerce over Communicators: Challenges and Solutions for User Interfaces. In *Proceedings of the 1st ACM Conference on Electronic Commerce (EC)* (pp. 177–186). New York, NY, USA: ACM.
- Spassova, L., Wasinger, R., Baus, J., & Krüger, A. (2005). Product Associated Displays in a Shopping Scenario. In *Proceedings of the 4th IEEE / ACM International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 210–211). Vienna, Austria: IEEE Computer Society.
- Stahl, C., Baus, J., Brandherm, B., Schmitz, M., & Schwartz, T. (2005). Navigational- and Shopping Assistance on the Basis of User Interactions in Intelligent Environments. In *Proceedings of the IEE International Workshop on Intelligent Environments (IE)* (pp. 182–191). University of Essex, Colchester, UK.
- Stahl, C., Baus, J., Krüger, A., Heckmann, D., Wasinger, R., & Schneider, M. (2004). REAL: Situated Dialogues in Instrumented Environments. In *Workshop on Invisible and Transparent Interfaces (ITI) at AVI* (pp. 10–15). Gallipoli, Italy.
- Stahl, C., & Hauptert, J. (2006). Taking Location Modelling to New Levels: A Map Modelling Toolkit for Intelligent Environments. In M. Hazas, J. Krumm, & T. Strang (Eds.), *2nd International Workshop on Location- and Context-Awareness (LoCA)* (pp. 74–85). Berlin, Heidelberg, New York: Springer.
- Stock, O. (1991). ALFRESCO: Enjoying the Combination of Natural Language Processing and Hypermedia for Information Exploration. In *AAAI Workshop on Intelligent Multimedia Interfaces* (pp. 197–224). Anaheim, CA, USA.
- Stork, D. G., & Hennecke, M. E. (Eds.). (1995). *Speechreading by Humans and Machines: Models, Systems, and Applications* (1st ed.). Berlin, Heidelberg, New York: Springer.
- Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, 26(2), 212–215.
- Tan, T., Shi, Y., & Gao, W. (Eds.). (2000). *Advances in Multimodal Interfaces - ICMI 2000* (Vol. 1948). Berlin, Heidelberg, New York: Springer.
- Tessitore, L., & Hahn, W. v. (2000). Functional Validation of a Machine Interpretation System: Verbmobil. In W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation* (pp. 611–631). Berlin, Heidelberg, New York: Springer.
- Turk, M., & Robertson, G. (Eds.). (2000). Perceptual User Interfaces. *Communications of the ACM (Special issue)*, 43(3), 32–70.

- Ullmer, B., & Ishii, H. (2001). Emerging Frameworks for Tangible User Interfaces. In J. M. Carroll (Ed.), *Human-Computer Interaction in the New Millennium* (pp. 579–601). Boston, MA, USA: Addison-Wesley.
- W3C-EMMA. (2005). *EMMA: Extensible MultiModal Annotation markup language*. (W3C Working Draft, 16 September 2005, <http://www.w3.org/TR/emma/>, last accessed: 22.10.2006)
- W3C-InkMarkup. (2004). *Ink Markup Language*. (W3C Working Draft 28 September 2004, <http://www.w3.org/TR/InkML/>, last accessed: 22.10.2006)
- W3C-MMIReqs. (2003). *Multimodal Interaction Requirements*. (W3C Note, 8 January 2003, <http://www.w3.org/TR/mmi-reqs/>, last accessed: 22.10.2006)
- W3C-NLSML. (2000). *NLSML: Natural Language Semantics Markup Language for the Speech Interface Framework*. (W3C Working Draft 20 November 2000, <http://www.w3.org/TR/nl-spec/>, last accessed: 22.10.2006)
- W3C-OWL. (2004). *OWL: Web Ontology Language*. (W3C Recommendation 10 February 2004, <http://www.w3.org/TR/owl-ref/>, last accessed: 22.10.2006)
- W3C-RDF. (2004). *RDF: Resource Descriptive Framework*. (W3C Recommendation, <http://www.w3.org/RDF/>, last accessed: 22.10.2006)
- W3C-RDFS. (2004). *RDFS: Resource Descriptive Framework Schema 1.0*. (W3C Recommendation 10 February 2004, <http://www.w3.org/TR/rdf-schema/>, last accessed: 22.10.2006)
- W3C-StochasticLanguageModels. (2001). *Stochastic Language Models (N-Gram) Specification*. (W3C Working Draft 3 January 2001, <http://www.w3.org/TR/ngram-spec/>, last accessed: 22.10.2006)
- W3C-VoiceXML. (2004). *Voice Extensible Markup Language*. (W3C Recommendation 16 March 2004, <http://www.w3.org/TR/voicexml20/>, last accessed: 22.10.2006)
- W3C-XForms. (2006). *XForms 1.0 (2nd ed.)*. (W3C Recommendation 14 March 2006, <http://www.w3.org/TR/xforms-datamodel/>, last accessed: 22.10.2006)
- W3C-XML. (2006). *XML: eXtensible Markup Language 1.0 (4th ed.)*. (W3C Recommendation 16 August 2006, <http://www.w3.org/TR/REC-xml/>, last accessed: 22.10.2006)
- W3C-XMLSchema. (2004). *XML Schema, Part 2: Datatypes*. (W3C Recommendation 28 October 2004, <http://www.w3.org/TR/xmlschema-2/>, last accessed: 22.10.2006)
- Wahlster, W. (1991). User and Discourse Models for Multimodal Communication. In J. W. Sullivan & S. W. Tyler (Eds.), *Intelligent User Interfaces* (pp. 45–67). New York: ACM.
- Wahlster, W. (2000). Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System. In W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation* (pp. 3–21). Berlin, Heidelberg, New York: Springer.
- Wahlster, W. (2002a). *Multimodal and Natural Language Dialogue Systems*. (Presented as a winter semester (WS 02/03) special lecture series at the University of Saarland)

- Wahlster, W. (2002b). SmartKom: Fusion and Fission of Speech, Gestures, and Facial Expressions. In *Proceedings of the 1st International Workshop on Man-Machine Symbiotic Systems* (pp. 213–225). Kyoto, Japan.
- Wahlster, W. (2003). SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In *Proceedings of the Human Computer Interaction Status Conference* (pp. 47–62). Berlin, Germany.
- Wahlster, W. (2004). Conversational User Interfaces. *it: Information Technology*, 46(6), 289–290.
- Wahlster, W. (2006a). Dialogue Systems Go Multimodal: The SmartKom Experience. In W. Wahlster (Ed.), *SmartKom: Foundations of Multimodal Dialogue Systems* (pp. 3–27). Berlin, Heidelberg, New York: Springer.
- Wahlster, W. (Ed.). (2006b). *SmartKom: Foundations of Multimodal Dialogue Systems*. Berlin, Heidelberg, New York: Springer.
- Wahlster, W., Blocher, A., & Reithinger, N. (2001). SmartKom: Multimodal Communication with a Life-Like Character. In *Proceedings of the 7th European Conference on Speech Communication and Technology (EuroSpeech)* (pp. 1547–1550). Aalborg, Denmark.
- Wahlster, W., Krüger, A., & Baus, J. (2004a). *Kognitive Technologien für Praktischen Anwendungen*. (Finanzierungsantrag, RENA: Resource-adaptive Navigation, p. 1-29)
- Wahlster, W., Krüger, A., & Baus, J. (2004b). *Resource-adaptive Cognitive Processes*. (Finanzierungsantrag, BAIR: User Adaptation in Instrumented Rooms, p. 1-24)
- Wahlster, W., Kröner, A., & Heckmann, D. (2006). SharedLife: Towards Selective Sharing of Augmented Personal Memories. In O. Stock & M. Schaerf (Eds.), *Reasoning, Action and Interaction in AI Theories and Systems* (pp. 327–342). Berlin, Heidelberg, New York: Springer.
- Wasinger, R., Kray, C., & Endres, C. (2003). Controlling multiple devices. In *Physical Interaction (PI03) Workshop on Real World User Interfaces at MobileHCI* (pp. 60–63). Udine, Italy.
- Wasinger, R., & Krüger, A. (2004). Multi-modal Interaction with Mobile Navigation Systems. *Special Journal Issue Conversational User Interfaces, it - Information Technology*, 46(6), 322–331.
- Wasinger, R., & Krüger, A. (2005). Modality Preference: Learning from Users. In *Workshop on User Experience Design for Pervasive Computing (Experience) at Pervasive*. Munich, Germany.
- Wasinger, R., & Krüger, A. (2006). Modality Preferences in an Instrumented Environment. In *Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI)* (pp. 336–338). Sydney, Australia.
- Wasinger, R., Krüger, A., & Jacobs, O. (2005). Integrating Intra and Extra Gestures into a Mobile and Multimodal Shopping Assistant. In *Proceedings of the 3rd International Conference on Pervasive Computing (Pervasive)* (pp. 297–314). Munich, Germany.

- Wasinger, R., Oliver, D., Heckmann, D., Braun, B., Brandherm, B., & Stahl, C. (2003). Adapting Spoken and Visual Output for a Pedestrian Navigation System, based on given Situational Statements. In *Workshop on Adaptivity and User Modelling in Interactive Software Systems (ABIS)* (pp. 343–346). Karlsruhe, Germany.
- Wasinger, R., Schneider, M., Baus, J., & Krüger, A. (2004). Multimodal Interactions with an Instrumented Shelf. In *Workshop on Artificial Intelligence in Mobile Systems (AIMS) at UbiComp* (pp. 36–43). Nottingham, UK.
- Wasinger, R., Stahl, C., & Krüger, A. (2003a). M3I in a Pedestrian Navigation & Exploration System. In L. Chittaro (Ed.), *Proceedings of the 5th International Symposium on Human-Computer Interaction with Mobile Devices and Services (Mobile HCI)* (pp. 481–485). Udine, Italy: Springer.
- Wasinger, R., Stahl, C., & Krüger, A. (2003b). Robust Speech Interaction in a Mobile Environment through the use of Multiple and Different Media Input Types. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)* (pp. 1049–1052). Geneva, Switzerland.
- Wasinger, R., & Wahlster, W. (2006). Multi-modal Human-Environment Interaction. In E. Aarts & J. L. Encarnação (Eds.), *True Visions: The Emergence of Ambient Intelligence* (pp. 291–306). Berlin, Heidelberg, New York: Springer.
- Weiser, M. (1991). The Computer for the Twenty-First Century. *Journal of Communication*, 265(3), 94–104.
- Wikipedia. (2006a). *Animism - Wikipedia, The Free Encyclopedia*. (Online; accessed 22.10.2006, <http://en.wikipedia.org/wiki/Animism>)
- Wikipedia. (2006b). *Finite State Machine - Wikipedia, The Free Encyclopedia*. (Online; accessed 22.10.2006, http://en.wikipedia.org/wiki/Finite_state_machine)
- Wikipedia. (2006c). *Optical Character Recognition - Wikipedia, The Free Encyclopedia*. (Online; accessed 22.10.2006, http://en.wikipedia.org/wiki/Optical_character_recognition)
- Wikipedia. (2006d). *Perception - Wikipedia, The Free Encyclopedia*. (Online; accessed 22.10.2006, <http://en.wikipedia.org/wiki/Perception>)
- Wikipedia. (2006e). *Sense - Wikipedia, The Free Encyclopedia*. (Online; accessed 22.10.2006, <http://en.wikipedia.org/wiki/Sense>)
- Wikipedia. (2006f). *Speech Recognition - Wikipedia, The Free Encyclopedia*. (Online; accessed 22.10.2006, http://en.wikipedia.org/wiki/Speech_recognition)
- Wilamowitz-Moellendorff, M. von, Müller, C., Jameson, A., Brandherm, B., & Schwartz, T. (2005). Recognition of Time Pressure via Physiological Sensors: Is the User's Motion a Help or a Hindrance? In *Proceedings of the Workshop on Adapting the Interaction Style to Affective Factors, in conjunction with the User Modeling Conference* (pp. 43–48). Edinburgh, UK.

- Wu, L., Oviatt, S. L., & Cohen, P. R. (1999). Multimodal Integration: A Statistical View. *IEEE Transactions on Multimedia*, 1(4), 334–341.
- Xiao, B., Girand, C., & Oviatt, S. L. (2002). Multimodal Integration Patterns in Children. In J. Hansen & B. Pellom (Eds.), *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)* (pp. 629–632). Denver, CO, USA: Casual Productions.
- Xiao, B., Lunsford, R., Coulston, R., Wesson, M., & Oviatt, S. (2003). Modeling Multimodal Integration Patterns and Performance in Seniors: Toward Adaptive Processing of Individual Differences. In *Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI)* (pp. 265–272). New York, NY, USA: ACM.
- Xu, F. (2003). Multilingual WWW: Modern Multilingual and Crosslingual Information Access Technologies. In W. Abramowicz (Ed.), *Knowledge-Based Information Retrieval and Filtering from the Web* (pp. 165–184). Boston, MA, USA: Kluwer Academic.
- Zhu, W., Owen, C. B., Li, H., & Lee, J.-H. (2004). Personalized In-store E-Commerce with the PromoPad: An Augmented Reality Shopping Assistant. *Electronic Journal for E-Commerce Tools and Applications (eJETA)*, 1(3).

A	
Accuracy	
Gesture	98
Handwriting	98
Rate	97, 162
Speech	98
User feedback	167
Value	98, 167
Acoustic model	78, 114
Always Best Connected (ABC)	26, 114
Animism	115
Anthropomorphization	58, 64, 115
Anthropomorphized objects	21, 37, 80, 117, 202
B	
Backus-Naur Form	78, 137
Bayesian networks	173
Blackboard	103, 109, 137, 146, 150, 152, 156, 162
Data attributes	145 f.
Event filtering	169
Interaction node	146, 160
Method attributes	145 f., 148
BMW Personal Navigator (BPN)	31
BNF	<i>see</i> Backus-Naur Form
C	
Certainty factors	158, 172, 174
Chi-square	187
Communication	8
Bodily	10
Emblems	10
Illustrators	10
Many-to-many	8
Nonverbal	9
Auditory	9
Invisible	9
Visual	9
Now-to-future	8
One-to-many	8
One-to-one	8
Verbal	9
Visual	
Artefactual	10
Kinesic	9
Proxemic	10
Communication act	23, 58, 134, 136, 140, 145, 150, 154, 159, 162, 168
..... <i>see also</i> Frame-based structure	
Element	<i>see</i> Semantic constituent
Frame	136, 150
Slot	23, 134, 136, 150, 154, 162, 170
Communication modes	8, 14, 32, 64, 75
Gesture	64
Eye-gaze	64
Facial expressions	64
Hand movement	64
Pointing	64
Speech	64
Writing	64
Communication protocol	140
Computer	
Desktop	13
Desktop metaphor	24
Input device	10
Mobile	25, 28
... <i>see also</i> Personal Digital Assistant	
Personal	24

- Platform 64
- Confidence scoring 89, 162
- Confidence value . 89, 97, 148, 157, 162, 174
- Comparing 95, 99, 163
- Extra-gesture 95
- Handwriting 90
- Intra-gesture 92
- Re-weighted 162, 166, 174
- Speech 90
- Conflict resolution 162, 168, 171
- Conrad Electronic 117, 185
- Control 6
- Device 130
- Exclusive 131
- Mixed-initiative 131
- Shared 131
- System-initiated 131
- User-initiated 131
- Cross-selling 120
- D**
- Data attributes *see* Blackboard
- Data container 80, 182
- Decision theory 173
- Deixis 20
- *see also* Referring modes
- Pars-pro-toto 21, 47
- Person 21
- Place 21
- Spatial 21
- Dempster-Shafer theory of evidence 173
- Device 126
- Input 128
- Mobile 25
- .. *see also* Personal Digital Assistant
- Multiple 130
- Non-shared 130
- Output 128
- Private 128
- Public 128
- Shared 130
- Direct manipulation 29, 76, 84
- Discourse history 149, 156, 162
- *see also* History context
- E**
- EMMA 15, 108, 141
- Error 13, 66, 97, 104, 178
- Extra-gesture *see* Gesture
- F**
- Feature structure 49, 137
- *see also* Frame-based structure
- Frame-based structure 134, 150
- G**
- Gesture 35, 84, 98, 101, 196
- Extra-gesture 35, 87, 95, 206
- One-to-one mapping 84
- Intra-gesture 35, 73, 85, 92, 206
- Pickup 10, 87, 119
- Point 10, 85, 87
- Putdown 10, 87, 119
- Selection-gesture 10, 85
- Slide 35, 85, 94
- Givenness Hierarchy 107
- Grammar
- *see also* Language model
- Context-free 78
- Finite-state 78, 135
- Rule-grammar 79
- Grice Maxim of Quantity 19, 106
- H**
- Handwriting 82, 90, 98, 100, 196, 206
- History context 20, 154
- *see also* Discourse history
- Human-Computer Interaction (HCI) *see* Interaction
- I**
- I/O
- Computer input 6
- Computer output 6
- Human input 6
- Human output 6
- Peripherals 5
- System output 6

- User input 6
- Input 14
 - Active 76, 113
 - Composite 15
 - Non-overlapped 101, 103, 105
 - Overlapped 101, 103
 - Overlapped and conflicting ... 105, 111, 174
 - Overlapped and non-conflicting ... 105, 111
 - Passive 76, 113
 - Semantic 101, 105
..... *see also* Semantic
 - Sequential 15, 102
 - Simultaneous 15, 102
 - Temporal 101, 103
..... *see also* Temporal
- Instrumented
 - Device 128
 - Environment 75, 126, 128
 - Object 126
- Interaction 14, 31, 126
 - Collaborative 127
 - Computer-human 11
 - Direct 37, 58, 115, 118, 202
 - Distal 87
 - Human-computer 5, 11
 - Independent 127
 - Indirect 37, 58, 115, 118, 202
 - Map-based 45
 - Multimodal 14
 - Off-device 87, 121
 - On-device 85, 121
 - Proximal 87
 - Real-world 64
 - Sequential 102
 - Simultaneous 102
- Interaction manager 121, 138
- Interaction modifier 118
- Interaction node *see* Blackboard
- Interaction shopping 37
- Interface 126
 - Graspable user interface 84
 - Multi-device single-user 127
 - Multi-user multi-device 127
 - Multi-user single-device 126
 - Single-user single-device 126
 - Tangible user interface 29, 76, 84
 - Digital representation 84
 - Physical representation 84
- Intra-gesture *see* Gesture
- K**
- Knowledge 134
 - Data 135
 - Information 135
 - Knowledge 135
 - Understanding 135
 - Wisdom 135
- L**
- Language
 - Multimodal 14
 - Spoken 14
- Language model 78, 135, 137
 - Formal 78
 - N-gram 78
- M**
- M3L 56, 144
- Mann-Whitney U 187
- Markov decision processes 173
- Media fusion *see* Modality fusion
- Method attributes *see* Blackboard
- Mixed-initiative dialogue system 120
- MMIL 144
- Mobile ShopAssist (MSA) 37
- Modal *see* Modality
- Modality 8, 14
 - Combination . 65, 77, 96, 108, 153, 183, 189, 207
 - Intuition 193
 - Observability 198
 - Preference 199
 - Type 136
- Modality event 142, 160, 169
..... *see also* Interaction node
- Modality fission 16, 121
- Modality fusion 16, 152, 157, 162
 - Early fusion 17
 - Late fusion 17, 89

- Overlay 159
 Unification 159
 Multilingual 64
 Multimodal fusion *see* Modality fusion
 Multimodal input 14
 Adaptability 13
 Efficiency 12
 Flexibility 12
 Multimodal integration . *see* Modality fusion
 Multimodal interaction *see* Multimodal input
 Multimodality 14
 Complementary 108
 Supplementary 108
 Symmetric 10, 64, 121
 Mutual disambiguation 16, 157
- N**
- N-best lists 89, 154, 167, 173
 NLSML 143
 Node *see* Blackboard
 Number constraint 145
- O**
- On-the-go 30
 Ontology 78, 135, 138, 140
 Open-microphone 49
- P**
- Pars-pro-toto *see* Deixis
 Perception 6
 Personal Digital Assistant (PDA) 24, 32, 37,
 78, 82, 87, 127, 184
 Pickup *see* Gesture
 Point *see* Gesture
 Polymodal 55
 Presentation output planner 121
 Presentation planner 138
 Privacy 31, 132, 196
 Putdown *see* Gesture
- Q**
- query+feature 152, 168
 *see also* Sentential mood
- R**
- Recognition
 Distributed 114
 Embedded 143
 Extra-gesture 87
 Handwriting 82
 Intra-gesture 85
 Rate 90, 153, 182
 Speech 78
 User feedback 167
 Recognizer
 Black-box 166
 Glass-box 166
 Same-type recognizers 114
 Speech
 Always listening 81
 Push to activate 81
 Push to talk 81
 Redundancy 105, 112
 Crossmodal 105
 Reference 19
 *see also* Referring modes
 Reference resolution 19, 139
 Crossmodal 23, 139
 Referent 22, 156
 Type identifiable 106
 Unidentifiable 106
 Uniquely identifiable 106
 Referential terms 20, 22
 Common nouns 22
 Demonstratives 22
 Noun 22
 Noun phrase 22
 Pronouns 22
 Proper nouns 22
 Specifiers 22
 Referring modes 20
 Ellipsis 21
 Expansion 21
 Substitution 21
 Endophora 20
 Anaphora 20
 Cataphora 20
 Exophora 20
 Deixis 20, 190

- RFID 40, 68, 76, 88, 95, 182
 Active 26, 36, 77
 Passive 27, 88, 97
- S**
- Saliency 156
 Scalability 100
 Selection-gesture *see* Gesture
 Semantic
 Non-overlapped 111, 169
 Order 104
 Overlap
 Degree of 106
 Partial 106
 Overlapped 111, 170, 174
 Semantic space 106
 Synchrony 104
 Semantic constituent 136
 Command 147, 151, 168
 Feature 58, 81, 96, 103, 109, 147
 Object 58, 81, 96, 103, 109, 147
 Query 147, 150
 Semantic element . *see* Semantic constituent
 Semantic interpreter 147
 Semantic Web 60, 138
 Semantics 134
 Logical form 134
 World knowledge 134
 Senses
 Classical 6
 Hearing 6
 Sight 6
 Smell 6
 Taste 6
 Touch 6
 Distance senses 6
 Proximity senses 7
 Sense organ 7
 Touch
 Breakaway force 10
 Haptics 7
 Tactition 7
 Sentential mood 150
 *see also* Semantic constituent
 Assert 150
 Command 150
 Query 150
 Slide *see* Gesture
 Speech 78, 90, 98, 100, 196, 205
 Speech grammar
 Direct 119
 Indirect 119
 Keyword 119
 Statistical significance 187
 Super-additivity 17
 Symmetric multimodality *see* Multimodality
 Syntax 134
 System
 Multimedia 14
 Multimodal 14
- T**
- Tangible interaction 76
 Tangible User Interface (TUI) . *see* Interface
 Temporal
 Duration 102
 Order 103, 169
 After 103
 Before 103
 During 103
 Synchrony 102
 Time space 106
 Tense
 First person 118
 Second person 118
 Third person 118
 Thesaurus 78, 139
 Time
 Past 149
 Present 149
 Timeframe 153, 169
 Timestamp 27, 103, 146, 169
 Type constraint 145, 156
- U**
- Uncertain reasoning 172
 Unimodal 108
 Up-selling 120
 Usability
 Results

- Extra-gesture 206
 - Handwriting 206
 - Intra-gesture 206
 - Intuition 193
 - Modality combination 189, 207
 - Observability 198
 - Preference 199
 - Speech 205
 - Study 183 f.
- User
- Demographic 13, 187
 - Mobility 64
 - Preference 107, 112, 187
- User input 11, 76, 136, 159
- Utterance 14, 150
- V**
- Visual-WCIS *see* What-Can-I-Say
- Vocabulary 78
- W**
- WCIS *see* What-Can-I-Say
- What-Can-I-Say .. 39, 85, 94, 101, 116, 206
- Wilcoxon 187
- WIMP *see* Windows Icon Menu Pointer
- Windows Icon Menu Pointer 12