Point and Speak

Design und Evaluation einer mobilen multimodalen Interaktionstechnik zur orts- und orientierungsbezogenen Serviceauswahl

ausgeführt am
Institut für Gestaltungs- und Wirkungsforschung
der Technischen Universität Wien
unter Anleitung von
ao.Univ.Prof., DI Dr. Peter Purgathofer

in Zusammenarbeit mit dem
Forschungszentrum Telekommunikation Wien (ftw.)
betreut von
Mag. Peter Fröhlich und
Dr. rer. nat. Peter Reichl

durch
Andreas Brauneis
Kopalgasse 54/3/2, 1110 Wien

Danksagung

An dieser Stelle möchte ich all jenen danken, die durch ihre fachliche und persönliche Unterstützung zum Gelingen dieser Diplomarbeit beigetragen haben.

Der größte Dank gilt meinen Eltern, die mir dieses Studium ermöglichten.

Weiters bedanke ich mich bei ao.Univ.Prof., DI Dr. Peter Purgathofer, Mag. Peter Fröhlich und Dr. rer. nat. Peter Reichl für die Betreuung der Diplomarbeit.

Kurzfassung

Die rasche technologische Entwicklung im Bereich mobiler Geräte ermöglicht viele neue interessante Anwendungen. Die Diplomarbeit „Point and Speak“ erweitert das Interaktionskonzept „Point-to-Discover“ (p2d), das vom Forschungszentrum Telekommunikation Wien (ftw.) entwickelt wird, um ein Sprachdialogsystem. P2d erlaubt es Benutzern mit ihrem Handy auf digitale Informationen zuzugreifen, die mit physischen Objekten der unmittelbaren Umgebung verknüpft sind. Dies wird technisch durch integrierte Sensoren (GPS und elektronischer Kompass) und 3D-Modelle der Umgebung ermöglicht.

Ausgehend von einer Anforderungsanalyse, in der die zukünftigen Benutzer eingebunden wurden, wurde sowohl eine rein sprachliche Benutzerschnittstelle, als auch eine multimodale Benutzerschnittstelle entworfen. Bei der Implementierung handelte es sich um eingebettete Sprachein- und ausgabe, bei der die gesamte Verarbeitung am mobilen Gerät erfolgte. Eingebettete ASR (automatische Spracherkennung) und TTS (Sprachsynthese) wurde mit Komponenten des Projektes p2d kombiniert, wie etwa dem bereits entwickelten Sensorboard. Mit den Prototypen wurde eine abschließende Benutzerevaluierung durchgeführt. Das sprachliche Interface wurde von den Testpersonen als vielversprechende Alternative zu den bisherigen Interaktionsmöglichkeiten charakterisiert.

Abstract

The rapid technological development of mobile devices enables a wide range of new interesting applications. The diploma thesis “Point and Speak” expands the interaction concept of “Point-to-Discover” (p2d), with a speech interface. p2d, which has been developed at the Telecommunication Research Center Vienna (ftw.), facilitates the access to digital information related to objects in the closest surrounding via a cell phone. This is realized by sensors attached to the mobile phone, namely GPS and digital compass, as well as 3D environmental models.

Departing from a user-based requirements analysis, a speech-only and a multimodal interface have been designed and developed. An embedded speech application was implemented, facilitating all speech processing on the mobile device. The developed ASR (Automatic-Speech-Recognition) and TTS (Text-To-Speech) software was combined with components p2d, for instance the developed sensor board. A concluding user evaluation with the prototypes was carried out. The speech interface was accepted by the users as a promising alternative to the traditional interaction possibilities.

Inhalt

Abkürzungsverzeichnis

Es folgt ein kurzer Überblick über die in der Diplomarbeit häufig verwendeten Abkürzungen.

Abkürzung	Steht für	Erklärung
GUI	Graphical User Interface	Graphische Benutzeroberfläche
ftw.	Forschungszentrum Telekommunikation Wien	
ASR	Automatic Speech Recognition	Automatische Spracherkennung
TTS	Text To Speech	Sprachsynthese
LiLiPUT	Lightweight Lab Equipment for Portable User Testing in Telecommunications	Ausrüstung zur Aufzeichnung von mobilen Benutzertests im Bereich der Telekommunikation..
Bargein	Bargein	Unterbrechung einer Ausgabe durch einen Benutzerbefehl
p2d	Point 2 Discover	Interaktions-Konzept mit dem das Mobiltelefon zum interaktiven Zeigestab wird.
PDA	Personal Digital Assistant	Persönlicher, digitaler Assistent
AR	Augmentetd Reality	Erweiterte Realität
POI	Points of Interest	Interessante Punkte in der Umgebung

Tabelle 1: häufig verwendete Abkürzungen in der Diplomarbeit „Point and Speak“

1 Einleitung

1.0 Vorbemerkung zur Schreibweise

Beim Verfassen der Arbeit habe ich mich aufgrund der leichten Lesbarkeit dazu entschieden, immer die männliche Form der Geschlechtspronomen zu wählen. Es sei jedoch ausdrücklich darauf hingewiesen, dass Frauen und Männer im gleichen Maße berücksichtigt werden, solange es nicht anders vermerkt ist.

1.1 Überblick

In diesem Kapitel wird die Idee der Diplomarbeit „Point and Speak“ vorgestellt. „Point and Speak“ versucht anhand der aktuellen technischen Möglichkeiten, die bei mobilen Geräten bereits vorhanden sind, ein Sprachdialogsystem zu entwickeln. Da es sich bei diesem Dialogsystem um eine Erweiterung des am Forschungszentrum Telekommunikation Wien (ftw.) durchgeführten Projektes „Point 2 Discover“ („p2d“) handelt, wird das Projekt „p2d“ im Laufe dieses Kapitels genauer vorgestellt. Auch werden andere Projekte und Forschungsergebnisse zusammengefasst, die für diese Diplomarbeit relevant sind. Die Aufgabenstellung und Ziele werden beschrieben und es werden Fragen gestellt, die während der Konzeptentwicklung aufgetaucht sind. Diese gilt es, im Laufe der Diplomarbeit zu beantworten.

1.2 Motivation

Die technischen Möglichkeiten wie Rechenleistung, Display oder Speicherkapazität nehmen bei Handys und mobilen Computern, sogenannten Pocket PCs, nehmen ständig zu. Dadurch ist es möglich, immer neue Anwendungen und Services auf Handys zu integrieren. Hat man früher mit dem Handy nur telefonieren und SMS schreiben können, so gibt es heute schon viel mehr Möglichkeiten, ein Mobiltelephon einzusetzen: etwa als Terminkalender, Radio, MP3-Player, Kamera oder als Browser. Das Handy wird immer mehr zu einem persönlichen Assistenten, welches als tägliches Hilfsmittel nicht mehr wegzudenken ist.

Die vorliegende Diplomarbeit „Point and Speak“ befasst sich mit einer weiteren neuen Anwendungsmöglichkeit, die im engen Zusammenhang mit dem Projekt „p2d – Point 2 Discover“ steht [HPp2d][SiFr06]. Es stellt eine neue Möglichkeit zur Verknüpfung von

digitalen Informationen mit der physischen Realität dar. Mit Hilfe des Handys können Informationen zu interessanten Punkten in der Umgebung geholt werden, etwa historische Informationen über Sehenswürdigkeiten oder den Busfahrplan einer Haltestelle. Durch einfaches Zeigen mit dem Handy auf die für die Benutzer interessanten Punkte wie zum Beispiel Gebäude können diese verfügbare Daten und Services anfordern.

Die Diplomarbeit „Point and Speak“ stellt ein Erweiterungsmodul für „p2d“ dar. Zusätzlich zur herkömmlichen Interaktion mit dem Handy beziehungsweise Pocket PCs mittels Tastatur oder Touchscreens stellt „Point and Speak“ eine sprachliche Komponente zur Verfügung. Man kann dem Handy gesprochene Befehle geben, welche die Anwendung automatisch erkennt (ASR), und so durch das Programm navigieren. Die Ausgabe des Programmes erfolgt dann sowohl visuell als auch via Sprachsynthese (TTS). Da die Sprachinteraktion eine zusätzliche Möglichkeit darstellt und nicht ausschließlich akustisch mit dem Handy kommuniziert wird, wird hier von einem multimodalen Interface gesprochen. Ein multimodales Interface soll dem Benutzer die Handhabung des Programmes erleichtern. Wenn mit dem Handy gerade auf ein Gebäude gezeigt wird, ist es umständlich, Informationen auf dem Display zu lesen oder Befehle über die Tastatur beziehungsweise den Touchscreen einzugeben. Hier ist es sehr hilfreich, die Informationen vorgelesen zu bekommen und auch Befehle sprachlich eingeben zu können.

1.3 Aufbau dieser Arbeit

Diese Diplomarbeit ist in folgende Bereiche untergliedert:

- **Einleitung:** Im Laufe dieses Kapitels werden themenbezogene Forschungsergebnisse vorgestellt und das der Diplomarbeit „Point and Speak“ zugrundeliegende Interaktionskonzept „Point 2 Discover“ beschrieben. Außerdem werden Fragen aufgelistet, die im Laufe der Diplomarbeit beantwortet werden.
- **Anforderungsanalyse:** Zu Beginn des Projekts wurde am ftw. ein Benutzertest durchgeführt der im zweiten Kapitel beschrieben wird. Den Testpersonen wurde anhand eines Prototyps die Idee, die hinter „Point and Speak“ und „p2d“ steht, nähergebracht und deren Meinungen dazu eingeholt.
- **Design der Benutzerschnittstelle:** Nachdem die Benutzertests ausgewertet waren, wurde anhand der Ergebnisse die Anwendung entworfen. Szenarien, wo die Anwendung eingesetzt werden kann, und der Entwurf des Programms werden im dritten Kapitel beschrieben.
- **Implementierung des Prototypen:** Im vierten Kapitel der Diplomarbeit wird die Implementierungsphase beschrieben, wo Software von Nuance und Loquendo verwendet wurde. Auch die verwendeten Sensoren und die für die Anwendung wichtige Grammatik werden in diesem Kapitel erklärt.
- **Benutzerevaluation:** Am Ende der Entwicklung wurde ein weiterer Benutzertest durchgeführt. Hier hatten die Testpersonen die Gelegenheit ein Sprachinteraktionssystem auszuprobieren und ihre Meinung dazu kundzutun. Abgeschlossen wird die Diplomarbeit durch eine Zusammenfassung und einen Ausblick auf weitere Arbeiten.

1.4 Themenbezogene Forschungsergebnisse

Im Folgenden werden themenbezogene Forschungsergebnisse der letzten Jahre vorgestellt. Ausgehend von einer kurzen Übersicht über existierende Projekte der Verknüpfung von digitalen und realen Informationen wird auch der „p2d“ Ansatz erklärt. Schließlich werden Bezüge zu relevanten Arbeiten im Themenbereich Sprachinteraktion vorgestellt.

1.4.1 Die Verknüpfung von digitalen Informationen mit der physischen Realität

Max Egenhofer – [Egen99] stellte in den späten 90er Jahren das Konzept der Spatial Information Appliance vor. Er gab neue Impulse, um mit Hilfe der neuen Technologien innovative Anwendungen entwickeln zu können. Der Hintergrund hierfür war, dass neue Technologien wie das Internet oder GPS-Sensoren verfügbar waren, diese Entwicklungen aber im Rahmen von mobilen Geräten kaum genutzt wurden. Einige dieser von Max Egenhofer erdachten Impulse sind:

- „*Smart Compasses*“ sollen dem Benutzer die Möglichkeit geben, sich in einem für ihn unbekannten Gelände zurecht zu finden., ähnlich dem magnetischem Kompass. Der Unterschied liegt darin, dass der magnetische Kompass immer in dieselbe Richtung zeigt. Ein „Smart Compass“ zeigt in die Richtung, wohin der Benutzer gehen will. Auch wenn der Benutzer das Display dreht, zeigt der „Smart Compass“ immer noch zum vom Benutzer gewählten Zielpunkt.
- „*Smart Horizons*“ erlauben dem Benutzer, hinter das aktuell Sichtbare zu sehen. Informationen, wie das Wetter, Topographie oder öffentliche Verkehrsmittel, die nicht unmittelbar sichtbar sind, werden so dem Benutzer zur Verfügung gestellt.
- „*Geo-Wands*“ sind geographische Zeigestäbe. Man zeigt mit dem Handy auf ein physikalisches Objekt. Zu diesem Objekt werden dann Informationen zur Verfügung gestellt.
- „*Smart Glasses*“ sind eine Erweiterung der Realität. Zusätzliche Informationen sollten hierbei über das digitale Kamerabild gelegt werden.

Damals waren das alles nur Konzepte, die noch nicht realisiert waren. Viele Projekte im Bereich von mobilen Geräten beschäftigen sich mit den Ideen von Max Egenhofer und versuchen, diese zu verwirklichen

Mobiltelephone werden immer häufiger benutzt, um eine Verbindung zwischen der physikalischen Welt und digitalen Informationen herzustellen. Durch die ständig neuen Entwicklungen bei Mobiltelephonen können neue Möglichkeiten ausgenutzt werden. Mithilfe der schon in fast jedem Handy integrierten Kameras können visuelle Codes (Semacodes), die an physische Objekten angebracht sind, aufgenommen und interpretiert werden [SemO]. Near field communication [NFCR] erlaubt es dem Mobiltelephon, auf kurze Distanz mit anderen NFC-Geräten, wie etwa Ticketautomaten, berührungsfrei zu kommunizieren. Andere ortsbezogene Dienste haben schon die breite Masse erreicht. Hierzu zählt zum Beispiel die Routenplanung.

Semapedia – The physical Wikipedia [SemO], entwickelt von Alexis Rondeau und Stan Wiechers, erlaubt es, zu interessanten Gebäuden via Mobiltelephon Informationen zu beschaffen.
Dies funktioniert folgendermaßen: An Objekten werden so genannte Semacodes angebracht. Semacodes sind graphische Repräsentationen von URLs. Der Semacode wird mit dem Handy photographiert und mittels Semapedia Reader, der auf dem Handy installiert ist, in eine URL umgewandelt. Diese URL kann dann mit dem Handy, sofern eine Internetverbindung zur Verfügung steht, aufgerufen werden. Semapedia nützt hier die freie Internetenzyklopädie Wikipedia, da dort viel Information zu einzelnen Gebäuden wie zum Beispiel Hofburg oder Stephansdom frei zur Verfügung steht.
Der User kann aber nicht nur die Semacodes fotografieren und sich Informationen holen, er kann auch selber Semacodes kreieren und Gebäude damit kennzeichnen. Dazu muss er nur auf der Homepage die gewünschte URL eingeben und diese wird dann in den entsprechenden Semacode umgewandelt. Ausgedruckt und mit Erlaubnis des Objekteigentümers am Gebäude angebracht, hilft er den anderen Usern von Semapedia, sich Informationen über das Objekt einzuholen.

Abbildung 1: An der Säule vor dem Centre Pompidou ist ein Semacode angebracht. Mittels Handy wird dieser photographiert und per Semapedia-Reader in eine URL umgewandelt.

Da das Projekt noch nicht sehr bekannt ist, sind auch nur selten Semacodes an Gebäuden zu finden. Doch mit einer entsprechend großen Community könnte das ein durchaus erfolgreiches Projekt werden, da die aktuellen Handys am Markt schon die technischen Voraussetzungen erfüllen. Das Projekt ist allerdings zusätzlich davon abhängig, ob Wikipedia und Semacodes kostenfreie Projekte bleiben.

Beim Einbau von GPS-Sensoren in mobile Geräte ergeben sich viele interessante Möglichkeiten, die über GPS empfangenen Daten sinnvoll zu nützen. Einige Projekte, die Ideen oder schon vorhandene Anwendungen vorstellen, werden hier beschrieben.

LAMP3D [BuCh05] ist eine Anwendung für Pocket PCs. Mit Hilfe von einem GPS Modul wird die Position und die Orientierung des Benutzers festgestellt. Die Orientierung des Benutzers wird durch die Folge der letzten durch GPS ermittelten Positionen berechnet. Die Umgebung wird dann auf dem Display mittels VRML (=Virtual Reality Modeling Language, eine Beschreibungssprache für 3D-Szenen) dreidimensional dargestellt. Der Benutzer wählt dann auf dem Display ein Gebäude oder Objekt aus, zu dem er Informationen haben möchte. Diese Informationen werden dann in Textform auf dem Pocket PC dargestellt. Dieses System ist schon fertig entwickelt und es wurden auch bereits Tests mit Usern durchgeführt. Die Testpersonen fanden sich in dem System schnell zurecht. Durch die 3D-Präsentation der Umgebung auf dem Display kann der User schnell und ohne Probleme die richtigen Gebäude auswählen. Probleme gab es mit der Orientierung des Benutzers. Da diese, wie oben beschrieben, mittels den letzten ermittelten GPS-Punkten erfolgt, kann die Orientierung des Benutzers nicht festgestellt werden, wenn sich dieser auf dem selben Punkt in eine andere Richtung dreht. Um dieses Problem zu lösen, wird von den Entwicklern empfohlen, einen elektronischen Kompass in den Pocket PC zu integrieren. Auch die Genauigkeit und Verfügbarkeit von GPS wurde als Problem angegeben. Vor allem bei bewölktem Wetter sind die Daten des GPS-Systems ungenau. Ein weiters Problem ist die Anzeige der Verfügbarkeit von Informationen zu einzelnen Objekten der Umgebung in der virtuellen 3D-Ansicht.

CyPhone [Pyss00] ist schon eine etwas ältere Entwicklung (2000). Mittels GPS soll eine Navigation im Freien möglich sein. Da das Handydisplay als zu klein erachtet wird, bekommt der User einen speziellen Hut, einen sogenannten HMD (Head-Mounted-Display), aufgesetzt. Vordefinierte 3D-Umgebungen werden durch ein Display, das durch den Hut direkt vor den Augen platziert wird, dem User gezeigt. In der virtuellen Darstellung können auch Textlabels den Gebäuden beigefügt werden. Cyphone stellt auch andere Anwendungen zur Verfügung, wie zum Beispiel eine Verbindung zwischen den Usern, um Informationen von anderen Usern zu erhalten. Dafür wird auch eigens ein neues Protokoll eingeführt. Da diese Anwendung vom Thema dieser Arbeit aber stark abweicht, wird hier nicht näher darauf eingegangen.

Durch den Einbau von zusätzlichen Sensoren in mobile Geräte enstehen noch weitere Anwendungsmöglichkeiten, etwa das Scrollen durch ein Dokument mittels Bewegen des Handys [HiPi00]. Das nun folgende Projekt hat nicht direkt mit der Verknüpfung digitaler Informationen zur physikalischen Realität zu tun, stellt allerdings interessante Ideen zur Verwendung von Sensoren in mobilen Geräten vor.

Sensing Techniques for Mobile Interaction [HiPi00] aus dem Jahr 2000 versucht durch Integration von verschiedenen Sensoren einem PDA neue Möglichkeiten zur Interaktion mit dem User zu geben. Einem Cassiopeia E-105 Palm-sized PC wurden ein zweidimensionaler Tilt-Sensor, ein kapazitiver Berührungssensor und ein Infrarotsender zur Abstandsmessung eingebaut. Wenn der User den PDA zur Hand nimmt, was mit Hilfe des kapazitiven Berührungssensors festgestellt wird, wird der PDA eingeschaltet. Mit Hilfe der Tilt-Sensoren wird festgestellt, ob der PDA gedreht, nach links oder rechts, nach oben oder unten bewegt wird. Beim Drehen des PDAs wird das Display angepasst und die Information entweder im Hoch oder Querformat angezeigt. Durch Bewegen des Gerätes nach links/rechts oder oben/unten wird der Inhalt des Displays in eben dieselbe Richtung gescrollt. Die verschiedenen Möglichkeiten der neuen Interaktion mit einem PDA wurden auch mit Usern ausgetestet. Probleme bereiten bei neuen Anwendungen die noch unbekannte Handhabung damit. Auch gibt es manchmal noch mit den einzelnen Sensoren Probleme - zum Beispiel die doppelte Verwendung des Tilt-Sensors. Teilweise ist schwer zu erkennen, ob der User das Display nur scrollen oder die Ansicht von Quer- auf Hochformat ändern will.

Das Creative Histories Projekt [CrHi] [BaKu05] versucht, neue Interaktionsmöglichkeiten bei Mobiltelefonen und PDAs zu den bisher vorhandenen hinzuzufügen. Dadurch, dass die bisher vorhandenen Interaktionsmöglichkeiten, die Eingabe über die Tastatur, noch sehr beschränkt sind, ist es interessant, dem Benutzer durch neue Möglichkeiten die Interaktion zu erleichtern. In einem Handy ist heute bereits genug Rechenkapazität vorhanden, um etwa 3D-Modelle von der Umgebung anzeigen zu können. Nur ist es ziemlich schwer, mit den vorhandenen Eingabegeräten darin zu navigieren. Zu diesem Zweck wurden bei dem Projekt *Creative Histories* spezielle Orientierungssensoren eingebaut. Es wird versucht, die Bewegungen des Benutzers in der realen Welt einzufangen und in der virtuellen Welt nachzubilden. So macht es einen Unterschied, ob der User vor einem Gebäude steht und in die Höhe,

nach unten oder auf eine Seite sieht. Durch die integrierten Sensoren kann dies festgestellt und die virtuelle Welt auf dem Handy angepasst werden. So kann nun in der virtuellen Welt auf dem Display ein Gebäude beispielsweise so angezeigt werden, wie es vor 100 Jahren ausgesehen hat, oder vielleicht auch zukünftige Bauvorhaben, sofern Modelle davon zur Verfügung stehen.

UMAR Ubiquitous Mobile Augmented Reality [HeOl04] ist ein konzeptionelles Framework für die Darstellung von erweiterten Realitäten (AR = Augmented Reality) auf PDAs und Handys. Manche Forschungsprojekte wie etwa das weiter oben vorgestellte *CyPhone* arbeiten zur Darstellung von 3D-Umgebungen mit sogenannten HMDs (Head-Mounted-Displays), also Anzeigen, die wie ein Hut getragen und vor die Augen gehalten werden. Hierauf wird bei UMAR verzichtet, da die modernen PDAs schon genug Rechenkapazität zur Verfügung stellen, um die erweiterte Realität auf dem Display darzustellen. Zudem wird davon ausgegangen, dass das PDA über eine Kamera verfügt. Das PDA ist mittels WLAN zu einem Server verbunden, der die erweiterte Realität berechnet. Der Videostream von der Kamera wird via WLAN zum Server geschickt und dort verarbeitet. So kann festgestellt werden, auf welches Gebäude der User gerade blickt. Der AR-Stream wird zum PDA zurückgeschickt und dort auf dem Display dargestellt. Auf dem Server läuft ein ARToolKid, um die erweiterte Realität zu berechnen [ArKi]. Ein Nachteil von diesem System ist, dass der User eine schnelle Verbindung zum Server benötigt. Wenn kein WLAN zur Verfügung steht, muss der Video-Stream mittels GPRS/UMTS verschickt werden, wo allerdings pro übertragenes Byte oder pro Minute Kosten verrechnet werden.

Mit der Vorstellung des „Point 2 Discover“ Prjektes welches im engen Zusammenhang mit dieser Diplomarbeit steht wird nun der Überblick über die Verknüpfung von digitalen Informationen mit der physischen Realität abgeschlossen. *„p2d“ – „Point 2 Discover“ [HPp2d][SiFr06]* ist ein Projekt, welches das Handy als interaktives Zeigegerät verwendet. Mit einem „p2d“ -Handy ist es möglich, Information von der Umgebung einzuholen, in der man sich gerade befindet. Durch einfaches Zeigen auf eine Busstation kann ermittelt werden, wann die nächsten Busse abfahren. Eine weitere Möglichkeit besteht darin, durch Hinzeigen auf eine Sehenswürdigkeit Informationen darüber zu erhalten. Wird beim Schifahren auf einen Berg gezeigt, kann dessen Namen angezeigt werden. Es ist mit „p2d“ auch möglich, bei Preisausschreiben teilzunehmen, wenn auf eine Anzeigetafel gezeigt wird.

Der Großteil der im Internet verfügbaren Informationen sind einem geographischem Ort zugewiesen, wie etwa historische Information zu einem Gebäude oder der Fahrplan zu einer Bushaltestelle. Die Anwendung „p2d“ ist sich der Umgebung bewusst und stellt die gewünschten Informationen zu den geographischen Punkten zur Verfügung.

Die „p2d“ Anwendung besteht aus verschiedenen Komponenten:

- **P2d Hardware Sensor Module:** Das Sensormodul besteht aus einem digitalen, magnetischen Kompass, einem Tiltsensor und einem GPS-Modul. Die Genauigkeit der vom GPS-Modul gelieferten Daten kann mittels Differential-GPS verbessert werden. Diese Sensoren können via Bluetooth mit dem mobilen Gerät verbunden werden. Mittels des DGPS-Sensors kann die genaue Position des Handys festgestellt werden. Die Orientierungssensoren ermitteln die aktuelle Zeigerichtung des Handys.
- **P2d Server Platform:** Anhand der vom Sensormodul erhaltenen Daten wird ein Umgebungsmodell erstellt. Hier wird ermittelt, welche Gebäude oder Objekte der Benutzer im Blickfeld hat. Oft werden Gebäude von anderen Objekten verdeckt sein, sodass der Benutzer sie nicht sehen kann. Die technische Machbarkeit unter Betrachtung der Limitierung der Sensorengenauigkeit und Verzögerungen beim Datentransfer über das Netzwerk soll diese Plattform gewährleisten.

- **P2d Anwendungen:** Rund um die Plattform und das Sensormodul werden eine Reihe von Anwendungen entwickelt. Angefangen beim einfachen Zeigen auf das gewünschte Objekt bis hin zu einer erweiterten Realität, wo die interessanten Punkte über das Bild, das von der Kamera des Handys geliefert wird, gelegt werden. Ebenso gibt es Tools und Dokumentation, die anderen Entwicklern erlauben, eigene „p2d“ Anwendungen zu entwickeln.

Abbildung 2: Das „p2d“ Handy wird benutzt, um sich Informationen zu Gebäuden in der näheren Umgebung zu holen.

1.4.2 Auditive Interaktion in mobilen räumlich bewussten Anwendungen

Bei der hier vorliegenden Diplomarbeit „Point and Speak“ wird mit dem Benutzer hauptsächlich aber nicht ausschließlich sprachlich kommuniziert. Ein Interface, das mehrere Eingabemöglichkeiten zur Verfügung stellt, wird als multimodales Interface bezeichnet. Es gibt bereits einige vorhandene Arbeiten, die sich mit sprachlichen und multimodalen Interfaces und akustischen Ausgaben auseinandersetzen.

Sharon Oviatt beschreibt in [Ov03] multimodale Interfaces. Verschiedene Eingabemöglichkeiten, wie etwa Sprache, Pen, Gesten, Kopf- und Körperbewegungen ermöglichen eine neue Art von Programmentwicklung. Sobald nicht nur eine Inputvariante berücksichtigt wird sondern mindestens zwei, wird das Interface als multimodal bezeichnet. Eingabekombinationen wie Sprache und Pen, Sprache und Lippenbewegung oder Sprache und Handbewegungen werden bei multimodalen Systemen kombiniert. Eine der ersten multimodalen Anwendungen gab es bereits Anfang der 80er, wo mittels Sprache und Navigieren auf einem großen 2D Bildschirm Objekte bewegt und erzeugt werden konnten. Mit der Weiterentwicklung der Spracherkennung Ende der 80er wurden immer mehr Systeme mit einer alternativen sprachlichen Eingabe entwickelt. Mit Hilfe des Georal-Systems konnten sich Touristen durch einen berührungssensitiven Bildschirm und Spracheingabe Reiserouten planen lassen. Man ging dann von der Entwicklung dieser einfachen Multimodalität, wie auf etwas zeigen und sprechen, über in die Entwicklung komplexerer Multimodalität. Es wurde beispielsweise versucht, Spracheingabe durch den Vergleich der Lippenbewegung mit dem auditiven Signal zu ermitteln. Ziel der Entwicklung multimodaler Interfaces ist es, je nach Bedarf die entsprechenden Eingabemöglichkeiten zu nutzen. Maus und Tastatur sind nicht immer die geeignetsten Eingabegeräte. Ein System ist flexibler und kann von mehr Benutzeren verwendet werden, wenn mehrere Ein- und Ausgabemöglichkeiten zur Verfügung stehen.

Design and Development of an Indoor Navigation and Object Identification System for the Blind [HuDi04] ist ein neues System welches entwickelt wurde, um blinde Menschen bei der Navigation im Gebäudeinneren zu unterstützen. Mittels eines Sensormoduls, das wie eine Taschenlampe gehandhabt werden kann, wird im dreidimensionalen Raum navigiert. Das Sensormodul ist mit einem tragbaren Computer

verbunden, welches einen dreidimensionalen Plan der Umgebung gespeichert hat. Das Sensormodul besteht aus verschiedenen Sensoren wie etwa einer Stereo-Kamera, Richtungssensoren und auch einem elektronischem Kompass. Durch das Zusammenwirken der Sensoren mit dem Umgebungsplan kann eine genaue Positionierung des Users vorgenommen werden. Wird einem Blinden selbst in großen Gebäuden längere Zeit gegeben, um es genau zu erkunden, ist oft erstaunlich, mit welchem Detailgrad er das Gebäude beschreiben kann. Er orientiert sich dabei an Charakteristika des Raumes, wozu Gänge, Türen aber auch Steckdosen und jede Art von Einrichtung gehören können. Für blinde Menschen ist es besonders wichtig zu wissen, wo sie sich gerade befinden. Von der aktuell bekannten Position wird dann ein Weg gesucht, um das gewünschte Ziel zu erreichen. Dazu ist die genaue Information über Hindernisse und Gefahren notwendig. Zu solchen Gefahren gehören Aufzüge, Treppen, selbst schließende Türen und noch viele andere Dinge. Diese Gegenstände werden aus dem Umgebungsplan im Zusammenspiel mit den Sensoren extrahiert und mittels akustischer Ausgabe entweder über Lautsprecher oder Ohrhörer dem Benutzer erklärt.

Ein sehr interessantes Projekt ist *gpsTunes [StEs05]*, welches auf einem Pocket PC implementiert ist. Dieser enthält GPS-Sensoren, Orientierungssensoren und einen mp3-Player. Es wird hier gezeigt, dass es möglich ist, den User mittels Veränderung der Lautstärke und Schwenken der Musik in die gewünschte Richtung zu dirigieren und ihn den Weg zum Ziel finden zu lassen. Das Ziel kann auf einer Karte angeben werden. Je näher sich der User diesem Punkt nähert, desto lauter wird die Musik. Am Punkt angelangt, hat die Musik die vom Benutzer bevorzugt Lautstärke und das Erreichen des Zieles wird noch zusätzlich mit einem speziellen Ton (etwa einem Piep) angegeben. Durch Orientierungssensoren kann die Position des Kopfes ermittelt werden. Wenn sich der Benutzer dreht, wird nun auch das Musikstück mitgeschwenkt, sodass der User hört, in welcher Richtung sich das Zielobjekt befindet. Als Problem taucht bei diesem Projekt auf, das sich die GPS-Sensoren beim Test als nicht sehr genau und zuverlässig gezeigt haben. Vor allem in dicht bebauten Städten gibt es oft keine gute Sichtverbindung und dadurch Verzögerungen und Ungenauigkeiten des GPS-Signales. Um diese Probleme zu lösen, werden zusätzliche Sensoren wie etwa Beschleunigungsmesser, Gyroskop und Magnetfeldstärkmessgerät eingesetzt. Wie diese Sensoren optimal zusammenarbeiten wird zurzeit noch geforscht.

[Lai2004] beschreibt eine Studie von *Jennifer Lai,* die ein unimodales System mit einem multimodalen System vergleicht. Als Testablauf wird der Zugang zu Emails herangenommen. Die Testpersonen versuchen zuerst durch Tasteneingabe und eine graphische Ausgabe Emails zu lesen. Das multimodale System verwendet sowohl Spracheingabe als auch Eingabe über die Tastatur. Die Ausgabe findet sowohl graphisch als auch akustisch statt. Die Tester sahen einen größeren Nutzen im multimodalen System. Sie gingen dabei großteils so vor, dass sie zuerst die Spracheingabe wählten. Nur wenn diese auch nach mehrmaligen Wiederholungen falsch funktionierte, wurde auf die alternative Eingabemöglichkeit, die Tastatur, zurückgegriffen.

Für viele Menschen ist es üblich, mehrere hundert Emails am Tag zu bekommen. Der Empfänger geht dann normalerweise so vor, dass er die Emails sortiert, etwa nach Absendern oder interessanten Topics. Diese Auswahl wäre nur schwer zu treffen, könnte nur auf akustische Ausgabe zurückgegriffen werden. Aus diesem Grund ist ein multimodales Ausgabesystem von großer Bedeutung. Laut [Lai2004] kann es als das erste System für Handys angesehen werden, das multimodale Interaktionsmöglichkeiten, also sowohl Sprache als auch Text, unterstützt. Andere Systeme nutzen entweder Sprache oder das GUI, niemals aber beides. Am Ende der Studie gaben die Testpersonen an, dass das multimodale System sowohl benutzerfreundlicher, hochwertiger und auch interessanter ist.

Moderne Systeme greifen oft auf mehr Eingabemöglichkeiten zurück. Zum Beispiel werden bei der Benutzeridentifikation oder Benutzerverifikation Fingerabdrücke, Retinalscans, Handschrift und Sprache kombiniert.

Rainer Wasinger hat ein Navigationssystem *[WaOl03] [WaSt03] [WaStKr03]* für Fußgänger entwickelt, welches die Suche einer Route im Internet ermöglicht um diese dann auf ein PDA zu laden. Die Route kann entweder im Freien oder auch in Gebäuden sein. Via GPS wird der Standort im Freien ermittelt und mit Hilfe von Infrarot-Ortung wird die Position des Benutzers innerhalb von Gebäuden festgestellt. Hierfür wird eine genaue Beschreibung der Innenräume benötigt, die allerdings von Hand gefertigt werden muss.

Diese Anwendung verwendet sowohl Sprach- und Soundausgabe als auch graphische und textuelle Ausgabe. Der User sieht auf der graphischen Ausgabe einen Plan der Umgebung mit den Gebäuden und Straßen mit textuellen Beschreibungen dazu. Die

Sprachausgabe kann nun Befehle wie etwa „Gehen Sie 100 Meter und biegen Sie dann links in die Mühlgasse ab!“ oder auch Gebäudebeschreibungen ausgeben wie „Das Gebäude Nr.15 ist eine Mühle aus dem 16. Jahrhundert.“. Andere Ausgaben wie System-Pieps oder der Name der nächsten Querstraße unterstützen den Output. Die Spracheingabe und Sprachausgabe wurde mit „IBM Embedded ViaVoice speech synthesizer and recognizer“ implementiert. Der Benutzer des Systems kann verschiedene Parameter der Sprachausgabe anpassen, wie beispielsweise verschiedene Sprachen und Dialekte, Tonhöhe, Geschwindigkeit und Lautstärke. Auch die Prosodie kann verändert werden, wenn erwünscht können Pausen eingefügt und Betonungen gesetzt werden. Für die Instruktionen und Beschreibungen können verschiedene Sprachausgaben durch die jeweiligen Parameter bestimmt werden. Auch die Soundausgabe kann modifiziert werden. Mittels Tönen wird signalisiert, wie weit von einem Ziel entfernt sich eine Person aufhält und zusätzlich wird auch auf Sehenswürdigkeiten hingewiesen. Die Ausgabe von Tönen alleine reicht nicht, da keine Beschreibungen dargestellt werden können. Hierfür ist die Sprachausgabe zuständig.

Der Forschungsbericht geht noch genauer darauf ein, wie Sprachausgabe möglichst verständlich und natürlich dem Benutzer übermittelt werden kann. Zum Beispiel werden in dem Satz „Gehen Sie 100 Metern und biegen Sie dann links in die Mühlgasse ab!“ nicht alle Wörter gleich schnell gesprochen. Wichtige Informationen wie „100 Meter“, „links abbiegen“ und „die Mühlgasse“ werden langsamer gesprochen, um es dem Benutzer leichter zu machen, die relevanten Informationen zu extrahieren.

Es wurde auch festgestellt, dass die Erkennungsrate von synthetischer Sprache bei älteren Personen nachlässt. Dem kann entgegengewirkt werden, indem die Lautstärke erhöht und die Geschwindigkeit reduziert wird.

Psychologische Studien haben herausgefunden, dass Personen, die rein synthetische Sprache hören, versuchen, dieser eine Persönlichkeit zuzuordnen. Besonders Stimmen, die der eigenen in Sprachfrequenz, Grundfrequenz oder Intensität ähneln, werden als sympathisch wahrgenommen. Durch Verbinden dieser Eigenschaften des Benutzer mit der Sprachausgabe kann Vertrauen und Gefallen hervorrufen werden.

Bei *A MultiModal Architecture for Cellular Phones [NaOr04]* wird der Versuch unternommen, dem User multimodale Interaktion zu ermöglichen indem Spracheingabe und Sprachausgabe als auch direkter Input am Handy (in diesem Fall einem Ericsson P900) ermöglicht werden. Ein interessanter Aspekt dieses Projektes ist die serverseitige

Übernahme der Sprachverarbeitung. Das Sprachsignal und auch der normale Input werden via GPRS an einen Server geschickt und dort ausgewertet. Die GUI wird dann vom Server aktualisiert. Als Grund, warum die Sprachverarbeitung nicht auf dem Handy selbst gemacht wird, wird die große Speicherbelastung, verursacht durch die Sprachgrammatik zum Erkennen des Sprachinputs, genannt. Bei komplizierten Grammatiken und für die Erkennung mehrerer Sprachen kann diese viel Speicherplatz einnehmen. Als Nachteil sei zu erwähnen, dass bei Unterbrechung des GPRS-Datenstroms keine Befehle mehr verarbeitet werden können.

1.4.3 Möglichkeiten der synthetischen Sprachausgabe

Die Studie *Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction. [NaLe00]* versucht herauszufinden, ob Menschen beim Hören computererzeugter Stimmen auf dieselbe Art reagieren wie bei einem menschlichen Sprecher. Die Teilnehmer dieser Studie hörten Buchbesprechungen auf einer Webseite. Statt der textuellen Beschreibung des Buches konnte ein .wav-File angehört werden. Die Teilnehmer registrierten die durch den Computer synthetisierten Eigenschaften wie etwa ein introvertierter Sprecher im Vergleich zu einem extrovertierten Sprecher. Extrovertierte Sprecher sprechen schneller, lauter, in einer höheren Tonlage und variieren diese auch mehr als introvertierte Sprecher. Durch Anpassung der synthetisierten Stimme an die eigene (durch Geschwindigkeit, Sprachgrundfrequenz, …) wurde die Sprachausgabe als zusehends attraktiver angesehen. Obwohl die Sprache alleine nicht Aufschluss über den Sprecher geben kann, versucht der Mensch, da hier andere Merkmale wie das Aussehen oder die Gestik fehlen, den Sprecher alleine durch seine Stimme zu bewerten. Die Studie fand heraus, dass die Menschen dahingehend positiv beeinflusst werden können, ein Buch zu kaufen, wenn die Stimme des Computers der des eventuellen Käufers anpasst wird. Die Buchbesprechung wurde positiver aufgenommen, der Sprecher als vertrauenswürdiger angesehen und das Buch wurde eher gekauft.

Die Sozialpsychologie zeigt in der Studie *Can Computer-Generated Speech Have Gender? An Experimental Test of Gender Stereotype [LeNa00]*, dass das Geschlecht Unterschiede bei der Mensch-Mensch Interaktion macht, zum Beispiel bei der Beeinflussung von Gesprächspartnern. Ein männlicher Überredner erweckt mehr Konformität bei seinem Gegenüber und zusätzlich werden einem männlichen Sprecher mehr Kompetenz und ein höherer sozialer Status zugesprochen. In dieser Studie wird nun untersucht, ob geschlechtsspezifische synthetische Sprachausgabe diese geschlechtsspezifischen Merkmale auslösen kann, vor allem in Beziehung zur Überredungskunst. Die Testpersonen konnten so gut wie immer unterscheiden, ob ein männlicher oder weiblicher Sprecher simuliert wird. Dem männlichen Sprecher wurden dann automatisch männliche Eigenschaften zugesprochen, auch wenn die Stimmen denselben Inhalt sprachen. Die Teilnehmer der Studie haben den Vorschlägen der männlichen Computerstimme eher zugestimmt als den Vorschlägen der weiblichen. Allerdings befanden sie die Stimme, die dem eigenen Geschlecht entsprach, als die angenehmere im Vergleich zu der des anderen Geschlechts. Die Testpersonen ordnen also den computererzeugten Stimmen eben jene Eigenschaften zu, die sie auch menschlichen Sprechern zuordnen würden.

1.5 Aufgabenstellung und Forschungsfragen

Die Aufgabe dieser Diplomarbeit ist es, einem orts- und orientierungsbewussten System eine sprachliche Komponente hinzuzufügen. Es soll herausgefunden werden, wie eine sprachliche beziehungsweise multimodale Interaktionsmöglichkeit auf bestem Weg zu gestalten ist. Eine interessante Frage ist, inwiefern mit den aktuellen technischen Mitteln und Möglichkeiten der automatischen Spracherkennung (ASR) und der Sprachsynthese (TTS) ein solches Vorhaben verwirklichen werden kann. Die technischen Möglichkeiten bei Pocket PCs beziehungsweise Handys entwickeln sich rasch, allerdings sind sie dennoch begrenzt. Automatische Spracherkennung und synthetische Sprachgenerierung sind sehr komplexe Anwendungen, die hohe Anforderungen an die Hardware stellen. Viel Speicherplatz und eine ausreichende Verarbeitungsgeschwindigkeit sind gefragt. Inwiefern die zur Verfügung stehenden mobilen Geräte diese Ansprüche erfüllen, muss im Zuge der Diplomarbeit genauso herausgefunden werden, wie auch ein möglichst einfaches und intuitives Interaktionsdesign entwickelt werden muss.

Es ist ein natürliches Handeln, auf ein Gebäude zu zeigen und zu fragen, wann es erbaut worden ist. Bisher standen dafür ein Reiseführer oder auch Bekannte oder Freunde zur Verfügung. Beinahe jeder führt heutzutage ein Handy mit sich. Warum sollte dieses diese Aufgabe nicht übernehmen können? Da bei Handys auch die Voraussetzungen für Sprachkommunikation gegeben sind, ist das Projekt „Point 2 Discover“ eine hilfreiche Informationsquelle und das „Point and Speak“ Interface eine sinnvolle Ergänzung zu „p2d“. Die Möglichkeit, Informationen graphisch darzustellen, ist bei Handys aufgrund der Displaygröße eine sehr begrenzte. Dadurch, dass bei der Anwendung des Systems mit dem Handy auf ein Gebäude gezeigt werden muss, ist es sowieso schwer, auf das Display zu sehen. Das macht ein Sprachinterface noch interessanter.

Um eine Übersicht über die Fragen zu bekommen, die bei der Entwicklung eines Sprachsystems beantwortet werden, werden diese im Folgenden aufgelistet. Sie sind in verschiedene Bereiche untergliedert: Sprachausgabe, Spracheingabe, soziale Komponenten und auf Usability bezogene Fragen. Bei jeder Frage wird darauf hingewiesent, in welcher Phase der Diplomarbeit diese beantwortet wird.

1.5.1 Sprachausgabe

Frage 1:

Soll der User mittels räumlicher Sprachausgabe darauf hingewiesen werden, in welcher Richtung sich das Gebäude befindet und wie weit es weg ist?

Ad 1) Der erste Benutzertest, der im nachfolgenden Kapitel (Kapitel 3) beschrieben wird, versucht auf diese Frage eine Antwort zu finden. Eine Teilaufgabe simuliert eine räumliche Audioausgabe. Je nachdem, wo sich das Gebäude befand, wurde mit Hilfe der Lautstärke und interauralen Signalen dessen Standort simuliert.

Frage 2:

Soll eine Erklärung ausgeblendet und eine andere einblendet werden, wenn sich der Benutzer einem anderen Gebäude zuwendet?

Ad 2) Wird im ersten Benutzertest, der Anforderungsanalyse, beantwortet. Es gibt dort eine entsprechende Frage, die die Testpersonen beantworten müssen.

Frage 3:

Sind unterschiedliche sprachliche und visuelle Textausgaben zu viele Informationen auf einmal?

Ad 3) Diese Frage wird bei der abschließenden Benutzerevaluierung (Kapitel 5) zu beantworten versucht. Bei einem Task werden zu den visuellen zusätzlich ergänzende Informationen akustisch ausgegeben.

Frage 4:

Sollen akustische Hinweise bei verfügbaren Informationen, oder gleich die Informationen ausgegeben werden?

Ad 4) Die Beantwortung dieser Frage erfolgt in der Anforderungsanalyse durch den ersten Benutzertest.

Frage 5:

Wird eine weibliche oder männliche Stimme bevorzugt?

Ad 5) In der abschließenden Benutzerevaluierung werden der Testperson eine weibliche und eine männliche Sprachausgabe vorgestellt. Die Testperson kann dann begründen, welche ihr warum besser gefällt.

Frage 6:

Wie kann mit den zurzeit verfügbaren technischen Möglichkeiten die Audioausgabe möglichst interessant gestaltet werden?

Ad 6) Im Laufe der Implementierung wird sich zeigen, welche technischen Möglichkeiten zur Verfügung stehen und wie gut künstliche Sprache bis jetzt synthetisiert werden kann.

1.5.2 Spracheingabe

Frage 7:

Wie ist die Spracheingabe zu gestalten?

Ad 7) Auf die Gestaltung der Spracheingabe wird im ersten der beiden Benutzertests Bezug genommen. Die Testpersonen werden gefragt, wie sie mit dem Gerät vorzugsweise kommunizieren möchten.

Frage 8:

Gibt der Benutzer kurze Sprachbefehle oder spricht er ganze Sätze um mit der Anwendung zu kommunizieren?

Ad 8) In welcher Form der Benutzer mit einem Sprachinterface interagieren will, wird sowohl bei der Anforderungsanalyse als auch bei der abschließenden Benutzerevaluierung festgestellt.

Frage 9:

Welche Befehle sollen möglich sein?

Ad 9) Auf die verwendeten Befehle wird in beiden Benutzertests Bezug genommen. Der erste Benutzertest versucht vorab abzuklären, welche Befehle der zukünftige User dem System sagen will. Der abschließende Test stellt dann schon eine Spracherkennung zur Verfügung, die eine Reihe von Befehlen erkennen und ausführen kann. Hier wird untersucht, ob diese Befehle den Benutzern genügen.

Frage 10:

Ist ein reines Sprachinterface sinnvoll, ohne zusätzliche visuelle Ausgabe?

Ad 10) Bei der abschließenden Benutzerevaluierung lernt die Testperson ein visuelles, ein multimodales und ein rein sprachliches Interface kennen. Welche Interatkionsmöglichkeit am Besten aufgenommen wird, wird im fünften Kapitel dieser Diplomarbeit beschrieben.

Frage 11:

Wie müsste ein Dialog gestaltet sein, um dem Benutzer zu gefallen?

Ad 11) Der Test von Kapitel 5 zeigt, ob das bereits implementierte Dialogsystem den Anforderungen der Benutzer genügt.

Frage 12:

Soll Bargein möglich sein? Unter Bargein versteht man die Möglichkeit der Unterbrechung der sprachlichen Ausgabe.

Ad 12) Zuerst wurde bei der Implementierung untersucht, ob Bargein technisch möglich ist. Ob Bargein notwendig ist, zeigt der zweite Benutzertest. Dort ist es der Testperson möglich, Ausgaben zu unterbrechen.

Frage 13:

Gibt es einen Unterschied bei der Erkennung mit und ohne Headset?

Ad 13) Im abschließenden Benutzertest muss die Testperson verschiedene Aufgaben erledigen, wobei sowohl Erkenungen mit als auch ohne Headset ausprobiert werden.

Frage 14:

Soll der Benutzer durch die Betätigung einer Taste/eines Buttons die Spracheingabe starten, oder soll ständig auf mögliche Benutzereingaben gelauscht werden.

Ad 14) Hierzu werden zwei unterschiedliche Prototypen entwickelt, die beide den Testpersonen im abschließenden Benutzertest vorgestellt werden. Es ist hier auch interessant, ob eine gute Erkennungsrate bei ständigem Lauschen auf den Benutzer erreicht wird.

Frage 15:

Ähnliche Projekte lassen die Sprachverarbeitung von einem externen Server verrichten. [NaOr04] Benötigen die verschiedenen Komponenten (Sprache, GUI, Sensoren) zu viel Rechenkapazität für die momentan verfügbaren Pocket PCs und Handys?

Ad 15) In Kapitel 3 kann im Zuge der Implementierung der Sprachsoftware diese Frage beantwortet werden.

1.5.3 Soziale Komponenten

Frage 16:

Fühlen sich die Benutzer belästigt, falls neben ihnen jemand mittels Spracheingabe mit dem Handy kommuniziert?

Ad 16) Diese Frage wird im ersten Benutzertest geklärt. Den Testpersonen wird eine entsprechede Frage gestellt.

Frage 17:

Stört es, wenn Informationen sprachlich ausgegeben werden?

Ad 17) Auch diese Frage wird in der Anforderungsanalyse behandelt.

Frage 18:

Schafft die Verwendung eines Headset Abhilfe gegen Belästigung durch Sprachausgabe oder Spracheingabe?

Ad 18) Wird auch in der Anforderungsanalyse geklärt.

Frage 19:

Fühlt sich der Benutzer überwacht, wenn das Programm ständig auf seine Eingaben wartet?

Ad 19) Wird im abschließenden Benutzertest durch eine entsprechende Frage beantwortet.

1.5.4 Fragen zur Benutzerfreundlicheit

Frage 20: Können die Benutzer selbständig im Programm navigieren?

Ad 20) Im abschließenden Benutzertest bekommt der Benutzer eine fertige Anwendung. Mithilfe einer kurzen akustischen Programmeinführung wird dem Benutzer erklärt, wie das Programm funktioniert. Ob diese vom Programm vorgelesene Einführung ausreichend ist, klärt die abschließende Benutzerevaluierung.

Frage 21: Ist das Feedback ausreichend?

Ad 21) Auch darüber gibt der abschließende Benutzertest Auskunft.

1.6 Zusammenfassung

Mit der Minituarisierung von technischen Komponenten und der damit zusammenhängenden gesteigerten Leistungsfähigkeit von Mobiltelephonen entstanden viele neue Möglichkeiten. Der Einbau von zusätzlichen Sensoren in mobilen Geräten ermöglichte die Generierung und Realisierung einer Reihe interessanter Ideen. Einige davon wurden im ersten Kapitel vorgestellt wie beispielsweise Semapedia, die einen anderen Ansatz gefunden haben, um digitale Information mit der physischen Realität zu verknüpfen. Es wurde auch zusammengefasst, welche Anwendungen in Richtung Sprachtechnologie bei mobilen Geräten bereits existieren.

Das Projekt „Point 2 Discover“, welches eng mit der hier vorliegenden Diplomarbeit „Point and Speak“ zusammenhängt, wurde vorgestellt und auch die Idee, die hinter der Diplomarbeit „Point and Speak“ steckt, wurde näher erklärt. Es tauchten bei der Entwicklung der Idee dieser Diplomarbeit eine große Anzahl von Fragen auf. Diese Fragen, die es im Laufe dieser Diplomarbeit zu beantworten gilt, wurden am Ende dieses Kapitels aufgelistet.

Im nun folgenden Kapitel wird der Benutzertest beschrieben, der am Anfang der Entwicklung stand und einen Einblick auf die Wünsche der zukünftigen Benutzer gegeben hat.

2 Anforderungsanalyse

2.1 Überblick

Im zweiten Kapitel wird der Benutzertest beschrieben, der für das Projekt „p2d“ unter Einbeziehung des Projektes „Point and Speak“ durchgeführt wurde. Anhand eines für den Benutzertest entwickelten Prototyps konnte den Usern eine Vorstellung von dem Projekt „p2d“ gegeben werden. Dieser Prototyp wird im Laufe des zweiten Kapitels genauso beschrieben wie der Benutzertest selbst. Anhand dieses Testes können schon einige der im vorigen Kapitel genannten Forschungsfragen beantwortet werden. Bei der Zusammenfassung der Testergebnisse wird speziell auf die gestellten Fragen eingegangen. Auch der genaue Testablauf wird beschrieben, sodass eine bessere Vorstellung von der Testsituation möglich ist.

2.2 Einleitung

Das Projekt „Point and Speak“ wurde benutzerorientiert entwickelt. Hierfür wurde vom ftw. ein Benutzertest mit zwölf Personen, welcher nicht alleine für das Projekt „Point and Speak“ sondern im Zusammenhang mit dem Projekt „Point 2 Discover“ durchgeführt wurde. Die Tests wurden mit Erlaubnis der getesteten Personen für eine bessere Auswertung akustisch und visuell aufgezeichnet. Es wurde auch laufend auf einem Testplan mitprotokolliert, welche Antworten die Testperson auf die Fragen gegeben hat. Der Test wurde auch schon zusammengefasst und veröffentlicht [FrSi06]. Im Rahmen dieser Diplomarbeit wurde der Prototyp entwickelt und der Testplan um den Teil, der „Point and Speak“ betrifft, ergänzt. Auch an der Testdurchführung wurde mitgearbeitet, großteils als „Wizard of Oz“ - Operator.

Das Ziel dieses Tests war es, den zukünftigen Benutzer in die Entwicklung miteinzubinden. So wurden Informationen eingeholt, was die zukünftigen Anwender von so einem Programm erwarteten oder wünschten und wie die Interaktion am Besten zu gestalten war. Die Bedienung sollte für die Personen möglichst intuitiv und einfach sein. Es war wichtig, eine andere Sicht auf das Projekt zu bekommen als die von der Warte des Entwicklers. Ein Entwickler hat einen sehr technischen Blick auf das aktuelle Projekt und vergisst oft, auf die späteren Anwender Rücksicht zu nehmen.

2.3 Prototyp

Der Prototyp war eine in Macromedia Flash 8 entwickelte Anwendung, die auf einem Pocket PC lief. Das Testgerät bei dieser Benutzerevaluation war der VPA IV von Vodafone. Die Flashanwendung wurde darauf installiert.

2.3.1 VPA IV

Abbildung 3: VPA IV von Vodafone

Quelle: www.expansys.com

Der VPA IV von Vodafone wurde gewählt, weil er die für die Tests erforderlichen Ressourcen zur Verfügung stellte. Er hatte ein großes (480x640) Display und verfügte über genügend Rechenkapazität. Zusätzlich hatte er ein integriertes WLAN, was für den Test von großer Bedeutung war. Die Audioausgabe erfolgte über ein Stereo-Headset.

2.3.2 Programm

Die Anwendung war eine Flashkomponente. Sie konnte mittels Internet Explorer auf dem Pocket PC ausgeführt werden. Da noch keine Orientierungs- oder GPS-Sensoren verfügbar waren, mussten diese simuliert werden („Wizard of Oz“-Ansatz). Eine peer2peer WLAN Verbindung wurde von dem Pocket zu einem bei den Benutzertests mitgeführten Laptop hergestellt. Mittels „Remote Display Control for Pocket PC“ von Microsoft konnte die visuelle Ausgabe des Pocket PCs auch auf dem Laptop angezeigt werden. Zusätzlich zur Anzeige des Displays wurden auch Tastenbefehle und Touchscreenkommandos auf den Pocket PC übertragen. Der simulierte Touchscreen funktionierte allerdings nur unzureichend im Zusammenspiel mit dem Macromedia Flash Programm. Deshalb wurde das Programm ausschließlich per Tastatureingabe remote gesteuert. Es folgt eine kurze Auswahl von Tastenkürzel und den dadurch ausgelösten Ereignissen.

Tastenkürzel	Ausgelöstes Ereignis
1	Ansicht wechseln auf Radarmodus
2	Ansicht wechseln auf Karte
3	Ansicht wechseln auf erweiterte Realität, Standort 2
4	Ansicht wechseln auf erweiterte Realität, Standort 1
7	Ansicht wechseln auf Listenmodus
L	Radar nach links rotieren
K	Radar nach rechts rotieren
Pfeiltasten	Karte / Erweiterte Realität scrollen
P	Erweiterte Realität schnell nach rechts scrollen
O	Erweiterte Realität schnell nach links scrollen
Ö	Akustische Information zu Indoor-Golf links abspielen
Ä	Akustische Information zu Indoor-Golf rechts abspielen
0	Audioausgabe zu den POI aufdrehen/abdrehen

Tabelle 2: Ausgewählte Befehle zurRemotesteuerung der Anwendung

Es folgt nun eine Beschreibung der implementierten Darstellungsmöglichkeiten der Umgebung, zwischen denen mit Hilfe der oben genannten Befehle gewechselt werden konnte.

2.3.3 Darstellungsarten der Umgebung

Die folgenden Abbildungen repräsentieren die verschiedenen Darstellungsarten der Umgebung, die der Testperson vorgeführt wurden.

Abbildung 4: Listenansicht
Darstellung der in der Nähe befindlichen interessanten Gebäude in Form einer List.

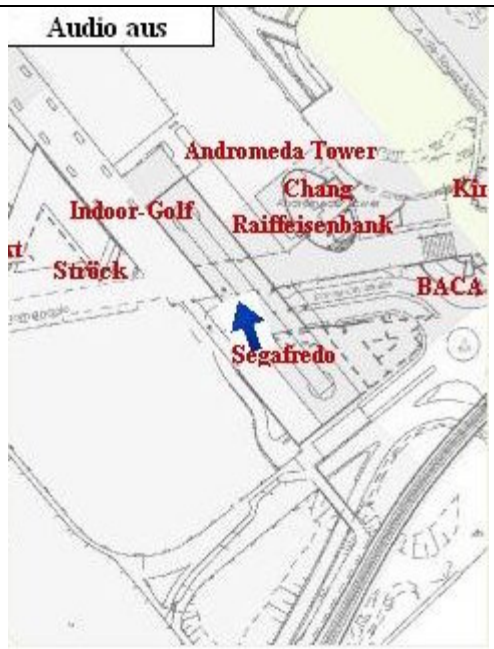


Abbildung 5: Kartenansicht
Eine Karte auf der die interessanten Punkte eingezeichnet sind.

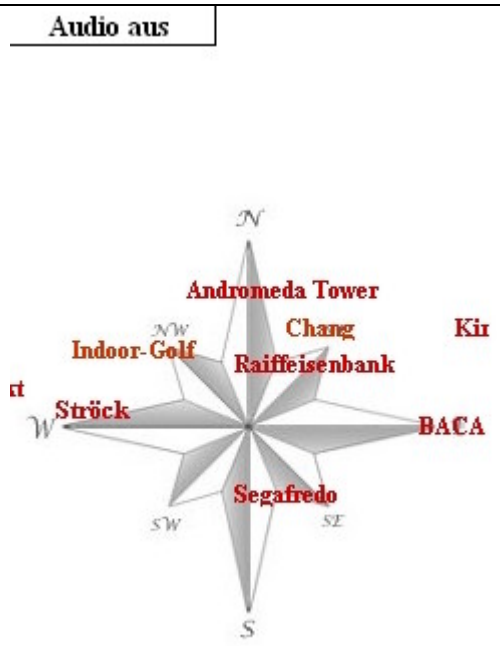


Abbildung 6: Radaransicht
Eine Kompassrose, auf der die interessanten Punkte eigetragen waren.

Abbildung 7: Erweiterte Realitätsansicht
Das von der integrierten Kamera gelieferte Bild, überlagert mit den interessanten Punkten.

Als einfachste Darstellungsart wurde die Liste (*Abbildung 4: Listenansicht*) konzipiert. Die interessanten Objekte der Umgebung (POI = points of interest) wurden der Reihe nach aufgelistet. Die Liste enthielt weder Informationen über die Entfernung der interessanten Objekte der Umgebung, noch über die Richtung, in der sich die Objekte befanden. Beim Auswählen der Objekte mittels Anklicken wurden Informationen darüber angezeigt.

Die nächste Ansichtsart war der Kartenmodus (*Abbildung 5: Kartenansicht*). Hier wurde eine Karte der Umgebung auf dem Display dargestellt und die „points of interest“ darübergelegt. Diese waren wieder einzeln auswählbar. Der Pfeil in der Mitte der Karte zeigte die aktuelle Sehrichtung des Benutzers an. Die Karte war genordet, es konnte die Sehrichtung des Benutzers remote geändert werden und auch die Karte in die Bewegungsrichtung verschoben werden.

Als dritte Art der Umgebungsdarstellung wurde eine Radaransicht (*Abbildung 6: Radaransicht*) konzipiert. Eine Windrose zeigte die Himmelsrichtungen an und wiederum wurden die POI darübergelegt. Der Radar und die darauf befindlichen

Objekte konnten remote mittels des Laptops gedreht werden, je nachdem, in welche Richtung der Benutzer blickte. Die Punkte konnten wiederum durch Anklicken ausgewählt und damit nähere Informationen darüber eingeholt werden.

Als letzte Darstellungsart wurde eine erweiterte Realität (*Abbildung 7: Erweiterte Realitätsansicht*) verwirklicht. Das Bild, das mittels der im Handy integrierten Kamera aufgenommen wurde, wurde am Display dargestellt und mit den POI überlagert. Beim Prototyp wurde das mittels eines vorher angefertigten Panoramabildes, dem die POI überlagert wurden, realisiert. Diese Darstellungsart war dadurch auf einen vorher definierten Umgebungspunkt eingeschränkt, was aber für die Testzwecke ausreichend war.

Unabhängig von der Darstellungsart wurden die POI immer mit roter Schrift dargestellt. Beim Anklicken dieser Punkte wurde eine visuelle und optional zusätzlich eine akustische Information ausgegeben. Ob die akustische Information ausgegeben wurde, konnte remote über den Laptop eingestellt werden.

2.4 Testablauf

2.4.1 Einleitung

Der Test wurde im Areal der Donaucity in Wien abgehalten. Es waren die Testperson, der Testleiter und der sogenannte „Wizard of Oz“ - Operator, welcher das Programm remote kontrolliert hat, bei den Tests beteiligt. Um noch zusätzlich mittels Videokamera und Digitalkamera zu dokumentierten, war noch eine weitere Person dabei.

Abbildung 8: Benutzertest mit dem remote über Laptop gesteuertem Prototypen. Personen von links nach rechts: Testleiter, Testperson, „Wizard of Oz“ Operator

Es hat sich bald herauskristallisiert, dass zu heller Sonnenschein ein Problem war. Sowohl am Display des Laptops, von wo aus die Remotesteuerung stattgefunden hat, als auch am Display des VPA IV waren bei direkter Sonneneinstrahlung die dargestellten visuellen Ausgaben oft nur sehr schwer zu erkennen. Das war schon das erste Zeichen, dass eine akustische Ein- und Ausgabe sinnvoll sein könnte.

Im folgenden wird der komplette Testablauf beschrieben, dem auditiven Teil aber besondere Aufmerksamkeit geschenkt. Die sprachliche Kommunikation mit der Software ist als zusätzliche Interaktionsmöglichkeit für die verschiedenen Darstellungsarten, die bereits beschrieben wurden, gedacht. Da die Spracherkennung nicht simulierbar war, wurde dem Tester die Möglichkeit einer sprachlichen Interaktion dargelegt und Fragen dazu gestellt. Etwa ob er diese für sinnvoll hält und ob er sie auch benutzen würde.

2.4.2 Teststationen

- *Besprechungsraum*: Die Testperson wurde im Besprechungsraum empfangen um demographische Fragen zu besprechen. Es wurde außerdem nach der Häufigkeit und Art der Handy- bzw. Internetnutzung gefragt.

- *Standort 1*: Die zweite Station für den Test war der erste Standpunkt, der direkt unter dem Tech-Gate lag (*siehe Abbildung 4: Listenansicht*). (Anm.: Im Tech-Gate befindet sich das ftw.) Der Testperson wurden die vier möglichen Darstellungsarten (Liste, Karte, Radar, erweiterte Realität) gezeigt und er musste bei jeder Darstellungsart eine kleine Aufgabe erfüllen. Eine Aufgabe war heraauszufinden, wann Veranstaltungen in den umliegenden Gebäuden stattfinden. Informationen wurden nach der Gebäudeanwahl durch den Benutzer nur schriftlich auf dem Display dargestellt. Nachdem er mithilfe jeder Darstellungsart Informationen über die Gebäude in der näheren Umgebung gefunden hat, wurden die Darstellungsarten von der Testperson subjektiv nach dem persönlichem Gefallen geordnet. Auch wurde nach Verbesserungsvorschlägen und Gründen für mögliches Missfallen der Darstellungsmöglichkeiten gefragt.

- *Auf dem Weg zum zweiten Standort* wurde der Testperson (ohne Vorwarnung) beim Passieren eines interessanten Punktes, der in diesem Fall die Indoor-Golf Anlage in der Donaucity war, eine räumlich orientierte Audioinformation vorgespielt. Je nachdem, auf welcher Seite sich das Gebäude befand, hörte die Testperson auch die Information mittels der Stereo-Kopfhörer von dieser Seite. Die Entfernung des Gebäudes wurde durch die Lautstärke der Audioausgabe simuliert.

  Es wurde auch die Möglichkeit gezeigt, zuerst einen Hinweiston anstelle der direkten sprachlichen Audioausgabe auszugeben.

- *Standort 2*: Am zweiten Standort sollte die Testperson gezielt zu bestimmten Gebäuden Informationen einholen - auf ähnliche Art und Weise wie beim ersten Standort. Anstatt der Liste, wo die Punkte nur aufgelistet waren, wurde diesmal mit dem Handy auf ein Gebäude gezeigt und durch Drücken einer Taste Informationen drüber eingeholt. Die anderen drei Darstellungsarten von Standort 1 (Karte, Radar, erweiterte Realität) blieben gleich. Dem User wurde lediglich gesagt, er soll, anstatt Veranstaltungen zu suchen, bestimmte Informationen zu einem Gebäude ermitteln.

  An diesem Standort wurde zusätzlich zu den visuellen Informationen auf dem Display dieselbe Information auch auditiv über die Kopfsprecher ausgegeben. Auch hier musste die Testperson wieder eine subjektive Reihenfolge der ausprobierten Selektionsarten erstellen. Auch den Nutzen der zusätzlichen auditiven Ausgabe und die Möglichkeit des räumlichen Klanges hatte er zu bewerten.

- *Standort 3*: Beim letzten Standort wurden dem Benutzer lediglich verschiedene Möglichkeiten präsentiert, durch Gebäude durchzusehen. Man konnte etwa mit Hilfe der Karte und des Radars durch Gebäude durchblicken. Als visuelle Hilfe wurden die nicht sichtbaren Gebäude mit einer anderen Farbe beschriftet. Hier wurde auch ein Google-Earth Blick präsentiert. Google Earth [GoEa07] ist eine kostenlose 3D-Anwendung für Breitbandanschlüsse die detaillierte Satellitenbilder von der Erde zur Verfügung stellt. Bei der erweiterten Realität wurde das Durchblicken von Gebäuden durch eine Art Diashow ermöglicht. Die Gebäude wurden entweder langsam weggeblendet, oder wie ein Dia gleich weggezogen und so das dahinter liegende Gebäude sichtbar gemacht.

  Abschließend wurden noch weitere, testbezogene Fragen gestellt.

2.4.3 Testergebnisse

An dem Test nahmen zwölf Personen Teil. Es wurde darauf geachtet, möglichst die Zielgruppe der Anwendung zu treffen. Es wurden fünf weibliche und sieben männliche Testpersonen im Alter von 17 bis 49 (Durchschnitt: 28) Jahren, die unterschiedliche Fertigkeiten mit Pocket PCs hatten, ausgewählt. Es waren Studenten als auch Personen aus verschiedenen Berufsgruppen dabei. Außerdem war den Testpersonen die Gegend Donaucity rund um das Techgate unbekannt, sodass sie für die Aufgaben des Tests, Informationen in einer unbekannten Gegend zu finden, gut geeignet waren.

Die Ergebnisse des Tests, standortspezifisch aufgelistet:

- *Standort 1:* Die Benutzer erstellten eine subjektive Reihenfolge der vier möglichen Darstellungsarten der Umgebung (Karte, Radar, Liste, erweiterter Realitätsmodus) und begründeten ihre Wahl. Bei den Aufgaben an Standort 1 wurden die Informationen noch nicht auditiv ausgegeben.
  - **Karte (siehe Abbildung 5: Kartenansicht):** Die Karte schnitt bei der subjektiven Beurteilung insgesamt am besten ab. Die Begründung ist darin zu sehen, dass die Handhabung einer Karte bei den meisten Testpersonen schon Gewohnheit war. Diejenigen, die sich mit dem Kartenlesen schwer taten, reihten diese bei den Tests natürlich auch schlechter. Die Kartenansicht wurde als gute Lösung gesehen, Informationen übersichtlich darzustellen.
  - **Liste (siehe Abbildung 4: Listenansicht):** Auch die Liste wurde als informativ angesehen, obwohl keine Informationen über Entfernung und Richtung, in der das Gebäude zu finden war, angegeben waren.
  - **Radar (siehe Abbildung 6: Radaransicht):** Der Radar wurde von den Testpersonen generell nicht angenommen. Dies wurde damit begründet, dass die Darstellung unübersichtlich war und Entfernungen nur schwer abschätzbar waren. Auch wenn sich der Radar immer der Blickrichtung des Benutzers (durch Remotesteuerung) angepasst hatte, fanden sich nur wenige Testpersonen damit zurecht.

- **Erweiterte Realitätsmodus (siehe Abbildung 7: Erweiterte Realitätsansicht):** Der erweiterte Realitätsmodus fand ein breites Spektrum der Akzeptanz. Einerseits wurde er als interessante Möglichkeit gesehen, Informationen über die umliegenden Gebäude zu erhalten. Andererseits wurde auch angemerkt, dass nicht mehr Gebäude sichtbar wurden als ohne Pocket PC schon sichtbar waren.

- *Auf dem Weg zu Standort 2:* Auf dem Weg zu Standort 2 wurden erstmals akustische Informationen abgespielt. Beim Vorbeigehen an der Indoor-Golf Anlage erhielt die Testperson, je nachdem, in welche Richtung sie unterwegs war und wo sich daher das Gebäude befunden hat, akustische Informationen von der entsprechenden Seite zugespielts.
  - *Forschungsfrage 1:* Viele Testpersonen drehten sich bei Abspielen der Information intuitiv in die Richtung, wo sich das Gebäude befand. Die Idee, die Information aus der Richtung des interessanten Punktes kommen zu lassen, wurde als sehr gut angesehen. Die Unterscheidung der Richtung wurde von den meisten Testpersonen erkannt, nur die Entfernung zur Informationsquelle war nicht abzuschätzen. Ziel sollte es hier sein, die Möglichkeiten des räumlichen Klanges mit Stereokopfhörern zu erforschen.
  - *Forschungsfrage 4:* Es wurde dem Benutzer hier ein Hinweissound vorgespielt. Dieser wurde von den Benutzern als positiv aufgenommen. Es wird generell vorgezogen, vorher nur die Information, dass gerade ein interessanter Punkt mit zusätzlichen Informationen passiert wird, mithilfe eines kurzen Signals zu erhalten.
  - *Forschungsfrage 2:* Man sollte nicht gleich mit möglicherweise für den User irrelevanter Information aus allen Richtungen überrollt werden. Anstatt des oder auch zusätzlich zum akustischem Hinweis wurde auch die Vibrationsfunktion des Handys als gute Möglichkeit gesehen, Aufmerksamkeit zu erregen. Der Benutzer will ja das Gerät nicht ständig in der Hand halten.

- *Standort 2:* Bei Standort 2 wurde jetzt zusätzlich zu den Informationen am Display diese auch akustisch über die Kopfhörer zur Verfügung gestellt. Die Testpersonen gaben auch hier eine Wertung der Interaktionsmöglichkeiten mit der Umgebung ab. Bis auf die Tatsache, dass die Liste durch bloßes Zeigen auf ein Gebäude ersetzt wurde, waren die Darstellungsmöglichkeiten dieselben wie beim ersten Standort. Hier schnitten alle Darstellungsarten gleich gut ab. Nur die Möglichkeit, die Umgebung mithilfe einer Windrose (Radar) zu erkunden, wurde überhaupt nicht angenommen.
  - **Einfaches Zeigen:** Die Idee mit dem Handy einfach durch Zeigen auf ein Gebäude Informationen darüber zu erhalten, wurde als sehr einfache und schnelle Art der Informationsgewinnung gut aufgenommen. Allerdings sollte die Aktion benutzergesteuert (etwa durch Bestätigung via Button) ablaufen, sodass nicht laufend Informationen erhalten werden, wenn sich die Testperson mit dem Handy dreht.
  - **Karte:** Auch beim zweiten Standort wurde die Karte wieder aufgrund der Übersichtlichkeit und der intuitiven Handhabung als gutes Hilfsmittel gesehen, die Umgebung zu erkunden.
  - **Radar:** Der Radar wurde hier aufgrund der Unübersichtlichkeit und ungewohnten Handhabung abgelehnt.
  - **Erweiterter Realitätsmodus:** Beim erweiterten Realitätsmodus wurden auch viele Eigenschaften als positiv empfunden. So ist es etwa sehr hilfreich, dass die Zuordnung der Gebäudenamen zum Gebäude, das gerade angezeigt wird, eindeutig ist.
  - **Akustische Information:** Dass die Information nicht nur textuell auf dem Display sondern auch akustisch bereitgestellt wird, wurde vom Großteil der Testpersonen als hilfreich eingestuft. Am Nützlichsten ist die akustische Information wenn mit dem Handy auf ein Gebäude gezeigt wird, da hier nicht dauernd auf das Display gesehen werden muss. Von den Benutzern wurde verlangt, dass die akustische Information auch abschaltbar sein soll. Es wurde auch angemerkt, dass akustisch nicht so viele Informationen dargestellt werden können, wie dies textuell möglich ist. Auch wurde vorgeschlagen, dass die akustischen Informationen nicht dieselben sein sollten wie die visuellen.

Am zweiten Standort wurde der Testperson auch spezifische Fragen zu Sprachein- und Sprachausgabe gestellt. Sie wurden gefragt, wie sie mit dem Handy sprachlich interagieren würden, falls dieses Spracheingabe beherrschen würde. Bei diesem speziellen Fall musste die Testperson sagen, wie sie mit Hilfe sprachlicher Befehle die nächsten Veranstaltungen im Austria-Center herausfinden würde.

- *Forschungsfrage 7:* Der Benutzer sollte erklären, wie er mittels Sprachbefehlen zum Austria-Center herausfinden wollte, wann die nächste Technologiemesse darin stattfindet. Es wurden durchwegs nur kurze, eindeutige Befehle angegeben.
- *Forschungsfrage 8:* Keine Testperson versuchte, in ganzen Sätzen mit dem Pocket PC zu kommunizieren. Etwa: „Wann findet die nächste Technologiemesse im Austria-Center statt?“. Eine solche Art der Kommunikation wird anscheinend einem Computer beziehungsweise Handy noch nicht zugetraut.
- *Forschungsfrage 9:* Die Testpersonen würden sprachlich durch das Programm navigieren, indem sie etwa: „Austria-Center“ -> „Veranstaltungen“ oder: „Austria-Center“ -> „Technologiemesse“ sagten.
- *Forschungsfrage 16:* Die Benutzer würden sich gestört fühlen, wenn jemand neben ihnen das Programm verwendet und diesem sprachliche Befehle gibt, auch wenn keine sprachliche Rückmeldung erfolgt.
- *Forschungsfrage 17:* Wenn zusätzlich auch noch die Sprachausgabe für die Umgebung hörbar wäre, würden sich die Testpersonen noch geringfügig stärker belästigt fühlen.
- *Forschungsfrage 18:* Es machte für die Testpersonen keinen großen Unterschied, ob die Audioausgabe via Headset oder den normalen Lautsprecher des Handys erfolgen würde. Das Headset wurde nicht als Lösung für die Belästigung angesehen, weil alleine die Audioeingabe als sehr störend empfungen wurde.

- *Standort 3:* Die Testperson wurde gebeten, die Möglichkeiten, durch Gebäude „durchzusehen“, wieder in eine subjektive Reihenfolge zu bringen. Hier wurde die Karte als am hilfreichsten angesehen. Die Darstellung der Umgebung mit einem Blick von oben (einem Google-Earth-Blick) wurde von einigen Testpersonen als interessante Möglichkeit empfunden. Die Meinungen schwanken hier von „guter Idee“ bis zu „unsympathisch“. Für viele ist diese Darstellung eine neue und dadurch ungewohnte Alternative. Als „Spielerei“ ist sie nicht schlecht, zur Orientierung aber eher ungeeignet. Auch hier wurde der Radar wieder wegen Unübersichtlichkeit und Überladung als nicht nützlich bezeichnet. Ein Großteil der Testpersonen bezeichnete die Darstellung der nicht sichtbaren Gebäude in einer anderen Farbe als nützlich und übersichtlich. Dahinterliegende Gebäude sichtbar zu machen, indem die davorliegenden Gebäude im erweiterten Realitätsmodus weggeblendet werden, wurde als unbefriedigend befunden. Die allgemeine Aussage war, dass ein zu kleiner Ausschnitt zu sehen ist und sich die Testpersonen wenig bis überhaupt nicht zu Recht fanden.

2.5 Zusammenfassung

Der Benutzertest, der im zweiten Kapitel der Diplomarbeit beschrieben wurde, stellte einen wichtigen Punkt im Zuge der Entwicklung einer intuitiven und benutzerfreundlichen Anwendung dar. Mithilfe dieses Tests wurde herausgefunden, was sich der zukünftige Benutzer von solch einer Anwendung erwartet, wie er damit umgehen und ob er sie überhaupt verwenden würde. Die wichtigsten Erkenntnisse hierbei waren, dass die Programmidee, Informationen aus der näheren Umgebung holen zu können, als interessant angesehen wurde. Einige Forschungsfragen wurden bereits beantwortet, etwa dass sich die Testpersonen gestört fühlen, wenn ein User in ihrer näheren Umgebung ein System mit Sprachinterface benutzt. Diese hier gesammelten Ideen gilt es in das Design der Benutzerschnittstelle und in der darauf folgenden Implementierung zu übernehmen und zu berücksichtigen.

3 Design der Benutzerschnittstelle

3.1 Überblick

Aufgrund der Testergebnisse des im Kapitel 2 beschriebenen Benutzertest wurde die Benutzerschnittstelle entworfen. Das Design der Benutzerschnittstelle bezog sich auf das Zeigen und Sprechen mit dem Handy bzw. Pocket PC. Auf die Darstellung der visuellen Elemente wie beim obigen Benutzertest wurde verzichtet. Es ging hier primär um die Gestaltung der Spracheingabe und Sprachausgabe. Wenn mit dem Handy auf ein Gebäude gezeigt wird und sprachlich Informationen eingeholt werden, ist es nicht wichtig, ob auf dem Display eine Karte, eine erweiterte Realität oder eine andere Darstellungsform zu sehen ist, da ohnehin nicht darauf geschaut wird.

Bei der Entwicklung des Programms standen zwei verschiedene Zielsysteme zur Verfügung. Beide Zielsysteme waren Pocket PCs. Zum einen ein hp iPAQ h5550 und zum anderen der VPA IV von Vodafone, auf dem schon die Benutzertests durchgeführt wurden.

Deshalb bezog sich der Entwurf der Benutzerschnittstelle auch vorerst auf Pocket PCs. Der wesentliche Unterschied von Pocket PCs zu normalen Handys ist der Touchscreen. Mittels eines Pens können vom User Eingaben gemacht werden. Bei den meisten Handys ist das noch nicht möglich. Hier steht auch nur ein kleineres Display zur Darstellung der Informationen und eine Zifferntastatur zur Eingabe bereit. Natürlich ist bei beiden Gerätetypen eine Spracheingabe und eine Sprachausgabe verfügbar. Der Hauptaspekt beim Entwurf der Benutzerschnittstelle lag darin, ein Dialogsystem zu gestalten, um dem Benutzer eine möglichst einfache Interaktion mit dem System zu ermöglichen.

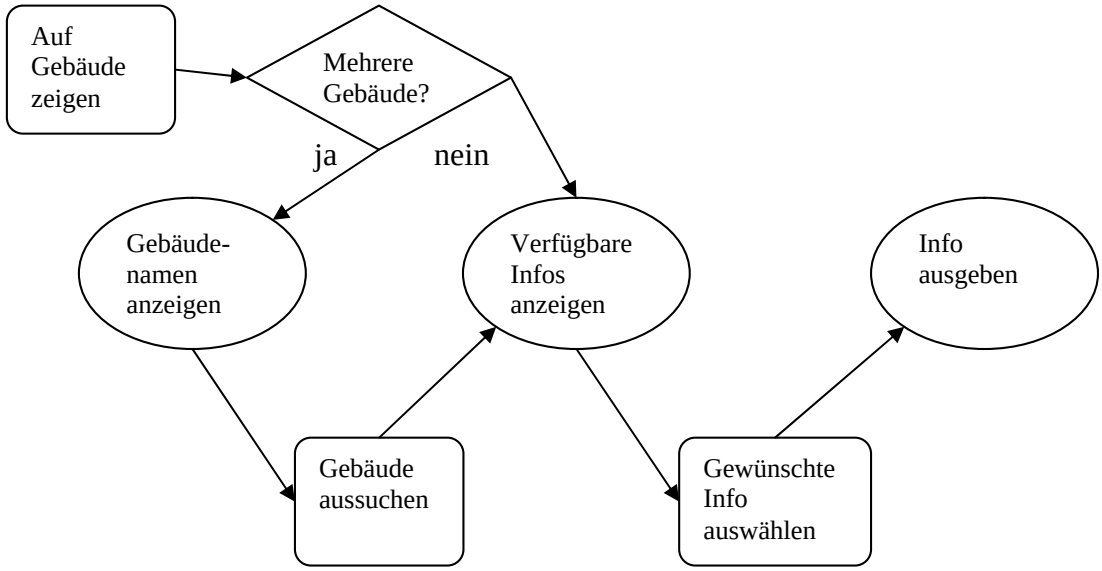
Im Folgenden werden nun einige mögliche Szenarien beschrieben, wie das System in Zukunft eingesetzt werden könnte.

3.2 Szenario 1: "Städtetourismus"

3.2.1 Szenarienbeschreibung

Das erste Szenario beschreibt die Möglichkeit, spezifische Informationen über Gebäude einzuholen. Ein User, zum Beispiel ein Tourist, kommt in eine ihm unbekannte Stadt, Wien, und möchte Informationen zu einem für ihn interessanten Gebäude einholen. Er ist nicht mit einer Reisegruppe unterwegs und es steht ihm daher auch kein Reiseleiter zur Verfügung. Einen Fremdenführer will er auch nicht engagieren - das würde ihm zu teuer kommen. Außerdem steht er jetzt schon vor einer Sehenswürdigkeit. Also verlässt er sich auf das neue, hier vorgestellte System und holt folgendermaßen Erkundigungen ein. Er zeigt mit dem Handy, auf dem das Programm bereits läuft, auf den Stephansdom und fragt: „Gibt es Informationen zu dem Gebäude?". Das Proramm sucht alle möglichen Gebäude, die sich im Sichtbereich des Benutzers befinden und listet diese auf. Der Benutzer sucht das für ihn interessante Gebäude aus. Um die Auswahl zu erleichtern, wird hier eine Kurzbeschreibung angegeben. Zum Beispiel „Haas-Haus": „Das Haas-Haus ist ein gläsernes Geschäftsgebäude.". Falls nur ein interessantes, dem System bekanntes, Gebäude in Sehrichtung des Benutzers verfügbar ist, entfällt dieser Auswahlschritt. Danach werden die möglichen Informationen, die gespeichert sind, aufgelistet. Diese Informationen werden textuell auf dem Bildschirm des Pocket PCs und auch akustisch ausgegeben. Die akustische Ausgabe könnte folgendermaßen lauten: „Es sind historische Informationen und die Termine verschiedener Veranstaltungen zum Stephansdom vefügbar". Nach der Auswahl der gewünschten Alternative wird dem User diese ausgegeben. Die Eingabe wird sprachlich oder durch direkte Eingabe über den Touchscreen möglich sein. Allerdings sind hier die Sprachausgabe und die Spracheingabe ein wichtiger Aspekt. Schließlich wird mit dem Handy gerade auf ein Gebäude gezeigt. Da ist es schwer, auch gleichzeitig auf das Display zu sehen. Es folgt ein kurzes Ablaufprogramm, um die einzelnen Interaktionsschritte besser vorstellbar zu machen.

3.2.2 Ablaufdiagramm



3.3 Szenario 2: „An der Haltestelle“

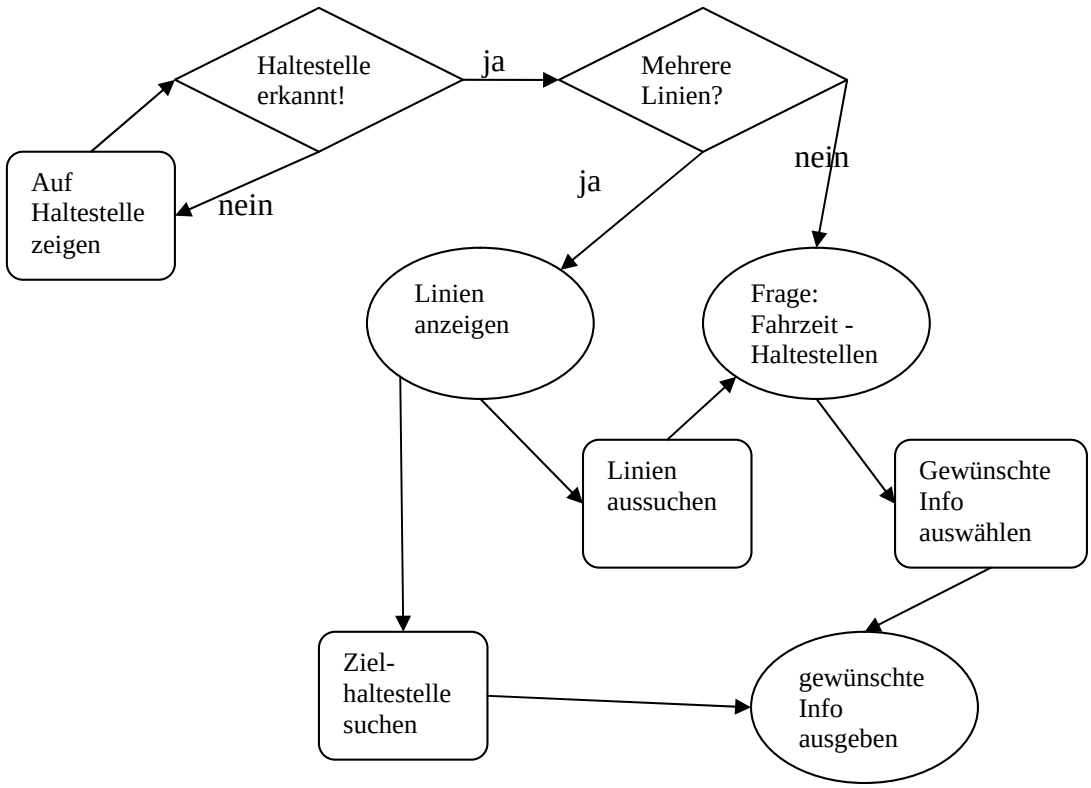
3.3.1 Szenarienbeschreibung

Beim zweiten Szenario kommt der Benutzer zu einer Haltestelle und versucht den Fahrplan zu lesen. Da es schon spät und die Nacht bereits hereingebrochen ist, kann der User den Plan nicht mehr entziffern. Er denkt daran, dass er das neue Programm auf seinem Handy installiert hat, das ihm Informationen darüber liefern kann, startet es, zeigt mit dem Handy auf die Haltestelle und sagt: „Infos?“. Wenn das Programm eine Haltestelle erkannt hat, liefert es als Rückgabe: „Straßenbahnhaltestelle Rathaus. Straßenbahnlinien 2 in Richtung Ring-Kai-Ring und D in Richtung Südbahnhof.“ Der Benutzer fragt nach dem Fahrplan für die Straßenbahnlinie 2. Im Folgenden kann sich der Benutzer aussuchen, welche Informationen er genau haben möchte: Abfahrtszeiten, weitere Haltestellen der Straßenbahnlinie mit oder ohne Fahrtdauer. Ein Routenplaner oder ähnliches ist hier nicht verfügbar. Eine Überlegung bei diesem Szenario ist, nicht nur die aktuellen Abfahrtszeiten (etwa die nächsten fünf Züge) ausgeben zu lassen, sondern auch Abfahrszeiten, die in der Zukunft liegen. Das Problem hierbei könnte sein, dass die Sprachsoftware Zeitangaben nicht genau interpretieren kann, was nach ersten Versuchen zu befürchten ist. Möglicherweise müssen hier die Eingaben per Tastatur

bestätigt werden. Eine mögliche Rückgabe des Programmes ist hier: „Nächste planmäßige Abfahrt um 1315.“

Hier wäre auch eine Eingabe wie „Fährt diese Linie zum Karlsplatz?“ möglich. Dann werden die Stationen nach der gewünschten durchsucht. Es wird sich auch realisieren lassen, alle Linien der ausgewählten Haltestelle nach dem gewünschten Fahrziel zu durchsuchen. Wenn dann das gewünschte Fahrziel durch das Programm gefunden wird, könnte die Ausgabe etwa so aussehen: „Linie 2 fährt zu Karlsplatz. Nächster Zug um 1315!“.

3.3.2 Ablaufdiagramm



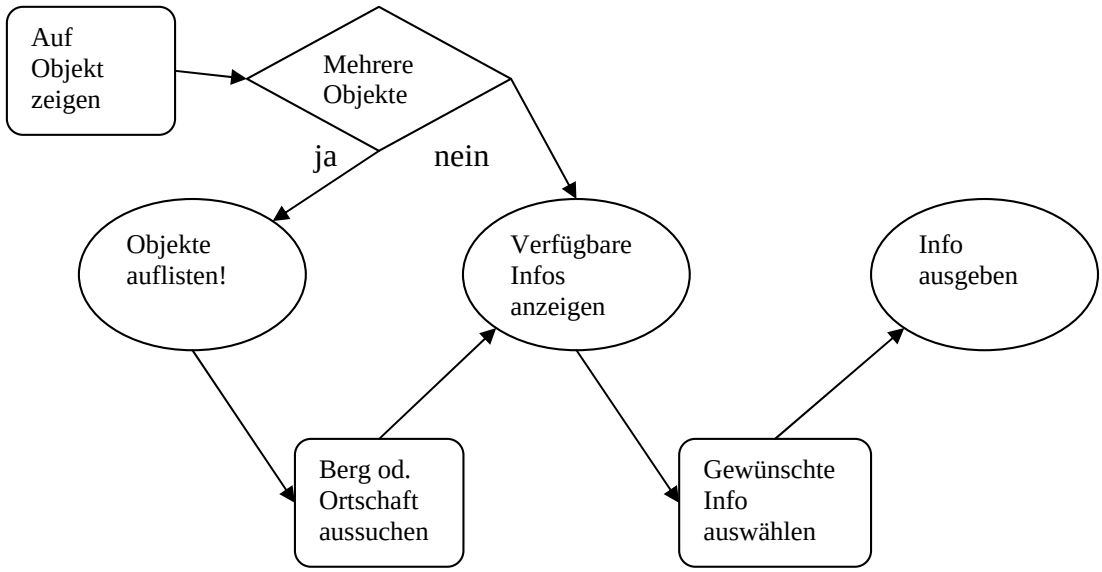
3.4 Szenario 3: „Beim Wandern“

3.4.1 Szenarienbeschreibung

Es gibt auch noch weitere denkbare Szenarien, um die neue Anwendung einzusetzen. Vorstellbar wäre hier zum Beispiel ein Wanderausflug, wo Informationen über einen Berg, den der Benutzer vor sich sieht gewünscht werden. Oder den Namen der Ortschaft, die in der Ferne zu sehen ist. Vielleicht wäre es hier interessant herauszufinden, ob es dort eine Gaststätte gibt, um sich nach dem mehrstündigen Trip wieder ordentlich zu stärken, falls der Ausflug im vorhinein nicht aufs Genaueste geplant wurde. Hier spielt die Entfernung eine ganz andere Rolle als etwa in der Stadt, wo der Benutzer meist direkt vor einem Gebäude oder einer Haltestelle steht. Das Modell der Umgebung, das hierfür berechnet werden muss, ist allerdings nicht Teil dieser Diplomarbeit. Es wird hier davon ausgegangen, dass das Modell schon funktioniert und sowohl in der Stadt als auch im freien Gelände die erwarteten Ergebnisse liefert. Es wird sowohl eine Haltestelle, die sich unmittelbar vor dem Benutzer befindet, richtig identifiziert, als auch ein Berg oder eine Ortschaft in einigen Kilometern Entfernung erkannt. Informationen über einen Berg könnten die Höhe des Berges, mögliche Wander- oder Kletterrouten sein. Oder vielleicht auch, ob sich auf dem Berg ein Skigebiet befindet. Welche Informationen verfügbar sind, wird wieder visuell oder sprachlich ausgegeben und der Wanderer kann die genauen Informationen anhören oder auch vom Display ablesen.

Als Erweiterung dieses Szenarios ist einiges denkbar. Befindet sich der Benuzer etwa in einem Skigebiet, könnten die zu befahren möglichen Pisten genauer vorgestellt werden wie zum Beispiel den Schwierigkeitsgrad, ob ein gutes Restaurant am Ende der Piste wartet oder welche Lifte zu Verfügung stehen, wenn die Piste bewältigt wird.

3.4.2 Ablaufdiagramm

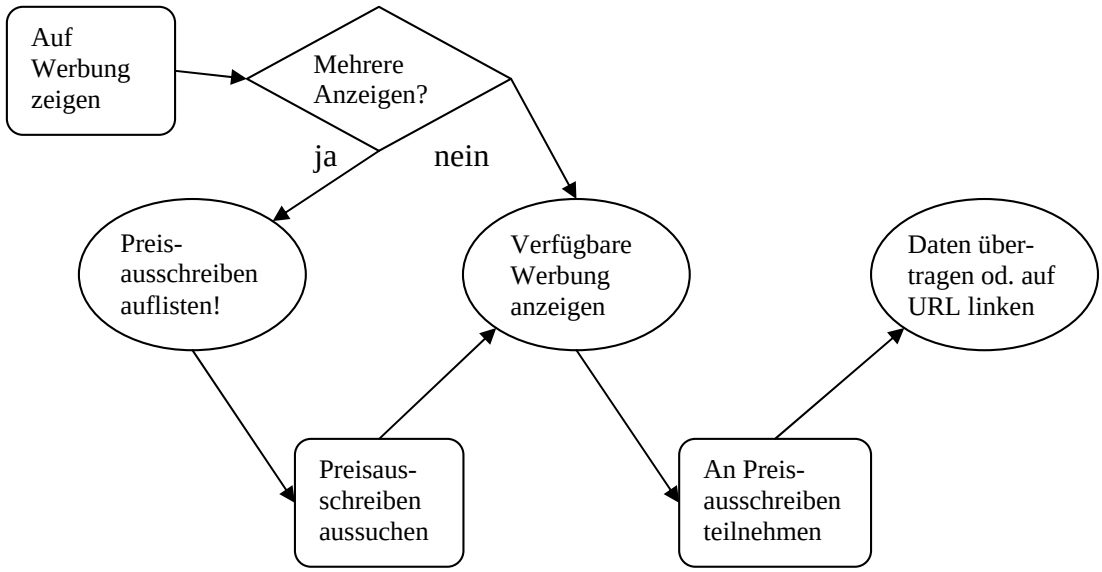


3.5 Szenario 4: „Teilnahme an Preisausschreiben“

3.5.1 Szenarienbeschreibung

Mit der neuen Applikation kann auch an Preisausschreiben teilgenommen werden. Hinter manchen Objekten wie zum Beispiel Werbetafeln oder auch bei Restaurants können sich entweder sichtbar darauf oder auch nur für „p2d“-User gedachte Preisausschreiben verbergen. Beim Einholen der Informationen wird der Benutzer dann darauf aufmerksam gemacht, dass die Möglichkeit besteht, an einem Preisausschreiben teilzunehmen. Dies kann geschehen, indem der Benutzer auf die entsprechende Internetseite verlinkt wird, oder indem von der Applikation die benötigten Daten übertragen werden. Es könnte der Teilnehmer mit der Handynummer eindeutig identifiziert werden und danach bei einem etwaigen Gewinn per SMS informiert werden.

3.5.2 Ablaufdiagramm



3.6 Abgeleitete Programmkonzepte

Im Prinzip ähneln sich die vier Szenarien. Es werden dem Benutzer nur jeweils andere Informationen gegeben. Es ist nicht unbedingt ein Nachteil, wenn die Szenarien gleich aufgebaut sind. So kann eine einfache Benutzerschnittstelle bereitgestellt werden, bei der sich der Benutzer möglichst schnell zurechtfindet. Der Unterschied liegt zum einen bei den verschiedenen zu gebrauchenden Wörtern für das Sprachdialogsystem. Die hierfür verwendete Grammatik muss allen Szenarien genügen und die möglichen Eingaben abdecken. Zum anderen gibt es einen Unterschied zwischen einem möglichen Preisausschreiben und anderen interaktiven Varianten wie Zimmerreservierungen, wo der Benutzer Daten nicht nur in Form von Informationen empfängt, sondern auch Daten von sich selbst weitergibt. Auch beim Szenario an einer Haltestelle gibt es Besonderheiten. Dort wird dem Benutzer ein zusätzlicher Suchmodus nach einer bestimmten Haltestelle zur Verfügung gestellt.

3.6.1 Entwurf der Benutzerschnittstelle

Um die von den oben genannten Szenarien gestellten Anforderungen zu erfüllen, gab es zwei verschiedene Ansätze, um das Programm zu konzipieren. Es wurden zwei verschiedene Programme mit verschiedenen Interaktionsmöglichkeiten entworfen. Diese werden hier nun kurz beschrieben.

- **Programmentwurf „Wordspotting“:** Das erste Programm wurde als multimodales Interface, wo Eingaben sowohl via Pen als auch sprachlich getätigt werden können, konzipiert. Bei entsprechendem Programmstatus, wo der Benutzer eine Eingabe tätigen kann, wartet das Programm auf einen Befehl, der entweder mit dem Pen oder sprachlich gegeben werden kann. Um einen sprachlichen Befehl geben zu können, muss bei diesem Programm das Schlüsselwort „Eingabe“ gesprochen werden. Wenn dieses verstanden wurde, kann der Benutzer den eigentlichen Befehl sagen. Das Programm wechselt dann in den für den Befehl vorgesehenen Programmzweig.

  Ein potentieller Nachteil von diesem Design besteht darin, dass das Programm längere Zeit auf sprachliche Usereingaben wartet, ohne sicher zu sein, dass diese überhaupt getätigt werden. Dadurch entstehen für die Spracherkennungssoftware Probleme, weil sie nicht genau weiß, wann ein Input beginnt. Durch etwaige Hintergrundgeräusche oder auch durch die eigene Sprachausgabe des Systems kann es hier zu unerwünschten Erkennungen kommen oder auch zu vermehrten Verwerfen von Spracheingaben. Das Programm kann den Unterschied nicht erkennen, ob der User mit jemand anderem spricht oder dem Pocket PC einen Befehl erteilt. Es wird auf alle Fälle versuchen, das Schlüsselwort „Eingabe“ herauszuhören. Egal, ob es nun an das Programm gerichtet ist oder in einem anderen Zusammenhang ausgesprochen wird.
- **Programmentwurf „Push 2 Speak“:** Der zweite Entwurf ist das rein sprachlich gesteuerte Programm „Push 2 Speak“. Hier muss, um die Spracherkennung zu starten, eine Taste auf dem Zielsystem gedrückt werden. Durch einen kurzen Signalton wird dann darauf hingewiesen, dass nun dem Programm einen Befehl gegeben werden kann. Das Ende des Sprachbefehls wird wiederum durch einen Tastendruck bestätigt. Der Vorteil von diesem

Entwurf gegenüber dem multimodalen Interface besteht darin, dass der für die Sprachsoftware zu interpretierende Teil eindeutig festgelegt ist. So kann der Spracherkennungssoftware ein präziseres Signal überliefert und dadurch die Fehlerrate vermindert werden. Auf die Multimodalität wird hier verzichtet.

Bei beiden Systemen erfolgt die Ausgabe sowohl auf dem Bildschirm als auch sprachlich via Headset oder auch über die normalen Lautsprecher. Es ist möglich, auf dem Bildschirm auch andere Informationen auszugeben als die auditiven. Interessant ist hier herauszufinden, ob verschiedene sprachliche und textuelle Informationen auch aufgenommen werden können (siehe Forschungsfrage 3).
Natürlich sind auch Kombinationen der beiden Systeme vorstellbar. Etwa, vor jeder sprachlichen Eingabe eine Taste zu betätigen. Allerdings kann vielleicht darauf verzichtet werden, wenn die Sprachsoftware auch aus einem längeren Inputstream die richtigen Informationen extrahieren kann. Welches der beiden Systeme von den zukünftigen Usern besser angenommen wird, wird die Benutzerevaluation, die am Ende der Entwicklung der beiden Systeme steht, zeigen.

3.6.2 Fehlererkennung

Derzeit ist es noch so, dass die sprachliche Eingabe des Benutzers nicht immer richtig erkannt werden kann. Die Erkennung der komplexen menschlichen Sprache mit ihren individuellen Eigenheiten kann von der Software noch nicht komplett garantiert werden. Zu ähnlich sind sich viele Laute und auch viele Wörter. Hintergrundgeräusche können die Erkennung genauso beeinflussen wie etwaige Sprach- oder Sprechfehler des Benutzers. Das Programm weiß natürlich nicht, ob die Eingabe korrekt erkannt wurde. Die Software gibt zusätzlich zu der erkannten Spracheingabe einen Wert (Confidence-Value) aus, der die Wahrscheinlichkeit einer richtigen Erkennung darstellt. Es ist genau zu überlegen, ab welchem Wahrscheinlichkeitswert man die Erkennung als korrekt interpretiert.
Hier gibt es zwei verschiedene Arten von Fehlertypen[NiKo05]:
Falsche Akzeptanz (false acceptance ratio): Das System akzeptiert eine Erkennung aufgrund des erreichten Confidence-Values als korrekt, obwohl diese falsch ist.
Falsche Zurückweisung (false rejection ratio): Das System weist eine Erkennung aufgrund des zu geringen Confidence-Values zurück, obwohl diese korrekt ist.

Es ist nicht möglich, die Wahrscheinlichkeit beider Fehlerarten komplett zu minimieren. Bei zu hohem Schwellwert für den Confidence-Value werden viele richtig erkannte Eingaben aufgrund des zu geringen Wertes als falsch abgelehnt. Dafür kann davon ausgegngen werden, dass eine als richtig erkannte Eingabe aufgrund des hohen Vertrauenswertes ziemlich sicher dem Benutzerinput entspricht. Ein zu hoher Schwellwert wird die Usability des Programms ziemlich beeinträchtigen, weil die Eingaben öfters wiederholt werden müssen, bis sie als richtig erkannt werden.
Andererseits muss darauf geachtet werden, den Schwellwert nicht zu niedrig zu wählen. Vor allem bei ähnlichen Wörtern kann das falsche Wort erkannt werden. Dieses wird dann zwar einen niedrigeren Vertrauenswert, aber dann trotzdem akzeptiert werden. Auch hier leidet die Benutzerfreundlichkeit des Programmes, wenn dieses häufig gesprochene Eingaben falsch interpretiert und so anders reagiert, als durch die Eingabe erwartet wird.

3.6.3 Bargein

Interessant ist hier die Möglichkeit des Bargein. Bargein bedeutet die Unterbrechung der Sprachsynthese durch den Benutzer mit Hilfe einer Spracheingabe. Hier gilt es zu überprüfen, ob es möglich ist, gleichzeitig Sprachsynthese und Spracherkennung auf dem Testgerät laufen zu lassen. Leider ist zu befürchten, dass die Zielsysteme (vor allem der iPAQ h5550) zu wenig Rechenkapazität aufweisen, um eine Information ausgeben und gleichzeitig auf mögliche Benutzereingaben warten zu können.

3.7 Zusammenfassung

Im dritten Kapitel der Diplomarbeit wurden verschiedene Szenarien vorgestellt, wo das Programm in Zukunft Verwendung finden könnte. Es wurden Ideen für den Entwurf des Sprachinterfaces vorgestellt, die mit Hilfe zwei verschiedener Ansätze realisiert werden. Zum einen der „Wordspotting“ Ansatz, der vorsieht, dass das Programm zuerst ein Schlüsselwort erkennen soll, um dann einen ebenfalls gesprochenen Befehl zu erkennen und auszuführen. Dieses Programm solle zusätzlich ein multimodales Interface zur Verfügung stellen, also auch eine normale Pen-Interaktion erlauben.
Die zweite Anwendung, „Push 2 Speak“ erwartet vom Benutzer einen Tastendruck, ehe es mit der Aufnahme des gesprochenen Befehls beginnt. Am Ende des Befehls wird wieder ein Tastendruck erwartet, um den genauen Sprachbefehl exakt eingrenzen zu können. Welche der beiden Variationen den Benutzern besser gefältl, wird im abschließenden Benutzertest herausgefunden. Die Realisierung der beiden verschiedenen Anwendungen wird im folgenden Kapitel beschrieben.

4 Implementierung der Prototypen

4.1 Überblick

Im vierten Kapitel der Diplomarbeit wird die Entwicklung der Prototypen beschrieben. Hierfür wurde mit zwei Firmen zusammengearbeitet, Nuance und Loquendo. Diese beiden Firmen stellten Evaluierungslizenzen für ihre Software zur Verfügung. Zu Beginn des Kapitels werden die beiden Firmen vorgestellt und die zur Verfügung gestellte Software beschrieben. Ein wichtiger Teil bei der automatischen Spracherkennung ist die dahinter stehende Grammatik. Diese wird noch vor der eigentlichen Erklärung der Programmstrukturen der beiden im vorigen Kapitel entworfenen Anwendungen mit den verschiedenen Spracheingabemöglichkeiten vorgestellt. Auch im Zuge der Implementierungsphase konnten einige Forschungsfragen, die im zweiten Kapitel aufgelistet wurden, beantwortet werden. Diese Fragen waren eher technischer Natur, etwa wie interessant eine synthetische Sprache bei einem mobilen Gerät gestaltet werden kann.

4.2 Einleitung

Nachdem das Interface entworfen war, konnte mit der Implementierung begonnen werden. Zum einen wurden von Loquendo [Loq07] aus Italien und zum anderen von Nuance [Nu07] (ehemalig Scansoft), deren weltweites Hauptquartier sich in den Vereinigten Staaten befindet, Software im Bereich von ASR und TTS zur Verfügung gestellt.

Die beiden Programme, die beim Design der Benutzerschnittstelle entworfen wurden, wurden zuerst mit Hilfe der Software von Loquendo und danach auch mit der Software von Nuance verwirklicht. So konnten beide Softwareprodukte getestet werden und es wurde herausgefunden, inwiefern die entsprechenden Interaktionsentwürfe mit der jeweiligen Software realisiert werden konnten. Ein systematischer Vergleich der beiden Produkte war nicht Zielsetzung dieser Diplomarbeit. Deshalb wurde die abschließende Benutzerevaluierung nur anhand der fertigen Programme, die auf der Nuance-Software aufbauten, durchgeführt. Sämtliche Anwendungen wurden im Microsoft eMbedded Visual C++ 4.0 entwickelt. Das Testgerät, der HP iPAQ h5550, war mittels ActiveSync mit dem Desktop PC verbunden. In der Folge wird nun zuerst die Software von

Loquendo vorgestellt. Danach wird die Software von Nuance beschrieben und es wersden die Bewegungssensoren erklärt. Einen wesentlichen Teil der Anwendung ist die dahinter liegende Grammatik, auf die noch genauer eingegangen wird.

Zuerst nun zu Loquendo und einer Beschreibung der zur Verfügung gestellten Software.

4.2.1 Loquendo

Die italienische Firma Loquendo stellte sowohl eine ASR (Automatic Speech Recognition) als auch eine TTS-Komponente (Text-to-Speech) in Form einer Evaluierungslizenz zur Verfügung. Die beiden Komponenten waren als zwei verschiedene Softwareprogramme vorhanden.

Die von Loquendo entwickelte TTS-Komponente funktionierte nicht nur für Pocket PCs, sondern auch auf normalen Servern oder Desktop PCs [LTTSB]. Es konnten 16 verschiedene Sprachen mit Hilfe von 36 verschiedenen weiblichen und männlichen Stimmen synthetisch erstellt werden. Ein wichtiger Bestandteil für „Point and Speak“ war, dass verschiedene Ausgabequalitäten zur Verfügung gestellt wurden. Je besser die Qualität, desto größer war der Speicherbedarf und die geforderte Rechenleistung des Pocket PCs. Deswegen wurde auf das kleinste, qualitativ aber auch schlechteste, Ausgabeformat zurückgegriffen.

Auch die ASR-Komponente von Loquendo funktionierte für verschiedenste technische Geräte wie etwa Server, Desktop PCs und auch PDAs. Die automatische Texterkennung verfügte über eine Sprecherunabhängigkeit, ein offenes Vokabular, das ausgebaut werden konnte, sowie eine Geräuschunterdrückung. Hintergrundgeräusche, wie sie etwa bei der Benutzung von ASR im Freien unvermeidbar sind, wurden herausgefiltert und die reine sprachliche Komponente der Spracherkennung übergeben. Die Basistechnologie verwendete zur Spracherkennung neurale Netzwerke und versteckte Markov-Modelle. Die erkannten Wörter oder Sätze wurden mit Wahrscheinlichkeiten gewichtet. Wenn die Wahrscheinlichkeit unter einem bestimmten Wert lag, wurde die Erkennung zurückgewiesen (rejected). Die ASR-Komponente unterstützte 16 verschiedene Sprachen. Die Software erkannte sowohl einzelne Wörter, als auch fließende Sprache.

Die ASR- und auch TTS-Komponente von Loquendo wurden schon in einigen industriellen Anwendungen verwendet. Die sprachgesteuerte Navigationssoftware von

Giovanni Soldinis Trimaran [Loq07] greift auf die von Loquendo zu Verfügung gestellte Technologie zurück. Giovanni Soldinis konnte mithilfe des Systems das ganze Schiff sprachlich steuern und sich dazu frei auf dem Schiff bewegen. Auch ins „Bertone Birusa SuperCar“ wurde diese Technologie eingebaut. Dieses Auto war eine konzeptionelle Entwicklung, die die neuesten Technologien, wie auch die Sprachkontrolle von Loquendo, beinhaltete [Ber07].

Abbildung 9: Das „Bertone Birusa Supercar“, eine konzeptionelle Entwicklung, die die Sprachsoftware von Loquendo eingebaut hat.

4.2.2 Nuance

Bei Nuance waren Sprachein- und -ausgabe auch als zwei verschiedene Produkte verfügbar. Die getestete TTS-Komponente für Pocket PCs und Handys trug den Namen RealSpeak Solo 4.0 [NuRS4]. Die Software konnte textuelle Elemente, ob dies nun Email, SMS oder andere Textformen, in syntethische Sprache umwandeln. 22 verschiedene Sprachen waren hier verfügbar und es konnte aus über 30 verschiedenen Sprecher gewählt werden. Auch bei RealSpeak Solo 4.0 gab es verschiedene Ausgabequalitäten. Abhängig davon, ob eine Anwendung für Desktop PC, Server oder Pocket PC geschrieben wird, waren speichermäßig große und kleine Pakete vorhanden. Natürlich bestimmte die Größe der Komponenten auch die Qualität der Ausgabe. Auch hier wurde für die Entwicklung der „Point and Speak“ Anwendungen die kleinstmögliche Komponente genommen, wieder aufgrund der vorhandenen Speichermöglichkeit und Rechnerleistung des verwendeten Testgerätes.

VoCon 3200 [NuAS] ist jenes Produkt von Nuance, das die automatische Spracherkennung zur Verfügung stellt. Gedacht ist sie für sichere, nämlich sprachliche, Navigation durch Programme wie etwa beim Auto fahren. Als andere Verwendungsmöglichkeit konnten neue Anwendungen für mobile Geräte entwickelt werden. Anhand einer solchen Anwendung („p2d“) wurde VoCon 3200 ausprobiert. Die von Nuance zur Verfügung gestellte Software ist so wie die von Loquendo getestete Komponente sprecherunabhängig. Um VoCon 3200 ideal an die gewünschte Applikation anpassen zu können, waren viele Komponenten als Module verfügbar. Um etwa das Targetsystem zu entlasten, konnte die Grammatik schon im vorhinein kompiliert werden. Die Spracherkennung konnte an einen Sprecher angepasst werden, allerdings wurde auch auf dieses Feature verzichtet, einerseits um Ressourcen zu sparen und andererseits sollte die fertige Testapplikation möglichst viele unterschiedliche Sprecher erkennen.

Die ASR-Komponente von Nuance stellte für eine Spracheingabe nicht eine mögliche Erkennung zur Verfügung, sondern mehrere davon. Jede dieser möglichen korrekten Erkennungen wurde gewichtet und somit die Wahrscheinlichkeit jeder einzelnen angegeben. Bei den Testapplikationen von „Point and Speak“ wurden jeweils die höchstwahrscheinlichen Erkennungen berücksichtigt, die anderen, auch wenn sie oberhalb der Toleranzgrenze lagen, wurden verworfen.

Von Nuance wurde eine Reihe von hilfreichen Entwicklungstools zur Verfügung gestellt. So konnte etwa Grammatiken einfacher entwickelt werden, weil sie schon im Vorhinein auf dem Desktop PC austestbar war. Es musste dafür noch nicht eigens eine Anwendung entwickelt werden.

4.2.3 Bewegungssensoren

Für die Testapplikationen von „Point and Speak“ wurde ein vom ftw. entwickeltes Sensormodul verwendet. Der für diese Diplomarbeit verwendete Sensor war noch zu groß, um ihn in einen Pocket PC zu integrieren. Er diente hauptsächlich als Testgerät für die Entwicklung orientierungsbezogener Dienste.

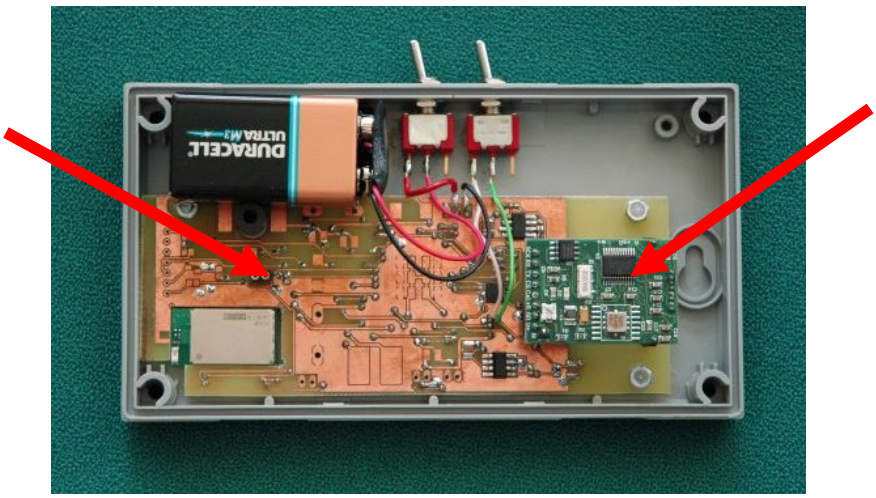


Abbildung 10: Das für die Entwicklung und die Tests verwendete Sensormodul. Der linke Pfeil zeigt auf das Bluetoothmodul. Der rechte Pfeil zeigt auf den eigentlichen Sensor. Beide konnten einzeln ein- und ausgeschaltet werden.

Das Sensorboard war in eine Plastikhülle mit Verschluss eingebaut. *Abbildung 10: Das für die Entwicklung und die Tests verwendete Sensormodul.* zeigt das Sensormodul ohne Verschluss. Das Sensorboard bestand aus einer Bluetooth- und einer Sensorkomponente. Der Sensor bestand aus einem elektronischem Kompass, einem Neigungs- und einem Kippsensor. Die wichtigste Information war die Richtung, in der mit dem Sensor gezeigt wurde. Mithilfe der Neigungs- und Kippwinkel konnte die Kompassrichtung korrigiert werdens. Es folgt eine schematische Darstellung des Sensormoduls.

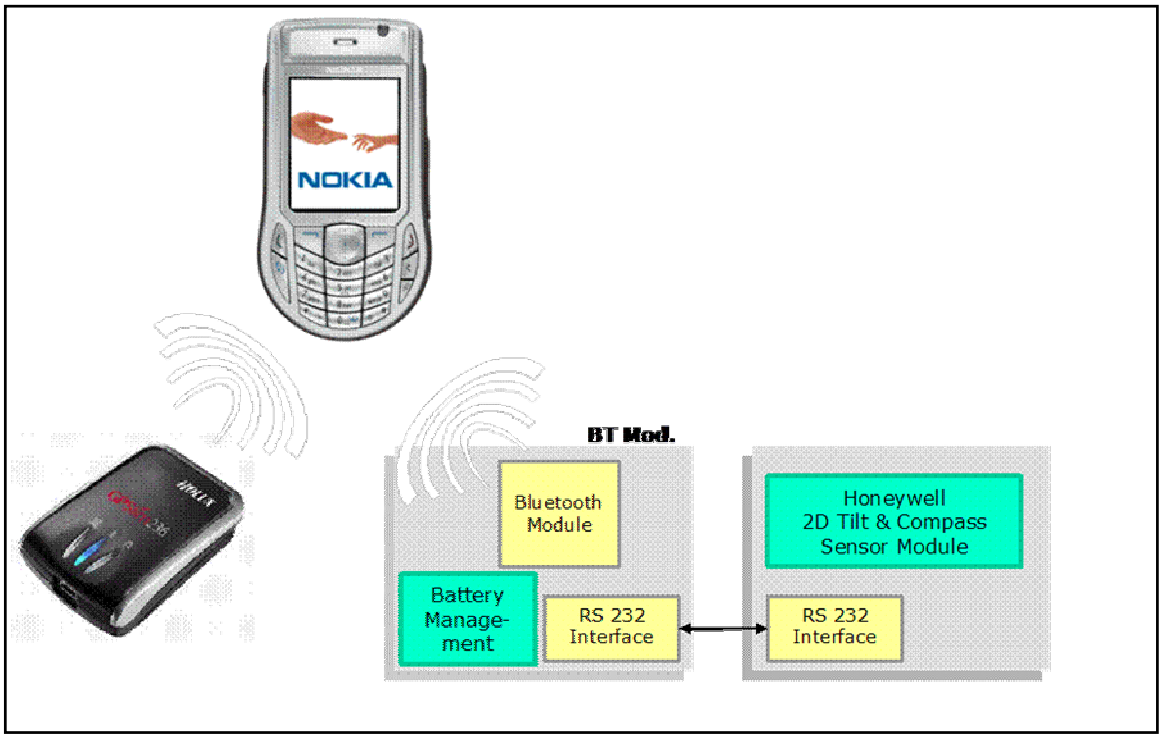


Abbildung 11: Schematische Darstellung des Sensorboards

Der Sensor wird mittels Bluetooth mit dem Pocket PC verbunden und verfügt über eine eigene Stromversorgung.

Bei den Testapplikationen wurde eine fixe Position angenommen, und im vorhinein die Winkel zu den Gebäuden ausgemessen. Da der Hauptaugenmerk der beiden Testapplikationen im Bereich der Spracheingabe und Sprachausgabe lag, konnte die im Testbereich geringe Fehlerrate des Sensormoduls vernachlässigt werden.

4.3 Programmentwicklung

Ein wesentlicher Grundstein für die automatische Spracherkennung stellt die Grammatik dar. Deshalb wird auf die für die fertigen Applikationen benötigte Grammatik zuerst eingegangen. Danach werden die beiden Programme „Wordspotting“ und „Push 2 Speak“ beschrieben. Da sich die Programmstruktur von Loquendo kaum von der Programmstruktur von Nuance unterscheidet, wird sie nur einmal vorgestellt. Unterschiede zwischen beiden Implementierungen werden bei der jeweiligen Programmbeschreibung erwähnt.

4.3.1 Grammatik

Die Grammatik stellt eine wichtige Komponente bei der automatischen Spracheingabe (ASR) dar. Die Grammatik erklärt der Spracherkennung, welche Vokabeln und welche Sätze es verstehen soll. Hier wird festgelegt, welche Sprache zu Grunde liegt, und welche Wortkombinationen möglich sein sollen. Es können in einer Grammatik einfache Befehle aufgelistet werden oder auch ganze Sätze angegeben werden. Aus welchen Wörtern sich diese Sätze zusammensetzen und welche Satzstellungen möglich sein sollen, bestimmt die Grammatik. Bei der Grammatik für die beiden Testapplikationen handelt es sich um eine relativ einfache Struktur. Die Programme sollten einfache Befehle verstehen und nicht etwa ganze Sätze.

Im Anhang (Annex 1) befindet sich die Grammatik, die für die abschließenden Benutzertests verwendet wurde. Hierzu ist folgendes anzumerken:

- Der Syntax dieser Grammatik entspricht dem von W3C entwickelten Standard für Grammatiken [W3C].
- Diese Grammatik wurde sowohl für das Programm „Wordspotting“ als auch für das Programm „Push 2 Speak“ verwendet.
- Die Grammatik wurde mit dieser Syntax für die Loquendo-Applikation verwendet. Die Syntax für Nuance wich davon ab, es wurde die von Nuance entwickelte Grammatiksprache verwendet. Da nur die Syntax von der im Anhang beschriebenen Grammatik abweicht, wurde darauf verzichtet, beide Grammatiken anzufügen.

- Die Grammatik wurde in drei Regeln unterteilt. Die Regel „wordspotting“ enthält das Schlüsselwort, welches für das Programm „Wordspotting“ gebraucht wurde. Die Regel „command“ enthält alle nötigen und in den beiden Programmen implementierten Befehle. Die Regel „information“ wurde dafür benötigt, die Regel „command“ übersichtlicher zu gestalten.
- Dem Programm wird nicht das erkannte Wort, sondern das bei <*@result* > unter Anführungszeichen gestellte Wort mit einer entsprechenden Gewichtung zurückgeliefert.

Beim Erstellen einer Grammatik ist es wichtig, dass sich die einzelnen Befehle möglichst stark voneinander unterscheiden, um die Klassifizierung des erkannten Wortes zu erleichtern. Wenn zwei Wörter, zum Beispiel „Sprecher“ und „Sprecherin“ nahezu gleich sind, kann es vorkommen, dass sie verwechselt werden. Darum wurde bei der hier verwendeten Grammatik als Befehl der Wortlaut „männlicher Sprecher“ beziehungsweise „weibliche Sprecherin“ ausgewählt. Hier ist eine Unterscheidung deutlich nachvollziehbar.
Auch ist es erstrebenswert, nur eine kleine Menge an Befehlen zur Verfügung zu stellen. Je mehr Befehle eine Grammatik erkennen soll, desto schwieriger ist es, diese alle auch zu unterscheiden. Bei einer großen Anzahl von Befehlen ist die Wahrscheinlichkeit, dass sich zwei ähneln und diese dann falsch erkannt werden, vergleichsweise hoch.
Ganze Sätze zu interpretieren, ist eine noch größere Herausforderung. Hierzu gibt es prinzipiell zwei Möglichkeiten: Einerseits kann die Grammatik so gestaltet werden, dass die Satzstellung und die Sätze exakt vorgegeben sind. Das hat den Vorteil, dass die Sätze mit großer Wahrscheinlichkeit richtig erkannt werden, insofern sie eben in genau der Satzstellung gesprochen werden, die vorgegeben ist. Es müssen viele Kombinationsmöglichkeiten und Satzstellungen berücksichtigt werden, um alle möglichen Eingaben abzudecken. Das hat aber den Nachteil, dass die Grammatik schnell unübersichtlich wird. Andererseits besteht auch die Möglichkeit, einen „Müllsammler“ (garbage collector) in die Grammatik einzubauen. Das bedeutet, die gewünschten Befehle, die erkannt werden sollen, anzugeben, die herum gesprochenen Wörter allerdings zu ignorieren. Bei dem Satz „Bitte geschichtliche Informationen vorlesen.“ wird alles, bis auf die relevanten Wörter „geschichtliche Informationen“,

ignoriert. So können theoretisch so gut wie alle Befehle, die in Sätzen eingepackt sind, erkennt werden, obgleich hier auch Negationen übersehen werden. Dies hat aber den Nachteil, dass die Erkennungsrate deutlich sinkt, wenn die Befehle von anderen Wörtern eingekleidet sind, die allesamt ignoriert werden. Es wird empfohlen, den garbage collector so wenig wie möglich zu verwenden.

4.3.2 Programm „Wordspotting“

Das Programm „Wordspotting“ hat ein multimodales Interface. Hier können sowohl sprachliche Befehle, als auch Befehle mittels Pen eingegeben werden. Die nächste Abbildung zeigt folgenden Status des Programmes: Beim linken Bild zeigt das Gerät auf das Austria-Center und der Benutzer hat bereits das Schlüsselwort „Eingabe“ gesagt. Das Programm wartet auf einen sprachlichen Befehl. Auch eine Peneingabe ist hier noch möglich. Rechts wurden geschichtliche Informationen zum Ares Tower ausgewählt und diese werden nun vorgelesen und angezeigt.

Abbildung 12: Programm „Wordspotting“ mit dem multimodalen Interface. Links wurde das Schlüsselwort „Eingabe“ bereits richtig erkannt und das Programm wartet auf einen akustischen Befehl. Rechts werden soeben geschichtliche Informationen zum Ares Tower akustisch und visuell ausgegeben.

Das Programm hat mehrere Status. Wenn eine Befehlseingabe erwartet wird, kann diese entweder sprachlich oder auch mittels Pen getätigt werden. Die sprachliche Eingabe funktioniert folgendermaßen: Das Schlüsselwort „Eingabe“ muss gesagt werden, wenn dieses richtig erkannt wird, was durch einen Hinweiston bestätigt wird, kann der gewünschte Sprachbefehl eingegeben werden. Bei der Beschreibung der verschiedenen Status wird keine Unterscheidung zwischen gesprochenem oder per Pen betätigten Befehl getroffen. Es wird jedesmal nur von Befehlseingabe gesprochen.

a) *Programmstart:* Beim Programmstart wird der Sensor via Bluetoothmanager ausgewählt. Nachdem der Sensor ausgewählt ist, wird zum nächsten Status gewechselt: Gebäude suchen.
b) *Gebäude suchen:* Bei diesem Status werden ständig die vom Sensor gelieferten Werte abgelesen. Falls sich in Blickrichtung ein Gebäude befindet, zu dem Informationen zur Verfügung stehen, wird in den Status „*Gebäude beschreiben*“ gewechselt. Hier wird ständig auf Befehlseingaben gewartet. In diesem Status können dem Programm einige Befehle, wie etwa die Ausgabestimme zu wechseln, gegeben werden.
c) *Gebäude beschreiben:* Nachdem ein Gebäude anvisiert ist, werden die verfügbaren Informationen aufgelistet. Hier wird auch weiterhin der Sensorwert überprüft. Falls das Gerät nicht mehr auf das gerade beschriebene Gebäude zeigt, wird wieder in den Status „*Gebäude suchen*“ gewechselt. Auch hier werden Befehlseingaben erwartet. Der Benutzer kann sich die Informationen vorlesen und gleichzeitig anzeigen lassen. Wenn der Benutzer Informationen auswählt, werden diese im Status „*Informationen anzeigen*“ zur Verfügung gestellt.
d) *Informationen anzeigen:* Hier werden dem Benutzer die von ihm gewünschte Information vorgelesen und auch angezeigt. Wenn das Vorlesen beendet ist, oder durch den Benutzer unterbrochen wird, wird zurück in den Status „*Gebäude beschreiben*“ gewechselt.

Die Initialisierung von ASR und TTS wird nach der Sensorauswahl vorgenommen. Die Statusabfragen laufen in einem eigenen Thread ab. Wenn das Programm eine Spracherkennung ausführen oder einen Text vorlesen soll, wird bei der Nuance-Implementierung ein eigener Thread aufgerufen, der diesen Task zu vollziehen hat. Bei

der Loquendo-Implementierung ist das im Statusthread eingebaut. Im Status „*Informationen anzeigen*“ wird sowohl ein Text vorgelesen als auch auf eine mögliche Eingabe (Bargein) gewartet.

4.3.3 Programm „Push 2 Speak“

Das Programm „Push 2 Speak“ kann alleine durch sprachliche Befehle gesteuert werden. Um einen Sprachbefehl geben zu können, muss der mittlere Knopf unterhalb des Displays gedrückt werden. Alle sprachlichen Ausgaben werden zusätzlich noch auf dem Display angezeigt. Das folgende rechte Bild zeigt das Programm in einem Status, wo Informationen über den Arestower ausgewählt wurden, und diese nun vorgelesen und angezeigt werden. Beim linken Bild hört der Benutzer soeben die Programmeinführung, wo die Bedienung der Anwendung erklärt wird.

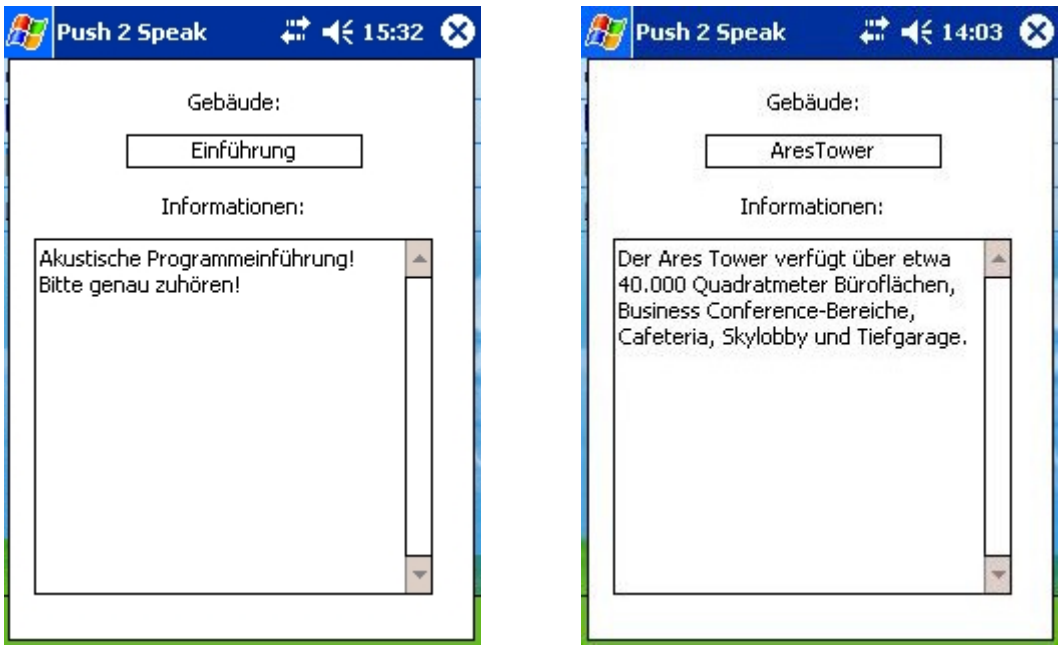


Abbildung 13: Programm „Push 2 Speak“. Links hört der Benutzer soeben die Programmeinführung. Rechts wurde bereits auf den Arestower gezeigt und Informationen darüber ausgewählt. Diese werden nun mittels TTS vorgelesen.

Der Programmaufbau von „Push 2 Speak“ ähnelt dem des Programms „Wordspotting“. Es hat etwa dieselben Status, bis auf den Unterschied, dass eine Einführung zur Verfügung steht. Die sprachliche Eingabe dieses Programms unterscheidet sich vom sprachlichen Interface des „Wordspotting“ Programms. Hier wird nicht auf das Schlüsselwort „Eingabe“ gewartet, sondern es kann ein Sprachbefehl erst dann eingeben werden, wenn zuerst der Hardwarebutton betätigt wird, der sich in der Mitte unterhalb

des Displays von iPAQ h5550 befindet (siehe *Abbildung 15: HP iPAQ h5550 Pocket PC*). Durch einen Signalton wird dem Benutzer mitgeteilt, dass das Programm auf einen Sprachbefehl wartet. Das Ende des Befehls muss wiederum mittels Betätigen desselben Hardwarebuttons bestätigt werden. Die Informationen, die bei dem Programm „Push 2 Speak“ ausgegeben werden, sind dieselben wie beim anderen Programm. Es werden nur einige Ausgaben erweitert, sodass sich die akustische Ausgabe von der visuellen unterscheidet. Damit kann im abschließenden Benutzertest die darauf bezogene Forschungsfrage beantwortet werden.

Folgende Status kann das Programm durchlaufen:

a) *Programmstart:* Beim Programmstart wird der Sensor via Bluetoothmanager ausgewählt. Nachdem der Sensor ausgewählt ist, wird zum nächsten Status gewechselt: „*Einführung*“.

b) *Einführung:* Hier wird dem Benutzer die Bedienung des Programms näher gebracht. Es wird ihm die Spracheingabe erklärt. Um zu testen, wie Sprachbefehle eingegeben werden sollen, muss er im Zuge der Einführung auf einen männlichen Sprecher umstellen. Nach erfolgreicher Erledigung dieser Aufgabe wird die Progrmmidee erläutert. Nach Beendigung der Einführung wird automatisch in den Status „*Gebäude suchen*“ gewechselt.

c) *Gebäude suchen:* Bei diesem Status werden ständig, wie beim Programm „Wordspotting“ die vom Sensor gelieferten Werte abgelesen. Falls sich in Blickrichtung ein Gebäude befindet, zu dem Informationen zur Verfügung stehen, wird in den Status „*Gebäude beschreiben*“ gewechselt. Es ist hier möglich, anhand der in der Einführung erklärten sprachlichen Befehlseingabe, einige Befehle an das Programm zu geben.

d) *Gebäude beschreiben:* Nachdem ein Gebäude anvisiert ist, werden die verfügbaren Informationen aufgelistet. Hier wird auch weiterhin der Sensorwert überprüft. Falls das Gerät nicht mehr auf das gerade beschriebene Gebäude zeigt, wird wieder in den Status „*Gebäude suchen*“ gewechselt. Eine eventuell stattfindende Sprachausgabe wird unterbrochen. Auch hier werden Befehlseingaben erwartet. Der Benutzer kann sich die Informationen vorlesen und gleichzeitig anzeigen lassen. Wenn der Benutzer Informationen auswählt, werden diese im Status „*Informationen anzeigen*“ zur Verfügung gestellt.

e) *Informationen anzeigen:* Hier wird dem Benutzer die von ihm gewünschte Information vorgelesen und auch angezeigt. Wenn das Vorlesen beendet ist, oder durch den Benutzer unterbrochen wird, wird zurück in den Status „*Gebäude beschreiben*“ gewechselt.

Da das Programm rein sprachlich gesteuert wird, kann es auch nur sprachlich beendet werden. Das ist in jedem Status möglich. Die Initialisierung und der Aufbau der Threads entspricht denen des Programms „Wordspotting“. Da das hier beschriebene Programm „Push 2 Speak“ dieselben Aufgabe erfüllt wie die erste Variante, unterscheidet es sich nur geringfügig. Beim Programm „Wordspotting“ wird der sprachliche Befehl durch eine Sprachpause beendet, die sowohl die Loquendo- als auch Nuancesoftware automatisch erkennt. Beim Programm „Push 2 Speak“ wird der Inputstream händisch durch Betätigung des Tastendrucks beendet. Wenn das Programm richtig bedient wird, ist so der zu erkennende Sprachbefehl eindeutig abgegrenzt.

4.4 Ergebnisse der Implementierung

Im Zuge der Implementierung konnte auch auf einige Forschungsfragen genauer eingegangen werden. Die Antworten werden nun zusammengefasst:

- *Forschungsfrage 6:* Die Frage, wie interessant mit den verfügbaren Möglichkeiten die Sprachausgabe gestaltet werden kann, wurde schon am Beginn der Implementierungsphase beantwortet. Aufgrund der verfügbaren Hardware musste auf die kleinstmögliche Ausgabequalität zurückgegriffen werden. Dementsprechend war die Qualität der Sprachausgabe nicht sehr hoch, obwohl auch die niedrigsten Qualitäten schon eine schöne Sprachmelodie zur Verfügung stellten. Es wurden allerdings noch nicht alle Wörter richtig betont. Logischerweise und vor allem bei Fremdwörtern tat sich die Sprachausgabe schwer, da diese nur auf Deutsch zur Verfügung standen.
- *Forschungsfrage 12:* Im Zuge der Implementierung wurde auch herausgefunden, ob Bargein technisch machbar ist. Im Falle von Nuance war diese Frage mit ja zu beantworten. Bei Loquendo wurden höhere Hardwareanforderungen gestellt, sodass dies nicht möglich war. Ob Bargein auch erwünscht und als sinnvoll angesehen wurde, wurde beim abschließenden Benutzertest geklärt.

- *Forschungsfrage 15:* Im Zuge dieser Diplomarbeit wurde herausgefunden, dass Spracherkennung und Sprachsynthese auch schon auf dem Targetsystem selbst machbar war. Allerdings musste auf einige Features verzichtet werden und etwa bei der Sprachausgabe auf das kleinste verfügbare Modul und somit die geringste Qualität zurückgegriffen werden.

4.5 Zusammenfassung

Im Zuge der Implementierung wurde die Software von zwei verschiedenen Anbietern (Nuance und Loquendo) evaluiert. Es wurde untersucht, was mithilfe von Sprachdialogsystemen möglich war. Eine deutliche Einschränkung erfuhren die entwickelten Prototypen von der Leistungsfähigkeit der Hardware. Die Sprachsoftware verlangte sowohl viel Speicherplatz als auch viel Rechenleistung. Wie gut die im Zuge der Diplomarbeit entwickelten Anwendungen angenommen wurden und inwiefern Sprachdialoge mit den zur Verfügung stehenden Möglichkeiten den Bedürfnissen der Benutzer genügten, zeigt das fünfte Kapitel, in dem der abschließend durchgeführte Benutzertest beschrieben wird.

5 Benutzerevaluation mit den entwickelten Prototypen

5.1 Überblick

Das fünfte Kapitel der Diplomarbeit beschäftigt sich mit der abschließenden Benutzerevaluierung. Es wurden Leute eingeladen, um die fertig entwickelten Anwendungen, die nach den Wünschen der im ersten Benutzertest getesteten Personen konzipiert wurden, ausprobieren und bewerten konnten. Für diesen Benutzertest wurde das LiLiPUT-System (Lightweight Lab Equipment for Portable User Testing in Telecommunications) verwendet, welches ebenfalls in diesem Kapitel vorgestellt wird. Die Durchführung der Benutzerevaluierung wird beschrieben und die Ergebnisse der Evaluierung werden besprochen. Es wurden nicht nur die Antworten der Benutzer protokolliert, sondern auch jede Erkennung beziehungsweise Fehlerkennung von Sprachbefehlen, um eine genaue Erkennungsrate errechnen zu können. Hierfür erwies sich das LiLiPUT-System als sehr hilfreich. Die verwendete Hardware, wie das Testgerät selbst und das Sensormodul werden beschrieben. Der genaue Testplan befindet sich im Anhang der Diplomarbeit.

5.2 Einleitung

Der Benutzer hat von Anfang an bei der Entwicklung der „Point 2 Discover“ Software und dadurch auch für die Entwicklung der „Point and Speak“ Anwendungen eine zentrale Rolle gespielt. Zuerst wurde er nach seinen Wünschen und Vorstellungen zu dieser Art von Diensten befragt. Nachdem die neue Software nun möglichst nach den Anforderungen der zukünftigen Benutzer gestaltet wurde, galt es nun zu untersuchen, ob der Benutzer damit auch zufrieden war. Schließlich ist es ein Unterschied, sich ein Programm vorstellen zu müssen und Kommentare dazu abzugeben, als dann einen fertigen Prototypen einer Anwendung kennen zu lernen und ausprobieren zu können.

5.2.1 LiLiPUT

Die Benutzertests wurden mit Hilfe eines zuvor am ftw. entwickelten mobilen Testlabors, das LiLiPUT [ReFr07] [FrRe06] (Lightweight Lab Equipment for Portable User Testing in Telecommunications) genannt wird, durchgeführt. Benutzertests bei mobilen Applikationen wie etwa bei der hier durchgeführten Benutzerevaluierung waren bisher meistens an stationäre Labore geknüpft. Wenn die Tests visuell und auditiv im Freien aufzeichnen werden wollten, musste man eine große Menge an Equipment mitführen.

Das Projekt LiLiPUT stellt für Tests eine neue Art der Aufzeichnung zur Verfügung, die vor allem für Benutzerstudien im Bereich der mobilen Mensch-Maschine Interaktion entwickelt wurde. Ziel von LiLiPUT war es, ein komplett drahtloses System zu entwickeln, das den Benutzer vom Tragen von Kabeln oder eines Rucksackes befreit. Er bekommt lediglich einen Hut aufgesetzt, der mit Kameras, Batterien und Sender ausgerüstet ist. Somit ist die Testperson in ihrer Bewegungsfreiheit nicht eingeschränkt und es ist eine möglichst reale Testsituation gegeben. Ein Beobachter, oder auch der Testleiter, trägt die restliche Ausrüstung, wozu etwa ein Notebook und die Empfänger zählen. *Abbildung 14*stellt das System anschaulicher dar:

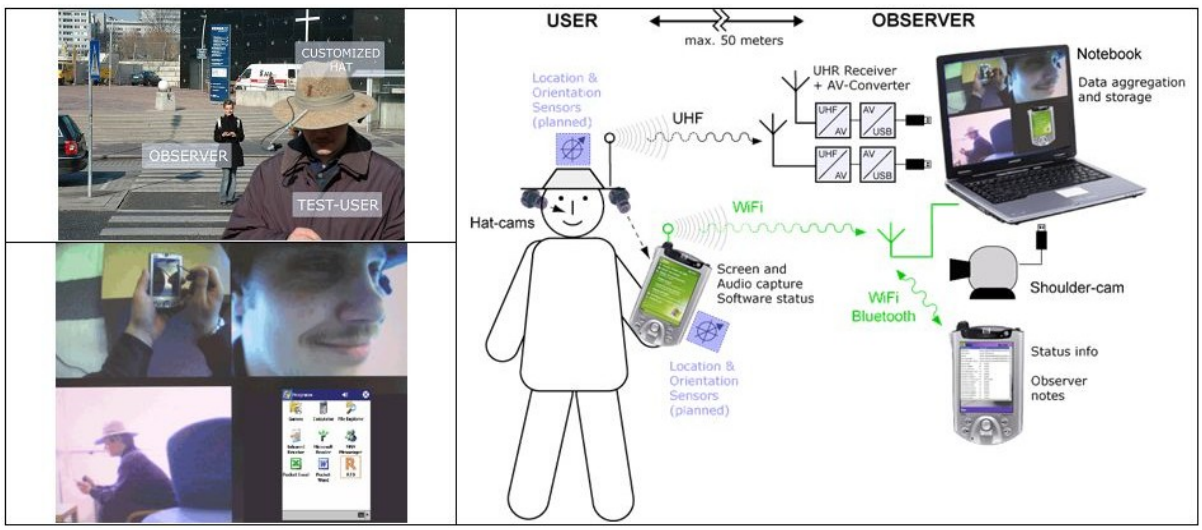


Abbildung 14: LiLiPUT im Einsatz. Das System im Freien (links oben), empfangene Videosignale (links unten) und die Systemarchitektur (rechts).
Einige Features sind noch in Planung.

Eine Kamera ist auf das Gesicht der Testperson gerichtet, um deren Reaktionen festhalten zu können. Die zweite Kamera filmt die Blickrichtung des Benutzers. Somit kann die Interaktion des Benutzers mit der Umgebung oder auch mit dem Testgerät aufgezeichnet werden. Es existiert auch eine Umgebungskamera, die der Observer trägt, um den Test von einer größeren Entfernung beobachten zu können.
Für die Audioaufnahme wurde Ein- und Ausgang des Pocket PC Headsets direkt abgegriffen. Auf diese Weise konnten sowohl die Benutzeräußerungen, als auch die Systemäußerungen mitgeschnitten werden.
Diese vielseitige Aufzeichnung von Benutzertests erleichtert deren Auswertung deutlich, da im Nachhinein einzelne Situationen gut nachvollziehbar sind und so mehr Informationen aus einem Benutzertest gefiltert werden kann, als wenn nur auf dem Testplan mitgeschrieben wird.

5.2.2 Testgerät und Sensoren

Die Anwendungen wurden unter eMbedded Visual C++ 4.0 für den iPAQ h5550 entwickelt.

Abbildung 15: HP iPAQ h5550 Pocket PC
Quelle: http://www.mobilezone.com.br

Der iPAQ h5550 Pocket PC von HP wurde gewählt, weil er von den zu Verfügung stehenden Geräten die gewünschten Anforderungen am Besten erfüllt. Der Vorteil des iPAQ h5550 gegenüber dem VPA IV, mit dem der erste Benutzertest durchgeführt wurde, liegt darin, dass er mehr Speicher bereitstellt. Die automatische Spracherkennung (ASR) und die Sprachsynthese (TTS) von Nuance benötigen viel Speicherplatz, der auf dem moderneren VPA IV nicht zur Genüge zur Verfügung steht. Allerdings liegt die Rechenleistung des iPAQ h5550 (400 MHz) unter die des VPA IV (520 MHz). Die Rechenleistung des Testgerätes war für die entwickelten Anwendungen ausreichend, allerdings musste auf ein zusätzliches Feature von LiLiPUT verzichtet werden: Das Abgreifen des Displays vom Pocket PC via WLAN war nicht möglich, da dieses zu viel Rechenkapazität benötigte und dadurch die Testapplikationen zu sehr beeinträchtigt hätte.

Als Sensor stand eine von den Technikern des ftw. entwickeltes Sensorboard zur Verfügung, das via Bluetooth an den Pocket PC angekoppelt wurde.

Da das Sensormodul noch nicht im Gerät eingebaut war, sondern als externe Sensorbox zur Verfügung stand, hatte der Testleiter die zusätzliche Aufgabe, die Box immer in Blickrichtung des Benutzers mitzudrehen, um im Pocket PC eingebaute Sensoren zu simulieren.

5.2.3 Testpersonen und Testgelände

Elf Testpersonen wurden gebeten, die Programme zu testen und ihre Kommentare dazu abzugeben. Es wurden verschiedene Tasks vorbereitet, die von den Usern ausgeführt werden mussten. Hierbei hatten sie die Gelegenheit, die verschiedenen Interaktionsmöglichkeiten kennen zu lernen. Einerseits war das die übliche Pen-Interaktion, wie sie bei Pocket PCs üblich ist, als auch die zwei neuen, sprachlichen Eingabemöglichkeiten. Sie mussten mit Hilfe der drei verschiedenen Interaktionsmöglichkeiten Informationen über in der Nähe befindliche Gebäude einholen. So lernten die Benutzer einerseits, wie das Programm funktioniert und andererseits auch mit der sprachlichen Komponente umzugehen.

Testpersonen	Weiblich	Männlich	Gesamt
Anzahl	4	7	11
Min Alter	22	24	22
Max Alter	29	30	30
Durchschnittsalter	24,75	26	25,5
SMS / Tag	2,5	1,5	1,9
Internetnutzung / Woche in h	14	14	14
Telefonieren / Tag in min	25	26,2	25,7

Tabelle 3: Über zu den wichtigsten demographischen Daten der Testpersonen

Es wurden Testpersonen ausgewählt, die der Zielgruppe der Anwendung entsprachen. Sowohl Studenten als auch Beamte und Angestellte im Alter zwischen 22 und 30 Jahren wurden gebeten, an den Tests teilzunehmen. Vier weibliche und sieben männliche Personen nahmen an dem Test teil. Alle waren Handybesitzer und regelmäßige Internetuser, sodass der prinzipielle Umgang mit technischen Geräten kein Problem darstellte.

Die Tests wurden in der Donaucity durchgeführt. Um eine ungefähre Vorstellung der Testsituation und der Testumgebung zu erhalten, befindet sich im Anhang (Annex 2) der Testplan, dem eine Karte beigefügt ist. In dieser Karte sind die für den Test wichtigen Orte und Gebäude beschriftet.

5.2.4 Software

Bei der Implementierung wurde zuerst die Software von Loquendo, danach auch die von Nuance zur Verfügung gestellte automatische Spracherkennung (ASR) und Sprachsynthese (TTS) getestet. Sie wurden, wie im vorigen Kapitel beschrieben, den Anforderungen des Projekts angepasst und entsprechende Programme entwickelt. Für die Tests selber wurden die beiden verschiedenen Prototypen von Nuance verwendet. Zum einen das Programm „Push 2 Speak“ und zum anderen die „Wordspotting“ Anwendung. Auf einen Vergleich der beiden in der Entwicklung getesteten Softwarepakete von Nuance beziehungsweise Loquendo wurde im Zuge des abschließenden Tests verzichtet.

Programm „Push 2 Speak“:

Dem Programm „Push 2 Speak“ konnten nur sprachliche Befehle gegeben werden. Um der Anwendung einen Befehl geben zu können, musste der mittlere Knopf unterhalb des Displays gedrückt werden. Nach einer Bestätigung des Programms mittels eines Pieptons wurde dem Benutzer mitgeteilt, dass er nun einen Befehl sagen konnte. Das Ende des Befehls wurde vom User mit der erneuten Betätigung der Taste gekennzeichnet. Somit war der Befehl eindeutig eingegrenzt und das Programm konnte nun versuchen, diesen zu interpretieren.

Programm „Wordspotting“:

Mit Hilfe des Programms „Wordspotting“ wurde eine alternative sprachliche Eingabe versucht. Die Anwendung hörte ständig auf den Benutzer. Dieser musste das Wort „Eingabe“ sagen, um dem Programm einen Befehl geben zu können. Nach der Erkennung des Schlüsselwortes wurde zur Bestätigung wieder ein Piepton ausgegeben, der dem Benutzer signalisierte, den eigentlichen Befehl nun sprechen zu können. Das Ende des Befehls musste bei diesem Programm nicht bestätigt werden. Die Software erkannte das Befehlende anhand einer längeren Sprechpause automatisch.

Dieses Programm war im Vergleich zu „Push 2 Speak“ multimodal ausgelegt. Es konnte durch dieses auch mithilfe des Zeigestabs (Pens) navigiert werden. Die Ausgabe erfolgte sowohl sprachlich als auch textuell auf dem Display, wobei die sprachliche Aus- und Eingabe auch deaktivieren werden konnte.

5.3 Testablauf

5.3.1 Einleitung

Die Benutzertests wurden an drei aufeinanderfolgenden Tagen in der Donaucity abgehalten. Am ersten Tag fand ein Vortest statt, wobei die letzten technischen Schwierigkeiten noch beseitigt wurden.

Die Hälfte der Personen wurde im Freien getestet. Am abschließenden Testtag wurden die Personen im Gebäudeinneren getestet, um einen Vergleich erzielen zu können, ob die Erkennungsrate im Inneren eines Gebäudes besser ist als im Freien. Im Freien sind viele Hintergrundgeräusche vorhanden, die das Ergebnis der Spracherkennung beeinflussen können. Dies stellte bezogen auf das Erkenntnisinteresse eine zumutbare Beeinträchtigung des Tests dar. Die Testperson musste sich nun die Gebäude, die sie beim outdoor-Test noch vor sich gesehen hat, vorstellen. Der Sensor funktionierte im Gebäudeinneren noch genauso, da für den Test ein bestimmter Punkt im Freien angenomnmen worden ist. Von diesem Punkt aus wurden im Vorhinein die interessanten Gebäude mittels des Sensors ausgemessen.

5.3.2 Demographische Fragen

Zuerst wurden den Testpersonen Fragen zu ihrem persönlichen Hintergrund gestellt. Alter, Beruf, Ausbildung wurden hier genauso abgefragt wie der Grad der Vertrautheit mit technischen Geräte. Es wurde ebenfalls in Erfahrung gebracht, ob sie das Handy, wovon jede Testperson eines besaß, nur dazu benutzen, um zu telephonieren und SMS zu schreiben oder auch für andere Dienste wie etwa Organizer oder als Navigationsgerät.

5.3.3 Block 1

Am richtigen Standort angekommen, wurde das LiLiPUT System zur Aufzeichnung bereitgemacht. Die Testperson bekam den LiLiPUT-Hut aufgesetzt und die Aufnahme wurde gestartet. Dann wurde der Testperson das Testgerät in die Hand gegeben und der Test begonnen. Die erste zu testende Anwendung war das „Push 2 Speak“ Programm. Hier gab es eine selbsterklärende Einführung vom Programm, wo die Testperson lernte, wie die Anwendung zu bedienen war. Hierauf folgten verschiedene Aufgaben. Der Benutzer musste Informationen zu verschiedenen Gebäuden einholen, was er sowohl mit als auch ohne Headset versuchte. Er wurde unter anderem gebeten, die Ausgabe zu unterbrechen. Bei einigen Informationen wurden visuell und auditiv verschiedene Informationen ausgegeben. Hier versuchte die Testperson, sich beide Ausgaben zu merken. Zum Schluss musste sie noch das Programm beenden, wieder mit Hilfe eines weiteren Sprachbefehls.

Am Anschluss an diesen Block wurden der Testperson Fragen zu dem Programm und dessen Bedienung gestellt. Ein genauer Testplan ist als Anhang (Annex 2) verfügbar. Dort sind die Aufgaben für den Benutzer genau beschrieben und auch die Fragen aufgelistet.

5.3.4 Block 2

Nachdem die Fragen zu Block eins beantwortet wurden, konnte der Benutzer das zweite Programm testen. Hierbei handelte es sich um das „Wordspotting“-Programm. Diese Anwendung hatte ein multimodales Interface. Der Testperson wurde erklärt, wie dieses Programm zu bedienen ist. Die Aufgabenstellung war hinsichtlich der Informationsbeschaffung ähnlich wie beim ersten Block. Der Unterschied lag großteils an den Interaktionsmöglichkeiten. Hier konnte die Testperson sowohl die Pen-Navigation (siehe *Abbildung 16*) als auch die zusätzliche sprachliche Navigation ausprobieren.

Nachdem auch das zweite Programm zur Genüge ausgetestet war, wurden weitere Fragen gestellt. Die Testperson musste etwa die drei verschiedenen Interaktionsmöglichkeiten in eine subjektive Reihenfolge bringen, was sie am Liebsten verwenden würde.

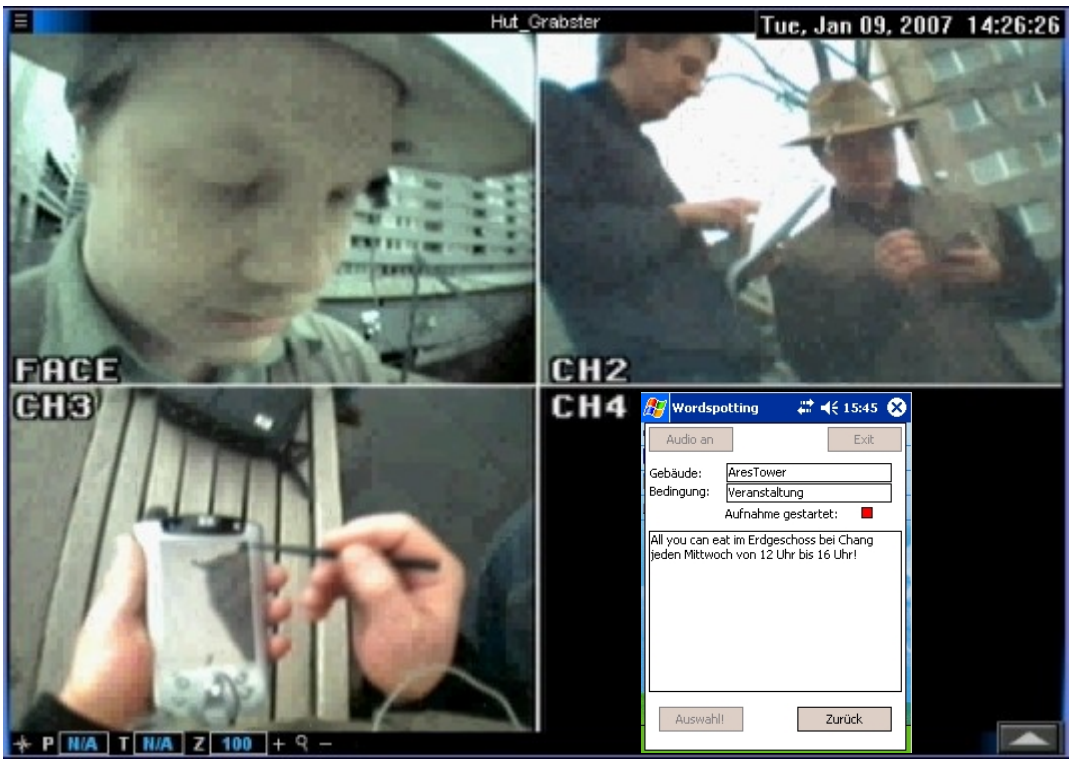


Abbildung 16: Aufzeichnung via LiLiPUT. Der Benutzer steuert soeben mit dem Pen durch das Programm. Es sind 3 verschiedene Kamerapositionen verfügbar und das Display des PocketPCs ist sichtbar.

Die Anwendungen, die bei den abschließenden Benutzertests getestet wurden, bezogen sich auf die einfachste Darstellungsart von digitalen Informationen in einer physikalischen Umgebung, der Listenform. Der Testperson wurden dann die restlichen verfügbaren Darstellungsarten anhand eines Ausdruckes erklärt. Diese Darstellungen sind dem Benutzertest, der im Anhang zu finden ist, beigefügt. Er konnte sie nicht so wie beim ersten Benutzertest, ausprobieren. Bei der Benutzerevaluierung wurde der Testperson, die nun die sprachliche Interaktion kennen gelernt hatte, gefragt, wie ihr die restlichen Darstellungsarten gefallen, und ob sie sich auch hier eine sprachliche Kommunikation mit dem Gerät vorstellen könnten. Diese Aufgabe war als Ideensammlung gedacht, um herauszufinden, ob es Sinn macht, etwa für eine Radaransicht ein sprachliches Interface zu entwerfen.

5.4 Ergebnisse

Jeder der beiden Blöcke deckte eine Reihe der in Kapitel zwei gestellten Forschungsfragen ab. Darum sind die Ergebnisse der abschließenden Benutzerevaluation den Blöcken zugeordnet.

5.4.1 Block 1

- *Forschungsfrage 3:* Die Idee, akustisch und visuell unterschiedliche Informationen auszugeben, wurde vom ersten Benutzertest aufgegriffen. Die Frage, ob unterschiedliche visuelle und akustische Informationen hilfreich sind konnte mithilfe des „Push 2 Speak“ Programmes beantwortet werden. Nachdem die Benutzer bei bestimmten Informationen visuell und akustisch unterschiedliche Ausgaben gehört hatten, sagten diese einstimmig, dass dies zuviel Information sei. Es erfordert eine zu große Konzentration, sich beides merken zu können.
- *Forschungsfragen 10:* Die erste Anwendung „Push 2 Speak“ konnte nur sprachlich gesteuert werden. Durch die Einführung hat der Benutzer die Programmsteuerung kennen und damit umgehen gelernt. Sämtliche Aufgaben bei Block 1 wurden sprachlich ausgeführt. Es wurden von dem Programm zu den interessanten Gebäuden immer alle verfügbaren Informationsarten auch vorgelesen, sodass die Ausgabe am Display nur eine zusätzliche Ergänzung darstellte. Bei den Fragen am Ende diese Blockes wurden die Benutzer nach der Gebrauchstauglichkeit des Programmes gefragt. Die Testpersonen waren durchwegs mit der Einfachheit des sprachlichen Interfaces zufrieden (Durchschnittsnote: 1.3).
- *Forschungsfrage 11:* Auch die Auswahl der Befehle und die Gestaltung des Dialoges wurde als intuitiv angesehen (Durchschnittsnote: 1.3), sodass die Steuerung durch das Programm keine größeren Schwierigkeiten machte.
- *Forschungsfrage 12:* Bei Block 1 gab es eine Aufgabe, bei der die Testperson das Bargein ausprobieren konnten. Bei einer längeren Ausgabe konnte der Benutzer versuchen, die Ausgabe zu unterbrechen. So wurden auch gleich die dafür am meisten verwendeten Befehle herausgefunden. „Stop“, „Abbrechen“ und „Beenden“ waren hier die am häufigsten verwendeten Kommandos, wobei

die Anwendung allerdings nur „Stop“ verstand. Die Möglichkeit des Bargein wurde auch vom Großteil der Tester als positiv aufgenommen (Durchschnittsnote: 1.6). Vor allem bei längeren Ausgaben oder bei einer falschen Verzweigung des Programmes durch einen missverstandenen Sprachbefehl wurde das als sehr nützlich empfunden.

5.4.2 Block 2

- *Forschungsfrage 19:* Eine Frage, die sich auf das Programm „Wordspotting“ bezog, war es, ob sich der Benutzer überwacht fühlt, wenn die Anwendung ständig auf Eingaben des Benutzers wartet. Bei dieser Anwendung konnte ein Befehl gegeben werden, wenn vorher das Schlüsselwort „Eingabe“ gesagt wurde. Zehn Testpersonen fühlten sich überhaupt nicht, eine Testperson eher nicht überwacht.
- *Forschungsfrage 8*: Eine interessante Frage ist die, wie der Benutzer mit der Anwendung kommunizieren will. Keine einzige Testperson wollte mit der Testapplikation mithilfe von längeren Sätzen kommunizieren. Die Anwendung wurde nicht als menschlicher Kommunikationspartner gesehen, sondern als Hilfsmittel, dem Befehle gegeben werden können. Je kürzer die Befehle sind, desto besser.
- *Forschungsfrage 9*: Auch die Frage, welche Befehle verwendet werden, wurde beantwortet: Möglichst einfach und intuitiv sollten sie sein. Zum Beispiel sagten die Testpersonen „Veranstaltungen“, wenn sie wissen wollten, welche Veranstaltungen in einem gerade anvisierten Gebäude in nächster Zeit stattfinden. Die Befehle, die in den beiden Anwendungen implementiert waren, wurden als ausreichend angesehen. Manchmal wurden mehr Alternativen gewünscht, etwa um eine Ausgabe zu unterbrechen.
- *Forschungsfrage 5*: Im Laufe der verschiedenen Aufgaben für die Testpersonen wurden zwei verschiedene Stimmen gehört. Eine weibliche und eine männliche Stimme. Bei den abschließenden Fragen wurden die Testpersonen gebeten, diese zu benoten und ihnen auch Eigenschaften zuweisen. Der männliche Sprecher bekam nach dem Schulnotensystem durchschnittlich 1.8, die weibliche Sprecherin 2.4. Auffallend war hier, dass sowohl die weiblichen, als auch die männlichen Testpersonen die männliche Stimme bevorzugten. Eigenschaften

waren den Stimmen nur schwer zuzuerkennen, da deutlich gehöret werden konnte, dass es keine echten Personen, sondern Computerstimmen waren. Eine einzige Testperson konnte den Stimmen Eigenschaften wie kompetent und professionell zuordnen.

- *Forschungsfrage 14:* Es wurden zwei verschiedene Möglichkeiten implementiert, die der Benutzer kennen gelernt hat, um sprachlich mit dem Pocket PC zu kommunizieren. Einerseits durch das Sagen des Schlüsselwortes „Eingabe“, andererseits durch das Einschließen des Befehls zwischen zwei Tastenbetätigungen. Von den beiden Sprachinteraktionen wurde „Push 2 Speak“ bevorzugt. Als Begründung wurde angegeben, dass das Schlüsselwort oft nicht gleich erkannt wurde und diese Interaktionsmöglichkeit wiederum schneller funktioniert als das „Wordspotting“, da die Eingabe des Schlüsselwortes entfällt.
- *Forschungsfrage 16:* Interessanterweise wurde die Frage nach der Belästigung durch Spracheingabe widersprüchlich zum ersten Benutzertest beantwortet. Haben im ersten Benutzertest noch so gut wie alle getesteten Personen gesagt, sie würden sich gestört fühlen, wenn jemand neben ihnen das Programm sprachlich steuert, so haben in der abschließenden Benutzerevaluierung die Testpersonen auf diese Frage geantwortet, dass es sie nicht stören würde. Eine einzige Person hat geantwortet, sie würde sich eher nicht belästigt fühlen. Alle anderen fühlen sich gar nicht belästigt. Die Durchschnittsnote (1 stand für starke Belästigung, 5 stand für keine Belästigung) beträgt 4,9.
- *Forschungsfrage 17:* Die Frage, ob die zusätzliche Belästigung durch Sprachausgabe stört, wurde auch eher mit nein beantwortet, allerdings nicht so klar wie bei Forschungsfrage 16. Die Durchschnittsnote beträgt hier 3,7. Auch hier ist ein klarer Unterschied zur Beantwortung dieser Frage bei der Benutzerevaluierung festzustellen. Eine Testperson würde eher zuhören, wenn jemand neben ihr das Programm verwenden würde. Es könnten ja auch für sie interessante Informationen dabei sein.
- *Forschungsfrage 18:* Beim Vergleich der Forschungsfragen 16 und 17 wird klar, dass die Verwendung eines Headsets sinnvoll ist. Es würden sich dadurch weniger Leute belästigt fühlen.
- *Forschungsfrage 20:* Nachdem die Benutzer das Programm kennen gelernt haben, konnten diese recht gut damit kommunizieren. Die gestellten Aufgaben,

wie etwa historische Informationen herauszufinden, wurden ohne größere Schwierigkeiten erledigt. Einzig die Erkennung der Sprachbefehle war manchmal noch ein Hindernis. Die Navigation im Programm selbst war kein Problem.

- *Forschungsfrage 21:* Auch das von den Anwendungen gegebene Feedback war ausreichend. Etwa der Signalton, der darauf hinwies, dass nun ein Sprachbefehl eingeben werden konnte.

Beim abschließenden Benutzertest wurden nicht nur die beiden verschiedenen entwickelten sprachlichen Interaktionsmöglichkeiten miteinander verglichen, sondern auch der normalen Pen-Interaktion gegenüber gestellt. Die Peninteraktion wurde von den meisten Benutzern bevorzugt. Als Grund dafür wurde angegeben, dass sie einfach und schnell auszuführen ist. Außerdem ist auch die Wahrscheinlichkeit einer Fehlerkennung wie bei der sprachlichen Interaktion nicht vorhanden. Die sprachliche Möglichkeit wurde als interessante Alternative angesehen. Bei vielen Handys etwa gibt es keine Pen-Interaktion sondern eine Steuerung per Joystick oder den verschiedenen Nummerntasten. Wenn hier ein ausgereiftes Sprachinterface zur Verfügung stehen würde, würden das viele Benutzer vorziehen. Es wurde von den getesteten Personen auch angemerkt, dass Sprachinteraktion sinnvoller zu verwenden ist, wenn zu viele Auswahlmöglichkeiten zur Verfügung stehen. Da müsste mittels Pen erst durch die Auswahlmöglichkeiten navigiert werden. Sprachlich wäre das durch Ausprobieren oder durch oftmalige Bedienung des Programms möglich. Als geübter Benutzer wäre die sprachliche Steuerung dann schneller, als etwa erst per Pen durch die Auswahlmöglichkeiten durchzuscrollen und die gewünschte Information zu finden.

Von den beiden sprachlichen Interaktionsmöglichkeiten wurde die „Push 2 Speak“ Anwendung der „Wordspotting“ Anwendung vorgezogen. Als Grund hierfür wurde angegeben dass das Erkennen des Schlüsselwortes oft mehrere Versuche benötigt. Das Einbetten des Sprachbefehls zwischen zweimaligem Drücken der Taste funktioniert schneller und auch sicherer. Hier musste das Programm nur einen Befehl richtig erkennen, bei der „Wordspotting“ Anwendung allerdings zuerst den Befehl „Eingabe“ und dann noch den eigentlichen Sprachbefehl. Wenn die Erkennung des Schlüsselwortes mit größerer Wahrscheinlichkeit funktionieren würde, wäre diese Möglichkeit interessanter einzustufen, weil die Anwendung „Wordspotting“ dann komplett sprachlich steuerbar wäre, ohne auch nur eine Taste drücken zu müssen.

5.4.3 Erkennungsraten

Während des Testes wurden die Testpersonen gebeten, sprachliche Eingaben sowohl mit als auch ohne Headset zu versuchen. Mit Hilfe des LiLiPUT-Systems erfolgte dann im Nachhinein die Auswertung, wie oft das System Befehle richtig erkannt hat. Die Falscherkennungen wurden in zwei Untergruppen geteilt, einerseits in falsche Akzeptanz und andererseits in falsche Zurückweisung. Da einige Test im Inneren des Gebäudes und einige im Freien gemacht wurden, folgt auch hier eine Unterscheidung. Es folgt eine kurze Übersichtstabelle.

	Indoor	Outdoor	Mit Headset	Ohne Headset	Gesamt
Richtig erkannt	78,08%	75,82%	81,25%	70,41%	74,49%
Falsche Akzeptanz	8.22%	12,42%	9,38%	13,27%	10,29%
Falsche Zurückweisung	13,70%	11,76%	9,38%	16,33%	15,23%
Summe	100,00%	100,00%	100,00%	100,00%	100,00%

Tabelle 4: Übersichtstabelle zu den Erkennungsraten der Anwendungsprototypen, gerundet

Aus *Tabelle 4* ist ersichtlich, dass bei der Erkennung mit Headset die Erkennungsrate höher war als die bei der Erkennung ohne Headset. Allerdings hat die Erkennung ohne Headset auch erstaunlich gut funktioniert. Die Vermutung, die im Zuge der Entwicklung entstanden ist, dass die Qualität des im Pocket PC eingebauten Mikrophons jener des Mikrophons im Headset stark unterlegen ist, hat sich nicht bestätigt.

Positiv ist zu erkennen, dass die Spracherkennung den Vorteil, den ein ruhiges Zimmer gegenüber eines belebten und von Hintergrundgeräuschen geprägten Platzes im Freien hat, nutzen konnte. Die Erkennung des Gesprochenen war im Gebäudeinneren etwas besser als bei den Tests im Freien. Dass der Unterschied nicht größer ist liegt daran, dass die Hintergrundgeräuschunterdrückung bei der Spracherkennungssoftware schon sehr ansprechend funktioniert.

Insgesamt wurde eine Erkennungsrate von etwa 75% erreicht, also jeder vierte Befehl nicht erkannt. Allerdings konnte im Zuge der Tests festgestellt werden, dass die Erkennung mit Dauer des Testes zunahm, sich die Testperson also auf die Software und die Handhabung des Gerätes eingestellte. Es wurde deutlicher in das Mikrophon gesprochen oder auch das Mikrophon näher herangeholt. Auch die Fehlbedienungen, die in obiger Tabelle nicht aufgelistet sind, wurden weniger.

5.5 Zusammenfassung

Das fünfte Kapitel der Diplomarbeit beschreibt den abschließenden Benutzertest, der anhand der beiden entwickelten Anwendungen „Wordspotting“ und „Push 2 Speak“ durchgeführt wurde. Die Anwendungen und die dahinterstehende Idee der Informationsbeschaffung zu den umliegenden Gebäuden wurde gut angenommen. Die sprachliche Interaktionsmöglichkeit, die für viele Benutzer neu war, wurde als interessante Möglichkeit angesehen, dem Handy oder Pocket PC Befehle zu geben. Es gab kaum Schwierigkeiten, die verschiedenen sprachlichen Eingabemöglichkeiten zu erlernen. Die Erkennungsraten der zur Verfügung stehenden Software wurde unter möglichst realen Umständen ermittelt und es wurde festgestellt, dass der Großteil der Benutzer das Programm und das Sprachinterface mit der erreichten Genauigkeit auch verwenden würde.

6 Resumee und Ausblick

6.1 Überblick

Abschließend werden noch die interessanten Aspekte, die im Laufe dieser Arbeit gesammelt werden konnten, zusammengefasst. Der dafür besonders wichtige Benutzertest war die abschließende Benutzerevaluierung, wo die Testpersonen schon zwei fertige Prototypen mit funktionierenden Sprachdialogsystemen ausprobieren konnten. Es wird auch ein kurzer Ausblick gewagt, was die weitere technologische Entwicklung bringen kann und was für zukünftige Arbeiten die über diese Diplomarbeit hinausführen noch offen stehen.

6.2 Zusammenfassung

Im Rahmen der Entwicklung der Prototypen und beim abschließenden Benutzertest konnten interessante Fakten im Bereich der Spracherkennung und Sprachsynthese bei mobilen Geräten gesammelt werden. Die Entwicklung solcher Systeme ist in letzter Zeit schnell fortgeschritten, auch als eines in mobilen Geräten eingebautes Modul funktioniert die Spracherkennung und die Sprachsynthetisierung schon sehr gut. Die Sprachmelodie von technischen erzeugten Stimmen ähnelt schon der einer menschlichen, auch wenn die Qualität der Stimmen noch an der Beschränkung durch die Hardware leidet. Im Zuge der Implementierung musste aufgrund der verfügbaren Hardware auf das qualitativ schwächste Ausgabemodul der zur Verfügung stehenden Software zurückgegriffen werden, was dazu führte, dass zwar bei der synthetischen Sprachausgabe die menschliche Sprachmelodie und Wortbetonung gut simuliert werden konnte, allerdings trotzdem noch stark zu hören war, dass es sich um eine vom Computer erzeugte Stimme handelte.

Trotz der technischen Beschränkungen durch die zur Verfügung stehende Software wurde bei den Benutzertests die verfügbare Sprachinteraktion positiv aufgenommen. Die Benutzer sahen zum Großteil das erste Mal ein rein sprachlich zu steuerndes Programm oder ein multimodales Interface und fanden sich sehr schnell mit dem System zurecht. Obwohl es noch einige Fehlerkennungen gab, würden die meisten getesteten Personen das System, so wie sie es vorgefunden hatten, auch benutzen. Sowohl die Idee sich Informationen über interessante Punkte in der Umgebung holen zu

können („p2d“) als auch das Sprachdialogsystem fanden Anklang, sodass sich die Benutzer vorstellen konnten dies auch wirklich zu benutzen, etwa als Tourist in einer fremden Stadt.

6.3 Ausblick und zukünftige Arbeiten

Da die Entwicklung sowohl der Software als auch der Hardware weiter fortschreiten wird, ist anzunehmen, dass sich in naher Zukunft die Qualität sowohl der Spracheingabe als auch der Sprachausgabe weiterhin verbessern wird. Mit besseren Erkennungsraten werden die Falschinterpretationen durch die Software immer weniger und somit die Programmbedienung einfacher und schneller. Im abschließenden Benutzertest wurde von vielen Testpersonen die normale Peninteraktion vor den beiden verschiedenen ausprobierten Sprachdialogsysteme gereiht, weil die Peninteraktion schneller und sicherer funktionierte. Auch ist die normale Interaktion wie mit Pen oder auch Tastatur noch geläufiger als der sprachliche Dialog mit einem elektronischen Partner. Je häufiger Sprachdialogsysteme benutzt werden und mit einer weiteren technischen Verbesserung können die Qualitäten und Vorteile von solchen Systemen bald die Qualitäten von der normalen Interaktion überwiegen. Bei einem reinen Sprachinterface ist man dann nicht mehr auf das kleine Display angewiesen. Bei einer intuitiven oder auch häufig benutzten Dialogführung kann auf zusätzliche visuelle Ausgaben verzichtet werden, vor allem in dem Fall, wo Informationen wie Gebäudebeschreibungen oder Busfahrpläne, die auch akustisch gut dargestellt werden können, zur Verfügung gestellt werden. Sprachdialoge und multimodale Systeme können als echte Alternativen zu den klassischen Interaktionsmöglichkeiten gesehen werden.

Da die bei „Point and Speak“ entwickelten Prototypen nur für eine Darstellungsart des „p2d“ Projektes, der Listendarstellung, ein Sprachdialogsystem implementieren, kann als mögliche zukünftige Arbeit versucht werden, zu den anderen Applikationen wie etwa der erweiterten Realitätsansicht von „p2d“ ein solches System zu implementieren. Anregungen dafür wurden schon beim abschließenden Benutzertest gesammelt.

Bei vielen anderen Anwendungen sind Sprachdialogsysteme vorstellbar, so wie etwa beim Autofahren, wo die Hände für andere Benutzeraktionen verwendet werden sollten. Es gibt dafür schon etliche Anwendungen, die sprachlich gesteuert werden können. Hier werden mit Sicherheit in Zukunft noch mehr Systeme, die sprachlich zu steuern sind, entstehen.

Literaturliste

[SemO] semapedia.org: Hyperlink your world. www.semapedia.org
Letzter Zugriff: September 2006

[NaOr04] Luca Nardelli, Marco Orlandi, und Daniele Falavigna. A MultiModal Architecture for Cellular Phones.
In: Proceedings of the 6th international conference on Multimodal interfaces (ICMI '04) (Oktober 13–15, 2004, State College, Pennsylvania, USA) Publisher: ACM Press

[NuAS] Nuance: Automotive solutions from Nuance Broschüre; Quelle: www.nuance.com

[Pyss00] Tino Pyssysalo, Tapio Repo, Tuukka Turunen, Teemu Lankila, Juha Röning. CyPhone – Bringing Augmented Reality to Next Generation Mobile Phones.
In: Proceedings of DARE 2000 on Designing augmented reality environments (DARE '00) (April, 2000 Elsinore, Denmark) Publisher: ACM Press

[BuCh05] Stefano Burigat, Luca Chittaro. Location-aware Visualization of VRML Models in GPS-based Mobile Guides.
In: Proceedings of the tenth international conference on 3D Web technology (Web3D '05) (2005 Bangor, United Kingdom) Publisher: ACM Press

[HiPi00] Ken Hinckley, Jeff Pierce, Mike Sinclair, Eric Horvitz. Sensing Techniques for Mobile Interaction.
In: Proceedings of the 13th annual ACM symposium on User interface software and technology (UIST '00) (2000 San Diego, CA USA) Publisher: ACM Press

[HuDi04] Andreas Hub, Joachim Diepstraten, Thomas Ertl. Design and Development of an Indoor Navigation and Object Identification System for the Blind.
In: ACM SIGACCESS Accessibility and Computing , Proceedings of the 6th international ACM SIGACCESS conference on Computers and accessibility (Assets '04) (October 18–20, 2004, Atlanta, Georgia, USA) Publisher: ACM Press

[BaKu05] Lynne Baillie, Harald Kunczier, Hermann Anegg, Rolling, Rotating and Imagining in a Virtual Mobile World
In: Proceedings of the 7th international conference on Human computer interaction with mobile devices & services (MobileHCI '05) (September 19–22, 2005, Salzburg, Austria) Publisher: ACM Press

[StEs05] Steven Strachan, Parisa Eslambolchilar, Roderick MurraySmith. gpsTunes Controlling Navigation via Audio Feedback.
In: Proceedings of the 7th international conference on Human computer interaction with mobile devices & services (MobileHCI '05) (September 19–22, 2005, Salzburg, Austria) Publisher: ACM Press

[Lai2004] Jennifer Lai. Facilitating Mobile Communication with Multimodal Access to Email Messages on a Cell Phone.
In: CHI '04 extended abstracts on Human factors in computing systems CHI '04 (April 24–29, 2004, Vienna, Austria) Publisher: ACM Press

[WaOl03] Rainer Wasinger, Dominika Oliver, Dominik Heckmann, Bettina Braun, Boris Brandherm, Christoph Stahl. Adapting Spoken and Visual Output for a Pedestrian Navigation System, based on given Situational Statements.
In: Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen (ABIS 2003) (October 6-8, 2003, Karlsruhe, Germany)

[WaSt03] Rainer Wasinger, Christoph Stahl, Antonio Krueger. Robust speech interaction in a mobile environment through the use of multiple and different media input types.
In: 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003) (September 1-4, 2003, Geneva, Switzerland)

[WaStKr03] Rainer Wasinger, Christoph Stahl, Antonio Krüger. M3I in a Pedestrian Navigation & Exploration System.
In: 5th International Symposium on Human Computer Interaction with Mobile Devices (MobileHCI 2003) (September 8-11, 2003, Udine, Italy)

[SiKu05] Rainer Simon, Harald Kunczier, Hermann Anegg. Towards Orientation-Aware Location Based Mobile Services.
In: 3rd Symposium on LBS and TeleCartography (November 28-30, 2005, Vienna, Austria)

[HeOl04] Anders Henrysson, Mark Ollila. UMAR Ubiquitous Mobile Augmented Reality
In: Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia (MUM '04) (October 27-29 2004 College Park, Maryland, USA) Publisher: ACM Press

[ArKi] ARToolKid. http://www.hitl.washington.edu/artoolkit/
Letzter Zugriff: 02.02.2007

[LeNa00] Eun-Ju Lee, Clifford Nass, Scott Brave. Can Computer-Generated Speech Have Gender? An Experimental Test of Gender Stereotype
In: CHI '00 extended abstracts on Human factors in computing systems (CHI '00) (April 01-06 2000 The Hague, Netherlands) Publisher: ACM Press

[NaLe00] Clifford Nass, Kwan Min Lee. Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction.
In: Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '00) (April 01-06 2000 The Hague, Netherlands) Publisher: ACM Press

[NiKo05] Georg Nikfeld, Markus Kommenda. Skriptum zu “Ein- und Ausgabe von Sprache” WS 2005

[HPp2d] „Point 2 Discover“ http://p2d.ftw.at/
Letzter Zugriff: 02.02.2007

[FrSi06] Fröhlich, P., Simon, R., Baillie, L., and Anegg H. (2006).
Comparing Conceptual Designs for Mobile Access to Geo-Spatial Information.
In: Proceedings of the 8th conference on Human-computer interaction with mobile devices and services (MobileHCI '06) (September 12-15, 2006. Espoo, Finland)
Publisher: ACM Press

[SiFr06] Rainer Simon, Peter Fröhlich, Hermann Anegg. Beyond Location Based – The Spatially Aware Mobile Phone
In: Proceedings of the 6th International Symposium on Web and Wireless Geographical Information Systems (W2GIS 2006) (December 4-5, 2006, Hong Kong, China)

[NFCR] Near Field Communication Research Project at FH Hagenberg, Austria.
http://www.nfc-research.at/
Letzter Zugriff: 10.10.2006

[Egen99] Max J. Egenhofer. Spatial Information Appliances: A Next Generation of Geographic Information Systems
In: 1st Brazilian Workshop on GeoInformatics, 1999.

[ReFr07] Peter Reichl, Peter Fröhlich, Lynne Baillie, Raimund Schatz, Antitza Dantcheva. The LiLiPUT Prototype: A Wearable Lab Environment for User Tests of Mobile Telecommunication Applications
In: ACM International Conference on Human Factors in Computing Systems (CHI 2007) (San Jose, California, USA) ACM Press.

[FrRe06] Peter Fröhlich, Peter Reichl, Raimund Schatz, Lynne Baillie, Wolfgang Weinberger, Florian Hammer. LiLiPUT: Lightweight Lab Equipment for User Testing in Telecommunications
In: British HCI conference: ENGAGE (HCI 2006) (September 2006, London, UK)

[Nu07] Nuance. www.nuance.com
Letzter Zugriff: 02.02.2007

[Loq07] Loquendo. www.loquendo.com
Letzter Zugriff: 02.02.2007

[LTTSB] Loquendo Embedded TTS Broschüre; Quelle: www.loquendo.com

[LASRB] Loquendo Embedded ASR Broschüre; Quelle: www.loquendo.com

[Ber07] http://www.bertone.it/en/birusa_birusa_en.htm
Letzter Zugriff: 02.02.2007

[NuRS4] Nuance Realspeak™ Solo 4.0 Broschüre; Quelle: www.nuance.com

[W3C] JSpeech Grammar Format W3C Note 05 June 2000. http://www.w3.org/TR/jsgf/
Letzter Zugriff: 02.02.2007

[CrHi] “Creative Histories – The Josefsplatz Experience” project homepage.
http://www.josefsplatz.info/
Letzter Zugriff: 06.02.2007

[Ov03] Oviatt, S.L. Multimodal interfaces.
In The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, J. JACKO AND A. SEARS, Eds. Lawrence Erlbaum Assoc., Mahwah, NJ, 2003, chap.14, 286-304

[GoEa07] „Google Earth – erforschen, suchen und entdecken“ homepage
http://earth.google.de/
Letzter Zugriff: 17.02.2007

Anhang

Annex 1: Grammatik in JSGF Format

```
#JSGF V1.0 ISZ-8859-1;
grammar meinVersuch;


<wordspotting> =
(
                    Eingabe {<@result "Eingabe">});


<command> =
(                   zurück {<@result "zurück">}|
                    stop {<@result "stop">}|
                    Programm beenden {<@result "ende">}|
                    Programm schließen {<@result "ende">}|
                    weibliche Sprecherin {<@result "weiblich">}|
                    männlicher Sprecher {<@result "männlich">}|
                    (geschichtliche <information>       {<@result "geschichtlich">}|
                    historische <information>    {<@result "geschichtlich">}|
                    <information> über das Gebäude {<@result "Gebäude">}|
                    Gebäudeinformationen {<@result "Gebäude">}|
                    Entfernung {<@result "Entfernung">}|
                    Interessante <information> {<@result "Interessantes">}|
                    Interessantes {<@result "Interessantes">}|
                    <information> für Mieter {<@result "Mieter">}|
                    Mietinformationen {<@result "Mieter">}|
                    <information> über Veranstaltungen {<@result
"Veranstaltung">}|
                    Veranstaltungen {<@result "Veranstaltung">});

<information> =
(
                    info |
                    information |
                    infos |
                    informationen);

public <meinVersuch> = <wordspotting>|<command>;
```

Annex 2: Testplan von den abschließenden Benutzertests

Audio & Videoerlaubnis

Vielen Dank für die Teilnahme an der p2d Benutzer-Studie. Wir beabsichtigen, für Forschungszwecke sowohl Audio- als auch Videoaufnahmen von diesem Test zu machen. Diese erleichtert uns die Datenanalyse und ermöglicht es uns, Ausschnitte der Tests Projektpartnern oder anderen Projektmitarbeitern zu zeigen. Bitte lesen Sie sich die untenstehende Erklärung durch und unterschreiben Sie bitte an der dafür vorgesehenen Stelle.

Erklärung:

Ich bin mir bewusst, dass während des Benutzertests Video- und Audioaufnahmen gemacht werden und bin damit einverstanden, dass diese Aufnahmen für Forschungszwecke bzw. zur Vorführung bei Projektpartnern oder Projektmitarbeitern verwendet werden.

Name in Blockschrift _____

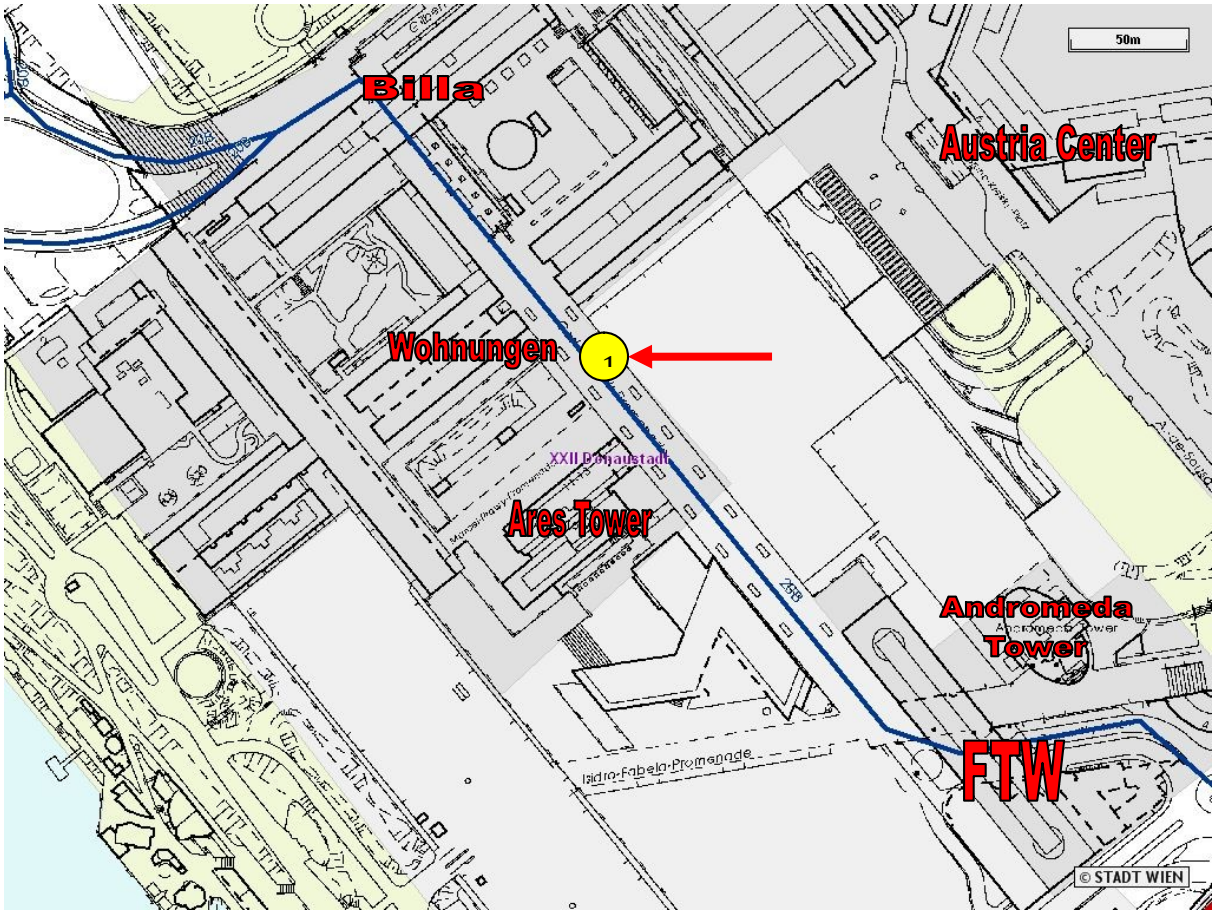
Unterschrift: _____

Datum: _____

Demografische Fragen

- Alter
- Beruf
    - Geschäftsinhaber
    - Selbständig / Freiberuflich
    - Angestellter mit leitender Funktion
    - Beamter mit leitender Funktion
    - Angestellter ohne leitende Funktion
    - Beamter ohne leitende Funktion
    - Arbeiter
    - Landwirt
    - Arbeitslos
    - Hausfrau / -mann
    - Student, Schüler
    - Pensionist
    - Sonstiges:
- Höchste abgeschlossene Schulbildung
    - Volksschule
    - Hauptschule
    - Abgeschlossene Lehre
    - Abgeschlossene weiterführende Schulbildung ohne Matura
    - AHS – Matura
    - BHS – Matura
    - Universitätsabschluss
    - Sonstiges:
- Computer- und Internetzugang: ja nein
- Internet-Nutzung: beruflich / privat / beides
- Internet-Nutzung pro Woche (Angabe in Stunden)
- Nutzung Handys
    - Wie viele SMS ca. pro Tag?
    - Wie lange telefonieren ca. pro Tag?
    - Haben Sie schon einmal mit dem Handy im Internet gesurft?
    - Benutzen Sie ihr Handy noch für andere Sachen (z.B.: Navigation, Organizer)

Ort des Geschehens:



Block 1:

Beim Weg zum Standort erklären:
Vielen Dank für die Teilnahme an unserer Benutzerstudie. Das ftw. (Forschungszentrum Telekommunikation Wien) befasst sich mit allen relevanten Bereichen der Telekommunikation, unter anderem auch Usability.
Hierzu wurden neue Systeme entwickelt und Sie dürfen diese nun ausprobieren. Es geht dabei um eine neue Ein- und Ausgabemöglichkeit. Und zwar einer sprachlichen Steuerung.

ASR (Automatic Speech Recognition) erklären.
TTS (Text to Speech) erklären.
LiliPUT erklären, falls vorhanden.

Bei der Parkbank:
Der Sensor muss vom Testleiter oder vom Beobachter immer mit der Blickrichtung der Testperson mitgedreht werden.
Erklärung: Beim ersten Programm wird nicht viel erklärt. Es gibt eine Einführung, wo sich das Programm selbst erklärt.
Headset aufsetzen. Programm starten, Sensor auswählen.
Einführung startet automatisch.
Achtung, beim Sensor auswählen muss der Sensor selbst abgeschaltet, Bluetooth aber eingeschaltet sein.

Programm push2speak!

Sie werden im Folgenden einige Aufgaben ausführen und verschiedene Anwendungen ausprobieren. Im Nachhinein werden Sie dazu befragt.

1. Einführung durchmachen: Hier wird dem Benutzer vom Programm erklärt, wie es zu bedienen ist. Und er probiert es gleich einmal aus. Er muss dabei auf den männlichen Sprecher umschalten.

2. Sie sind neu in dieser Umgebung und wollen sich Informationen über für Sie interessante Gebäude holen. Zeigen Sie mit dem Pocket PC auf Gebäude in der Umgebung und schauen Sie, ob und welche Informationen es dazu gibt.

3. Sie haben das Austria-Center gesehen und wollen wissen, wann es gebaut wurde. Versuchen sie, geschichtliche Informationen darüber einzuholen.

4. *Headset ausstecken.* Sie wissen nun, wann das Austria-Center gebaut wurde. Jetzt interessiert es Sie, ob das Austria-Center als erstes Gebäude in der Donaucity gebaut wurde. Versuchen Sie herauszufinden, wann die anderen errichtet wurden, zu denen das Programm Informationen bereitstellt.

5. Es besteht auch die Möglichkeit, Ausgaben des Programms mittels Sprachbefehle zu unterbrechen. Holen Sie sich eine Information und versuchen Sie das doch einmal!

6. Die Anwendung gibt zurzeit visuell und sprachlich dieselben Informationen aus. Suchen Sie den Ares Tower und informieren Sie sich, welche Veranstaltungen es gibt.

7. Versuchen Sie nochmals, sich beide Ausgaben zu merken. Dieses Mal bei den Informationen: Interessantes!

8. Beenden Sie das Programm nun bitte mittels einer Spracheingabe.

Fragen zu Block 1:
Usability:
- Ist das Programm einfach zu bedienen? Notenskala 1-5.

Notenskala	1	2	3	4	5
	<>	<>	<>	<>	<>

- Wenn nicht, was könnte man besser machen?

- Wie gut ist die Programmeinführung? Notenskala 1-5.

Notenskala	1	2	3	4	5
	<>	<>	<>	<>	<>

- Was fehlt, was ist unverständlich?

- Bei unterschiedlichem visuellen und auditiven Output?
Ist das sinnvoll oder sind das zu viele Informationen?

Erkennung:
- Welche Erkennung hat besser funktioniert? Mit Headset, ohne Headset?

- Wenn beide gleich gut funktionieren würden, welche würden Sie bevorzugen?

Notenskala	mit	eher mit	egal	eher ohne	ohne
	<>	<>	<>	<>	<>

- Warum?

Bargein:

- Gefällt Ihnen die Möglichkeit der Unterbrechung der Ausgabe? Ist das sinnvoll?

Notenskala	1	2	3	4	5
	<>	<>	<>	<>	<>

- Welche Befehle würden Sie hier verwenden?

Soziales:

- Würden Sie sich belästigt fühlen, wenn jemand neben Ihnen das Programm mittels Headset bedient?

Notenskala	ja	eher ja	egal	eher nein	nein
	<>	<>	<>	<>	<>

- Würden Sie sich belästigt fühlen, wenn jemand neben Ihnen das Programm ohne Headset bedient?

Notenskala	ja	eher ja	egal	eher nein	nein
	<>	<>	<>	<>	<>

Block 2:

Programm Dauerlauscher!

Erklärung: Ein anderes Programm, multimodales Interface.
Dieselbe Programmstuktur. Nur, um diesmal Befehle einzugeben, muss man vorher „Eingabe“ sagen. (word-spotting). Nicht mehr den Knopf betätigen! Dann nach dem Piepton sprechen.
Man kann auch per Pen durch das Programm steuern.
Headset wieder anstecken.

9. Pen-Interaktion: Sie sind auf der Suche nach einer Wohnung. Schauen Sie, ob es in der Nähe Informationen diesbezüglich gibt.

10. *Headset aufsetzen, Audio einschalten*
    Die neue Eingabemöglichkeit ausprobieren: auf weibliche Sprecherin umstellen!

11. Sie haben einen Bekannten, der im Andromeda Tower arbeitet. Wenn Sie ihn das nächste Mal treffen, wollen Sie ihn mit Informationen über den Andromeda Tower überraschen. Versuchen Sie, den Andromeda Tower zu finden und mittels Spracheingabe Informationen darüber einzuholen.

12. *ohne Headset:* Versuchen Sie, an die Informationen, die sie soeben gehört haben, nochmals, dieses Mal ohne Headset, heranzukommen.

Fragen zu Block 2:
Sie haben jetzt 3 verschiedene Interakionsmöglichkeiten kennen gelernt. Bringen Sie die in die Reihenfolge, wie sie Ihnen am besten gefallen:
Zur Erinnerung:
Pen-Interaktion
Sprache mittels „Codewort“
Sprache mittels Tastendruck.

Sie haben sowohl eine männliche Stimme, als auch eine weibliche Stimme gehört. Beurteilen Sie die Stimmen nach der Notenskala!

Notenskala	1	2	3	4	5
Männliche Stimme	<>	<>	<>	<>	<>
Weibliche Stimme	<>	<>	<>	<>	<>

Welche gefällt ihnen besser?
Können Sie der Person Eigenschaften bescheinigen, wenn Sie nur die Stimme gehört haben?

Wenn ja, welche?
Männliche Stimme:
Weibliche Stimme:

Finden Sie die Befehle, mit denen Sie diese Programme steuern konnten, für geeignet?

Notenskala	ja	eher ja	egal	eher nein	nein
	<>	<>	<>	<>	<>

Wenn nein, welche hätten Sie verwendet?
Würden Sie eher in ganzen Sätzen mit dem Pocket PC kommunizieren?
Das zweite Programm lauscht ständig auf Benutzereingaben. Stört das? Fühlen Sie sich da überwacht?

Notenskala	ja	eher ja	egal	eher nein	nein
	<>	<>	<>	<>	<>

Würden Sie das Programm bei so einer Erkennungsrate verwenden?

Notenskala	ja	eher ja	egal	eher nein	nein
	<>	<>	<>	<>	<>

Finden Sie solch ein Programm generell für sinnvoll?

Notenskala	ja	eher ja	egal	eher nein	nein
	<>	<>	<>	<>	<>

Halten Sie die Sprachdialoge mit Handys/Computer für sinnvoll?

Notenskala	ja	eher ja	egal	eher nein	nein
	<>	<>	<>	<>	<>

Anhang:

Bild herzeigen und Fragen dazu stellen.

Erklärung von vier anderen Arten, Informationen von der Umgebung einzuholen:

a) Liste (wurde ausprobiert)
b) Radar (Bild zeigen und erklären)
c) Karte (Bild zeigen und erklären)
d) Erweiterte Realität (Bild zeigen und erklären)
e) Panoramardarstellung (Bild zeigen und erklären)

Fragen:
Bei welchen von diesen Anwendungen könnten Sie sich eine Sprachsteuerung vorstellen?

Notenskala	1	2	3	4	5
a) Liste	<>	<>	<>	<>	<>
b) Radar:	<>	<>	<>	<>	<>
c) Karte:	<>	<>	<>	<>	<>
d) Erweiterte Realität:	<>	<>	<>	<>	<>
e) Panoramadarstellung:	<>	<>	<>	<>	<>

Erklärungen: Warum finden Sie sie bei a) sinnvoll/nicht sinnvoll
Warum finden Sie sie bei b) sinnvoll/nicht sinnvoll
Warum finden Sie sie bei c) sinnvoll/nicht sinnvoll
Warum finden Sie sie bei d) sinnvoll/nicht sinnvoll
Warum finden Sie sie bei e) sinnvoll/nicht sinnvoll

Ideensammlung: Wie kann das funktionieren? Wie kann das Interface aussehen?