Integratives Konzept zur prototypischen Implementierung multimodaler Benutzerschnittstellen
Integrative rapid-prototyping for multimodal user interfaces

Dipl.-Ing. Björn W. Schuller
Prof. Dr. rer. nat. Manfred K. Lang

Lehrstuhl für Mensch-Maschine-Kommunikation
Technische Universität München
(schuller | lang)@ei.tum.de

Zusammenfassung

Das Ziel der vorgestellten Arbeit ist es, ein domänen- und anwendungs-unabhängiges universelles Interface-Konzept für die Schnittstelle zwischen Mensch- und Maschine zu schaffen. Ein solches Standard-Interface kann für prototypische Evaluationen oder sogar finale Implementierungen durch geringfügige Modifikationen in neue Einsatzbereiche übertragen werden. Die gemeinsame Basis soll dabei maximiert werden um den Portierungsaufwand auf die neuen Anforderungen minimal zu halten. Vorgesehen sind zwei-dimensionale Visualisierungen für Informationsausgaben und zur Darstellung möglicher Benutzeraktionen sowie Sprachausgaben. Als Eingabemöglichkeit sind neben natürlicher Sprache haptische Interaktionen via berührsensitiver Gestik und Selektion sowie konventionelleren Einheiten integriert. Eine übergreifende Schnittstellendefinition erlaubt die Integration weiterer Modalitäten. Besonders hervorzuheben ist die Fähigkeit konzipierter Schnittstellen benutzer- und applikationsspezifisches Wissen, auch im zeitlichen Zusammenhang, zu akkumulieren. Über eine weitere Schnittstelle können zusätzlich externe Einflüsse modelliert werden. Durch schritthaltend gebildete Wissensbasen können dynamsiche Erwartungshaltungen an die folgenden Benutzeraktionen für eine prädikative Beschränkung des Hypothesenraums angebundener stochastischer Erkennerinstanzen aufgebaut werden. Die Schwierigkeit dieser Aufgabe liegt in der a-priori unbekannten potenziellen Aufgabe des Systems. Es sind also Zustände einer nicht bekannten Applikation automatisch zu erfassen. Ermöglicht wird das gesamte Konzept im wesentlichen durch die Einführung einer Mark-up Sprache und globale Kommunikationsstrukturen. Der verfolgte Ansatz erwies sich als leistungsfähig unter den gegebenen Anforderungen. Generierte Interfaces zeigten darüber hinaus in Versuchen hohe Akzeptanz bei den Benutzern.

Abstract

The aim of the presented work is to build a generic human computer interface construction kit for potential future applications. The concept permits rapid-prototyping or even final implementations with either a basic set of interaction concepts or adapted solutions. Multimodal control is an integrative factor. Basic input modalities are natural speech and haptic control via touch gestures or conventional devices. Output is basically given on a two-dimensional display plus speech. Future modalities can be connected by a defined communication protocol. An interface description language called IDL helps to build an interface-, an intention and an output model for an interface. A special feature is the ability to profile the user and integrate contextual system and external knowledge online. The achieved data is used to constrain the hypotheses sphere of the recognition instances and provides more robust recognition results. First interfaces have been produced at our institute and proven robust. The high acceptance among first test persons claimed for further research in this area.

Quelle: VDI-Bericht 1678 "USEWARE 2002, VDI-Verlag, Düsseldorf

1 Einführung

Um Bedienkonzepte in einer frühen Phase testen zu können, ist es hilfreich ein Interface prototypisch mit geringem Aufwand implementieren zu können. Einzelne Funktionalitäten, ins Besondere auch Modalitäten, können dabei unter Umständen von einem menschlichen Versuchsleiter, einem sogenannten Wizard, simuliert werden (siehe auch /1/). Somit erscheint es naheliegend eine Entwicklungsumgebung zu schaffen die diesen Aufwand minimiert, und gleichzeitig möglichst allgemein einsetzbar ist. Im Idealfall kann ein solches Konzept sogar die Implementierung und Portierung finaler Schnittstellen-Lösungen leisten. Diese Arbeit stellt einen Ansatz hierzu vor und diskutiert neben der Umsetzung erzielte Ergebnisse. Angesichts des aktuellen Trends hin zu multimodaler Bedienung soll das Konzept von Grund auf multimodal gelöst sein und die Integration neuer Modalitäten vorsehen. Als Basis werden die Steuerung per natürlicher Sprache und Berührinteraktion und -Gestik präsentiert. Abbildung 1 veranschaulicht das Konzept an Hand des Beispiels eines Interfaces für eine mobile Kommunikationseinheit und zeigt Bedienungsmöglichkeiten. Zur Sprachbedienung ist neben dem grundsätzlich vorhandenen Basiskonzept eine angepasstere Lösung exemplarisch angegeben.

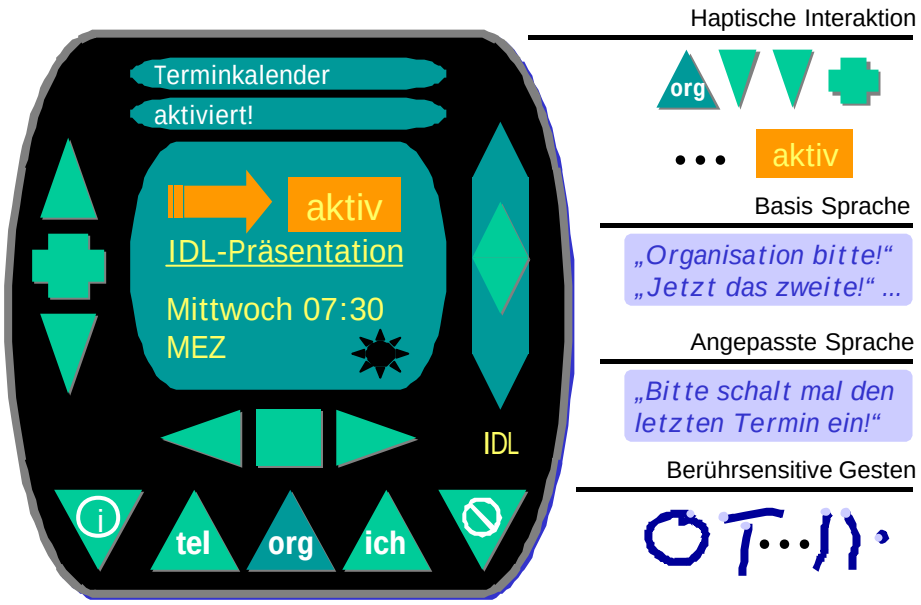

Abbildung 1: Beispielszenario mobile Kommunikationseinheit

Die Ausgabe konzentriert sich auf visuelle zweidimensionale sowie akustische Informationswiedergaben. Dabei sind sowohl bei der Ein- als auch bei der Ausgabe Standards vorgesehen, die eine schnelle generelle Lösung eines Interfaces erlauben. Durch ein weiterführendes Konzept ist jedoch auch eine angepasste Lösung unter hoher Bediengüte schnell erreicht. Mögliche Einsatzgebiete sind unter anderen mobile oder ortsfeste Infotainmentapplikationen auf Palmtops, Bordcomputer oder Desktops im Officebereich.

2 Realisierung

2.1 Skriptsprache für die Interfacespezifikation

Für eine generische Beschreibung der vorhandenen Benutzeraktionen eines beliebigen Interfaces führen wir eine Beschreibungssprache IDL („*Interface Description Language*“) ein. In ihr vorgesehen sind Verknüpfungen von Elementen zu Aktionen, Ausgaben und Statusmeldungen sowie Bindungen von dynamischen Inhalten. Hierarchische Menüebenen in Baum-Strukturen und mächtigere Shortcuts können implementiert werden. Für prototypische Evaluierungen im Rahmen von Wizard-of-Oz Benutzerstudien sind zusätzlich gedächtnisbehaftete Simulationsroutinen für zeitliche Verhältnisse vorgesehen.

2.2 Globale Kommunikationsstruktur

Eine zentrale Interpretereinheit kommuniziert über eine standardisierte formale Grammatik mit den multimodalen Erkenner-Instanzen und Ausgabeeinheiten über TCP/IP-Sockets. Im Falle eines realen Einsatzes kann auch die eigentliche Applikation angebunden werden. Dieses Vorgehen erlaubt eine Distribution über mehrere Rechner divergenter Plattformen. Die Erkennerinstanzen gehorchen einer übergreifenden Definition zur Bereitstellung von Hypothesen und zugehörigen Konfidenz-Bewertungen. Dabei werden jeweils n-beste Alternativen mit intentionsbasierten Wahrscheinlichkeiten über die geschätzte Absicht sowie deren Parameter im Detail bereitgestellt, was die Einschätzung der Sicherheit einer sinnvollen Aktion auf höchster Ebene ermöglicht. An dieser Stelle wird auch die später beschriebene Integration kontextuellen Wissens vollzogen. Als weiterer Vorteil können einzelne Modalitäten in einer späten semantischen Fusion unter Berücksichtigung weicher Entscheidungen optimaler zu einer Aktion ergänzt werden. Abbildung 2 gibt eine Übersicht über die Systemarchitektur sowie die gebildeten Wissensbasen von System und Entwickler.

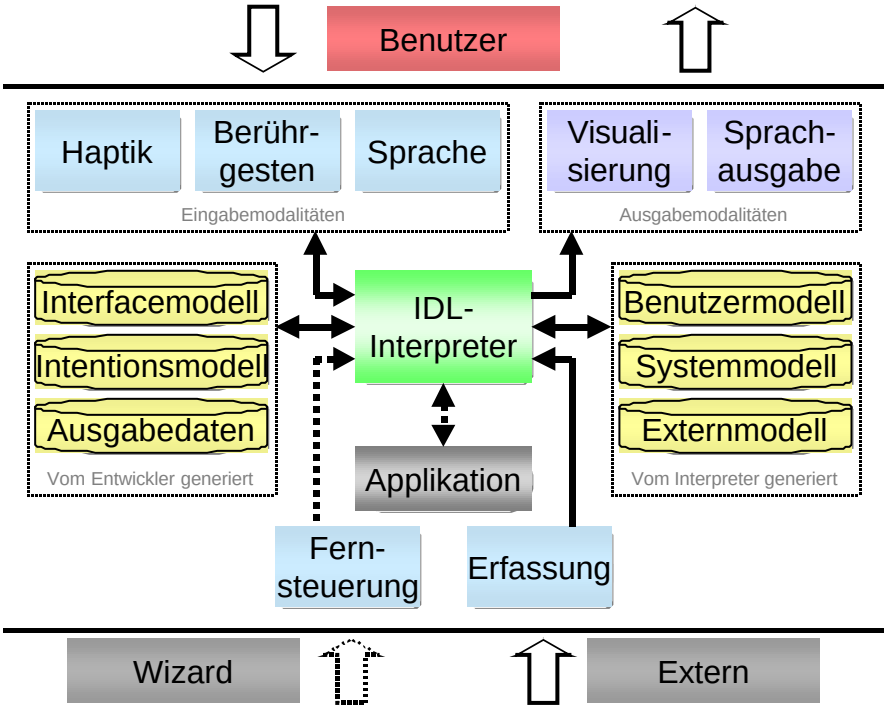

Abbildung 2: Übersicht über die System-Architektur

2.3 Basis des Bedienkonzepts

Bewährte Methoden der haptischen und natürlichsprachlichen Bedienung sowie generelle Style-Guides werden als gemeinsamer Ausgangspunkt bereitgestellt. Benutzerintentionen werden durch als Verknüpfung dargestellte Objekte erreicht. Diese können im Fließtext oder als Piktogramm in eine Displayseite eingebunden werden. Die Aktionsseiten sind ebenfalls lose verbunden. Durch dieses Prinzip lassen sich Menüebenen in Baumstruktur und Abkürzungen realisieren. In den Seiten kann mittels Verknüpfungswahl, sowie vorwärts- und rückwärts navigiert werden. Dabei können beliebig viele zunächst voneinander getrennte Hauptzweige realisiert werden, in die durch einen Direktsprung auf oberste oder letztgewählte Ebene gewechselt werden kann. Dargestellt werden diese in einem text- und grafikfähigem Display. Optional können zusätzliche Statuszeilen integriert werden. Über diese kann zum Beispiel die aktuelle Seite oder Systemaktion angezeigt werden. Sie eignen sich ins Besondere für die Präsentation dynamischer Inhalte. Durch einen online IDL-Codegenerator können aber auch im Hauptdisplay veränderliche Inhalte generiert werden. Als weitere Basisfunktionen stehen eine Hilfe-, eine Undo- und eine Stopp-Funktionalität jederzeit bereit, und erweisen sich gerade beim Einsatz unsicherer Erkennertechnologien als unerlässlich. Hierzu ist die Definition von adäquaten Hilfen und gegenläufigen Aktionen erforderlich. Eine durchgängig invariante Hilfefunktion ist ebenfalls vorgesehen.

2.4 Konventionelle haptische Bedienung

Als sichere Modalität wurde zunächst eine Bedienung über konventionelle haptische Einheiten integriert. Auf diese kann auch im Fehlerfall zurückgegriffen werden. Wie auch die weiteren Erkennerinstanzen sind sie in einer eigenen Instanz repräsentiert, die ihre Ergebnisse an die zentrale Einheit weiterleitet. Hier werden sie unter dem Kontext weiterer Modalitäten in Bezug zueinander gesetzt.

2.5 Sprachverstehen

Sprache ist die natürlichste Kommunikationsform zwischen Menschen. Neben ihrer Intuitivität erweist sie sich als sehr schnelle Interaktionsform und erlaubt berührungslose und blick- und ortsunabhängige Kontrolle. Vorraussetzung für einen geringen Lernprozess und eine minimale kognitive Last ist die Möglichkeit einer natürlich- oder spontansprachlichen Ausformulierung von Benutzerintentionen. Die Schwierigkeit birgt sich dabei in der unbekannten zukünftigen Applikation. Für Spracheingabe und Systemdialoge steht ein Basisbedienkonzept bereit. Dieses erlaubt die Navigation durch Menüebenen und die generelle Steuerung in Anlehnung an die haptische Bedienung wie Objektauswahl, -Aktivierung, -Speicherung, o.ä. Hinzu kommen global festgelegte Typen für Dialogabläufe, emotionale Benutzeräußerungen und Schlüsselphrasenaktivierung. Dies gewährleistet die Bedienung des vollen Funktionsumfangs auch per Sprache, was auf Grund einer individuellen Modalitätenselektion im Vordergrund der Modellierung steht. Das Basisbedienkonzept sieht sich im realen Einsatz jedoch in der Regel trotzdem mit domänenspezifischen, und ihm damit nicht bekanntem Vokabular und Äußerungen konfrontiert. Dies erfordert den Einsatz einer phrasengestützten Interpretation mit der Fähigkeit unsichere Elemente erkennen und überspringen zu können (siehe auch /2/). Um

aber das volle Potenzial dieser Kommunikationsform auszuschöpfen bedarf es eines aufgabenorientierten Sprachkonzepts. Hierfür kann der intentionsbasierte Decoder angepasst in seiner Modellierung erweitert werden. IDL bietet die Definition von potenziellen Benutzerintentionen in Anlehnung an die Aktionen, bestehend aus definierten Operator- und Operandenphrasen. Des Weiteren können sogenannte Superwörter festgelegt werden. Das Training kann dann in einem einfachen Schritt durchgeführt werden. Generell erlaubt die sprachverstehende Komponente eine Bedienung ohne manuelle Aktivierung einer Sprachinstanz mittels einer automatischen Interaktionserkennung. Für eine robustere Erkennung sorgt zusätzlich eine schritthaltende Sprecherverifikation, die sprachähnliche Geräusche filtert. Diese klassifiziert mittels eines kontinuierlichen Hidden-Markov-Modells (HMM) mit einem Zustand und Gaus-Mixturen (GMM) über das aus Kurzzeitspektren gemittelte Langzeitspektrum. Die höheren Frequenzanteile werden auf Grund ihrer Sprecherabhängigkeit kontinuierlich angehoben. Die sprachverstehende Komponente leistete in den betrachteten Versuchen über 93% Erkennungsleistung der gewünschten Benutzerintention.

2.6 Berührsensitive Interaktion

Bei vielen Anwendungen wie Palmtops oder öffentlichen Auskunftssystemen steht ein Touchpad oder -screen zur Verfügung. Über die Direktwahl hinaus bietet IDL die Möglichkeit Touchgesten einzusetzen. Szenarien sind hier unter anderen Buchstabieren oder der Einsatz von Shortcuts. Aber auch zusammengesetzte komplexe Anweisungsfolgen sind denkbar. Die Gestenerkennung ist mit Hilfe kontinuierlicher Bakis-HMMs mit GMMs realisiert. Die Zustandszahl der HMMs ist dabei variabel nach Geste optimiert. Merkmale sind die planaren Ortskoordinaten und ihre höheren Ableitungen. Dabei erfolgt die Extraktion schritthaltend, was auch einen eingeschränkten zeitlichen Bezug in die Betrachtung einbezieht. Die Merkmale erster Ordnung sind auf ein beschreibendes Rechteck normalisiert um eine Größeninvarianz und Ortsabhängigkeit sicherzustellen. Regelbasierte Plausibilitätsbeschränkungen optimieren weiterhin die Erkennung. Auf geeigneten Screens oder Pads wie mit akustischen Oberflächenwellen kann zusätzlich die z-Koordinate zur Regelung kontinuierlicher Größen eingesetzt werden. Diese Komponente kann auch zur Stützung der Segmentierung bei schwankender Andruckstärke mitberücksichtig werden. Eine semantische Schicht erlaubt die Interpretation von zusammengesetzten Gesten oder Gestenfolgen. So können Buchstaben in einem Zug, oder auch aus mehreren Einzelstrichen ausgeführt werden. Der Erkenner leistet bei 40 Gesten 98% Erkennung bei benutzerspezifischem Training und bei einer Versuchsreihe mit mehreren Anwendern im Schnitt 95% Erkennungsleistung.

2.7 Erwartungsbildung aus Benutzeraktionen

Ziel ist es die stochastischen Erkennermodule bei ihrer Entscheidung zu unterstützen. Durch das Vorausschätzen des Benutzerplans kann die Wahrscheinlichkeit einer Intention $I_k[t]$ des Benutzers B_i zu einem diskreten Zeitpunkt t als die bedingte Wahrscheinlichkeit $P(I_k[t]|B_i)$ verstanden werden. Die Modellierung temporaler Korrelationen erfolgt in erster Ordnung durch den Zusammenhang $P(I_k[t]|I[t-1],B_i)$. Durch die Erfassung dieser Benutzerpräferenzen können dem Benutzer gleichzeitig automatisch generierte Abkürzungen zu seinen

meistausgeführten Aktionen angeboten werden. Bei der dynamischen Profilierung des Benutzers hinsichtlich seiner Aktionen und Übergänge dieser ist bei der Erkennung unsicherer Modalitäten eine Überwachungsgröße sinnvoll. Dies kann ein gestuftes Vertrauen hinsichtlich einer Durchführung mit oder ohne Training sein, oder aus dem nachfolgenden Benutzerverhalten geschätzt werden.

2.8 Erwartungsbildung aus Zustandsübergängen des Systems

Eine sinnvolle Erwartungshaltung soll im aktuellen Systemzustand unsinnige oder nicht erlaubte Aktionen als Benutzeraktionen mit geringerer Auftrittswahrscheinlichkeit annehmen. Die Wahrscheinlichkeit einer Intention $I_k[t]$ wird somit analog durch den aktuellen Systemzustand $S[t]$ bedingt als $P(I_k[t]|S[t])$ gesehen. Eine Restwahrscheinlichkeit bleibt aber in jedem Fall einzuräumen. Tritt eine solch unerwartete Benutzeraktion erkannter Weise auf, kann in diesem Fall unter Umständen auf Hilfebedarf beim Benutzer oder einen Fehler in der Kommunikation geschlossen werden. Hierfür lernt das System selbstständig die auftretenden Applikationszustände kennen, und speichert in ihnen auftretende Aktionen und Übergänge zwischen diesen. In einer ersten Phase können entweder vom Entwickler einmalig möglichst alle vorhandenen Zustände vorgeführt werden. Alternativ lernt das System auch untrainiert im realen Einsatz.

2.9 Erwartungsbildung aus externen Einflüssen

Für die Adaption an umgebende Umstände ist ein Protokoll definiert. Das System kann diese so abbilden und lernen. Die Prädiktion einer Aktion I_k kann auf diese Weise durch die Berücksichtigung einer externen Einflussgröße X_l optimiert werden. Wenn wir die interpretierten Modalitäten als $M_1,..., M_N$ festlegen, lässt sich zusammenfassend die angenommene Intention $I_A[t]$ zum Zeitpunkt t wie folgt beschreiben:
$I_A[t]=arg\ max_k\ P(I_k|I[t-1],B_i,S[t],X_l,M_1,..,M_N).$

3 Anwendungen und Schlussfolgerungen

Das vorgestellte universelle Konzept wurde erfolgreich für Bedienkonzepte zu einer Infotainmentinstanz, einem multifunktionalen Audiogerät sowie einem Internet-Browser eingesetzt. Durch die Integration kontextuellen Wissens konnte eine über die Benutzer und Applikationen stark variierende mittlere Verbesserung von 6,3% der Erkennungsleistung erzielt werden. Im Rahmen der Evaluation zeigte sich als Vorteil, dass die gemeinsame Basis den Anwendern das Bedienen weiterer mit dem System erzeugter Interfaces erleichterte. Künftige Forschungen werden sich mit der weiteren Integration sozialer Kompetenz und langzeitlichen Versuchsreihen befassen.

4 Literatur

/1/ R. Nieschulz, B. Schuller, M. Geiger, R. Neuss: *"Aspects of Efficient Usability Engineering,"* IT+TI Vol. 44, Oldenburg, pp. 23-30, 2002

/2/ B. Schuller, F. Althoff, G. McGlaun: *"Navigation in virtual worlds via natural speech,"* HCII 2001, 9th Int. Conference on HCI, New Orleans, Louisiana, USA, Poster Session Abridged Proceedings, pp. 19-21, 2001