Mensch-Technik-Interaktion

Redselige Chips
Wenn Computer Gesten und Mimik zu deuten wissen, verstehen sie auch die Feinheiten der Sprache.

von Christine Ritschel

"Sie wünschen?" "Ich muss heute noch nach München, wann fährt der nächste Zug? Nein, warten Sie, können Sie gleich noch nachschauen, wie ich vom Bahnhof zur Universität komme?"

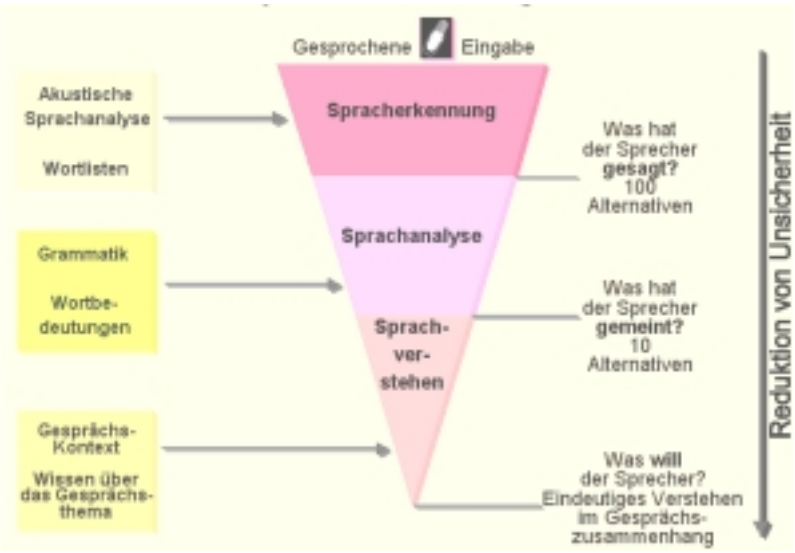
Ein solcher Dialog sollte dem Beamten der Bahnauskunft weiter keine Probleme machen. Er weiß, wo er sich gerade befindet, kann auf einer Uhr die Tageszeit ablesen, weiß die Fahrpläne zu lesen und vor allem - er versteht die Fragen. Doch in wenigen Jahren sollen computergesteuerte Kommunikations-Kioske auf Flughäfen und Bahnhöfen derartige Informationen ebenfalls geben können. Die erforderliche Technologie zu entwikkeln scheint den Unternehmen der Kommunikations- und Informationstechnik ein Gebot der Stunde. Microsoft-Gründer Bill Gates glaubt, eine gravierende Erweiterung des Computermarktes nur dann zu erreichen, wenn die nächste Generation von Anwendersoftware so entwickelt ist, dass selbst ein Computerlaie sie über eine auf ihn abgestimmte und intelligente Schnittstelle bedienen kann. Die geradezu universelle Einsetzbarkeit von Computerchips verleiht der Forderung Nachdruck, gibt es doch kein Handy, keinen Videorekorder und keine Waschmaschine ohne Prozessor- aber komplizierte Bedienungsanleitungen nerven den Konsumenten.

Wie eine "natürliche" Bedienung eines elektronischen Geräts aussehen könnte, erforscht Wolfgang Wahlster, Institutsleiter und Geschäftsführer des Deutschen Forschungsinstitutes für Künstliche Intelligenz (DFKI) in Saarbrücken. Im November des letzten Jahres erhielt er für das Projekt "Sprachverstehende Computer als Dialog- und Übersetzungsassistenten" (Verbmobil) den Deutschen Zukunftspreis – Preis des Bundespräsidenten für Technik und Innovation. Das Folgeprojekt "SmartKom" basiert auf dieser Entwicklung, soll aber auch Bildverarbeitung und Wissensdatenbanken miteinbeziehen, um die Spracherkennung zu verbessern sowie Assistenzfunktionen zu ermöglichen.

Verbmobil ist ein Sprachtechnologieprojekt im Bereich Maschinelle Übersetzung. Es erkennt Spontansprache mit allen ihren Ungereimtheiten und Versprechern, analysiert und übersetzt in eine andere Sprache, erzeugt einen Satz und spricht ihn aus. Gesteuert über einen zentralen Server, verarbeitet das System Spontansprache sprecherunabhängig und leistet Übersetzungshilfe in Dialogsituationen. Zudem wurden erste Prototypen für zukünftige Anwendungen erprobt, die auf transportablen Computern arbeiten (Spektrum der Wissenschaft, März 1994, S. 99). Nicht weniger als 30 deutsche Hochschulen, Forschungsinstitute und Unternehmen nahmen 1993 die Entwicklungsarbeit auf.

Bereits 1996 entstanden erste Produkte wie Diktiersysteme, Freisprecheinrichtungen und telefonische Informationssysteme, aber sie erwiesen sich als unausgereift. Vom amerikanischen Markt für Diktiersysteme wurde dazu bekannt, dass nach einem Jahr nur noch rund 10 Prozent der Käufer ihr Diktiergerät benutzten, weil trotz sprecherabhängigen Trainings die Fehlerrate noch bei 5 bis 10 Prozent lag. Je "freier" der Benutzer sprach, desto höher war die Fehlerrate.

Im nächsten Schritt bezogen die Wissenschaftler deshalb auch den Kontext eines Satzes in die Sprachverarbeitung mit ein, ebenso dialektische Sprachfärbungen, Betonungen sowie die Satzmelodie (Prosodie). Zum Beispiel erhält die Äusserung "Wir treffen uns vor der Tagung" – ist dies nun zeitlich oder örtlich gemeint? - nur im Kontext ihren Sinn. Oder problematische, dialektale Färbungen der saarländischen, pfälzischen Äusserung "Ich finde das nätt" sind erst durch Einbeziehung des Kontextes in "nett" oder "nicht" unterscheidbar. Nicht anderes erschließt sich auch der Mensch den Inhalt von Gesagtem
(siehe nebenstehende Grafik)



Doch das reichte immer noch nicht aus: Um einen Sachverhalt fehlerfrei zu erkennen und zu übersetzen, ist oft auch "Wissen" um den Gesprächsgegenstand erforderlich. Beispielsweise lässt eine Verabredung "zum Essen" im Deutschen die Tageszeit offen, eine Übersetzung ins Englische müsste aber zwischen "lunch" (Mittagessen) und "dinner" (Abendessen) unterscheiden. Dazu müssen ganze Wortwendungen, Satzbruchstücke und Idiome mit "Wissen" verknüpft in Datenbanken hinterlegt werden. Weil die große Datenmenge nicht anders zu bewältigen wäre, unterscheidet Verbmobil die Bereiche Reiseplanung, Hotel- und Gaststättenreservierung, Konferenzen und Terminplanung und greift für Deutsch, Englisch oder Japanisch auf jeweils andere Wissensquellen zurück.

Auf dieser Technologie basieren mittlerweile einige Telefonauskunftssysteme, ohne dass wir sie sonderlich wahrnehmen. Dazu gehört ein Börsentelefon von Sympalog (0190-590400), die Bahnauskunftssysteme TABA (0241-60 40 20) von Philips und OSCAR von DaimlerChrysler (0180-5 99 66 22), sowie ALF von DaimlerChrysler (0180-3 00 00 74), das über Abflug- und Ankunftszeiten der Lufthansa informiert. Der Automobilkonzern bietet außerdem als Sonderausstattung in seinen Modellen mit "Linguatronic" eine sprachgesteuerte Bedienung für Telefon und Klimaanlage an. Weitere sprachgesteuerte Bedienfunktionen nicht sicherheitsrelevanter Komponenten wie Navigationssystem, Radio, CD-Player oder Fensterheber kommen noch in diesem Jahr auf den Markt. Inzwischen liegt die Dialogerfolgsrate bei kommerziellen Systemen bei zirka 95 Prozent.

"Die drei Stufen der Sprachverarbeitung"

Das menschliche Gehirn verarbeitet Sprache in drei Stufen, wobei in jeder Stufe der Sprachverarbeitung neue Wissensquellen hinzu kommen, um die Unsicherheit, was der Sprecher eigentlich will, einzuschränken. Erfasst die erste Stufe nur akustisch über Wortlisten das Gesagte, kommen in der zweiten sowohl Grammatik als auch Wortbedeutungen hinzu – die Sprache wird analysiert. Eindeutiges Verstehen des Gesagten setzt aber voraus, die Äußerung im Kontext zu betrachten und benutzt deshalb weiteres Wissen zum Gesprächsgegenstand, über das Thema. So wie es der Mensch auch tut.

SmartKom-Mobil:
Der ständige Begleiter

Wer immer informiert sein möchte, wie die Börse steht oder wo er sich gerade befindet, sich in einer fremden Stadt bewegen will und ähnliche Serviceleistungen wie am "Infokiosk" in Anspruch nehmen will, braucht eine mobile Plattform, die beispielsweise den Zugang zum Internet über eine GSM-Handy-Verbindung herstellt, und mittels GPS die Eigenbewegung auf digitalen Karten verfolgt. Mikrophon, Kamera und Stift dienen der Befehlseingabe. Die Verarbeitung obliegt einem Zentralrechner, der über eine Funkverbindung mit dem Ein/Ausgabegerät kommuniziert. Derzeit in Erprobung ist ein multimodaler mobiler Touristenführer für Heidelberg und ein von DaimlerChrysler entwikkeltes mobiles Navigationssystem, dass per Dockung-Station sowohl im Auto als auch als mitnehmbares Kommunikationsgerät fungiert.

Doch bislang setzt Verbmobil immer noch voraus, dass das zu verarbeitende Sprachsignal ungestört ankommt. Was aber, wenn der Reisende seine Anfrage nach einem Zug in Richtung München undeutlich nuschelt? Und wie soll der Computer erkennen, dass ein "Wenn es geht, noch heute" ironisch gemeint ist, selbst wenn er mit Ironie und Sarkasmus umzugehen wüsste?

Menschen lösen dieses Problem, indem sie weitere Informationsquellen in die Sprachverarbeitung einbeziehen, und genau das soll das Folgeprojekt SmartKom auch: das Deuten, das Verstehen von Gestik und Mimik soll die Fehlerrate weiter senken. Auch das Wissen um den Gesprächsgegenstand soll ausgebaut werden. Wieder unter Federführung von Professor Wahlster erarbeiten Wissenschaftler Kernfunktionen für eine intelligente Kommunikationsplattform, die einen natürlichen anmutenden Dialog, eine intuitiv ablaufende Interaktion von Mensch und Technik ermöglicht.

Dazu erscheint es sinnvoll, die Benutzerschnittstelle in Form eines Kommunikationsagenten, "Smartakus" getauft, zu personifizieren oder einfacher gesagt: Dem Benutzer ein Gegenüber vorzuspielen. Informatiker Software-Programme bezeichnen als Agenten, die selbständig innerhalb eines Systems agieren können und auch über eine rudimentäre Intelligenz verfügen. Als Schnittstelle zur Maschine initiiert er dann die gewünschten Aktionen, ob das die Suche nach einer Zugverbindung ist, das Starten einer Klimaanlage oder die Programmierung des Videorekorders. Der hier forcierte Assistent versteht, was man ihm sagt, kann sogar fehlerhafte oder unvollständige Eingaben sinnvoll interpretieren, oder gegebenenfalls nachfragen, um so die Absichten des Nutzers zu erschließen.

Ein solcher Dialog wäre beispielsweise folgender:

„Smartakus" - der Kommunikationsssistent

Videobasierte Mimikerkennung auf der Basis von Eigenfaces

Sprecherunabhängige Emotionserkennung

(1) Smartakus: "Womit kann ich Ihnen dienen?"
(2) Benutzer: "Ich möchte morgen Mittag nach Frankfurt fahren."
(3) Smartakus: "Sie möchten also zwischen 11.00 und 13.00 von Saarbrücken-Hauptbahnhof abfahren?"
(4) Benutzer: "Ja, richtig."
(5) Smartakus "Es bestehen folgende Verbindungen: Mit dem Intercity Abfahrt ....."

Das System erkennt und interpretiert die Eingabe "Mittag" als Zeit zwischen 11.00 und 13.00 Uhr oder in Kenntnis des Standortes Saarbrücken des Benutzers sowie dass keine spezielle Bahnhofsangabe gemacht wurde, dass der Hauptbahnhof in Saarbrücken gemeint ist. Mit seiner Gegenfrage versichert sich das System - Smartakus -, ob es richtig "verstanden" hat, ansonsten kann der Benutzer ja widersprechen bzw. korrigieren.

Solche Klärungsdialoge spontansprachlicher Beratungssysteme sollen in Zukunft sowohl von Benutzer und System komplexe Interaktionen zu lassen, wie Rück- und Klärungsfragen stellen, Verstehensprobleme signalisieren oder den Dialogpartner unterbrechen können.

Dazu ist der Kommunikationsassistent in der Lage, sich jedesmal neu auf seinen individuellen Benutzer, auf die Gesprächsdomäne einzustellen. Man spricht von multifunktionaler Sprachtechnologie.

Die zusätzlichen Informationsquellen Gestik und Mimik erfassen je eine Infrarot- und eine Videokamera nebst graphischer Bildverarbeitung. Analysiert der Rechner das akustische Signal synchron zu den optischen Daten, kann er Mehrdeutigkeiten einer sprachlichen Äußerung oftmals schon erheblich reduzieren. Umgekehrt vermag er aber auch eine mehrdeutige Geste oder einen zweideutiger Gesichtsausdruck anhand der gesprochenen Information zu interpretieren.

SmartKom-Home/Office - Steuerung der Haustechnik, Haushalts- und elektronischer HiFi-Geräte etc.)

"Öffne das Garagentor, wenn ich die Einfahrt passiere und schalte das Licht an. Sobald ich die Garage verlasse habe, schließe das Tor und schalte das Licht wieder aus, aktiviere die Sicherheitseinrichtungen." Über den Software-Agenten "Smartakus" soll die vernetzte Haustechnik verbal zu steuern sein. Sony und Philips entwickeln im Projekt SmartKom-Home/Office auch sprachgesteuerte Benutzerschnittstellen für Videorecorder, DVD-Player und Fernseher. Dann sollte folgendes möglich sein "Nimm am Montag alle Nachrichten zur Bundestagsdebatte "Mautgebühren" auf, egal auf welchen Sender darüber berichtet wird! ... Auf welchem Sender läuft jetzt Sport? ...Aha, Bayern München gibt's im Zweiten und Skispringen im Ersten... dann schalt auf Bayern München um und nimm mir das Skispringen auf, ich schau' es später an! ...etc."

SmartKom-Public: Informationsstände im öffentlichen Verkehrsraum "Unterhaltung am Info-Kiosk"

In naher Zukunft sollen Kommunikations-Kioske an Flughäfen und Bahnsteigen Auskünfte zu Bahn- und Flugplänen, zum Stadtplan, sowie zu Hotels, Theatern, Kinos oder Restaurants erteilen. Neben der Sprachaufnahme mittels Mikrofon, registrieren eine Video-Kamera die Gesichtsmimik und eine Infrarot-Kamera die Gesten seines Benutzers. Um alles so lebensnah wie möglich zu gestalten, wird die Antwort auf ein Display projiziert und zudem gesprochen. Industriepartner wie Siemens entwickeln vandalismussichere Info-Kioske. Nach Entrichtung einer Gebühr beispielsweise mit einer Chipkarte, ähnlich wie bei einer öffentlichen Telefonzelle, kann der Mensch sie dann benutzen. Wie das im Einzelnen konkret aussehen wird, liegt noch in den Händen der Entwickler und Marketingstrategen.

Sprache, Gestik und Mimik sowie _ auf Seiten des Computers _ Piktogramme sind eigenständige Zeichensysteme. Sie sprechen zudem bestimmte Sinne (visuelle, akustische und taktile) des Menschen an; die Wissenschaftler sprechen von "Modalitäten". Dementsprechend enthält das SmartKom-Konzept als erste der Hauptkomponenten modalitätsspezifische Analysatoren für Sprache, Gestik, Mimik und Biometrie. Letzteres ist nur insofern Teil der Kommunikation, als durch Erkennung von Stimme, Handkontur oder Unterschrift bei nicht-öffentlichen Systemen eine Zugangskontrolle eingebaut werden soll. Schließlich dürfte es niemandem recht sein, wenn sein "SmartKom-Home/Office" in fremde Hände fällt und nun Unbekannte die Alarmanlage des trauten Heims ausschalten, die Türen öffnen oder aus Schabernack die im Hausnetz integrierten Geräte einschalten könnten. Die Zentraleinheit bildet die multimodale Interaktionskomponente, mit einer Schnittstelle zur jeweiligen Anwendung – das sind explizite Anwendungsmodelle gefüttert mit Informationen, Anwendungen und virtuellen Kommunikationspartnern - sowie das multimodale Mediendesign für die Ausgabeplanung.

Das Projekt “SmartKom” (Laufzeit vom 01.09.1999 – 30.09.2003) gliedert sich in mehrere Teilprojekte. In einem dieser Teilprojekte geht es um die Entwicklung modalitätsspezifischer Analysatoren, die dem Benutzer ermöglichen sollen, in der Kommunikation mit Maschinen und technischen Systemen unterschiedliche Eingabeformen benutzen zu können. Während in Vorgängerprojekten bereits für das Interaktionsmedium Sprache die Schnittstellen zwischen Analysatoren für die Spracherkennung und den interpretierenden und dialogverarbeitenden Komponenten erarbeitet worden sind, steht die Forschung bei Gestik und insbesondere bei Mimik und sprachlichen Emotionen noch am Anfang.

Die Gesamtaufgabe ist beliebig komplex: Die meist unbewußt eingesetzte Mimik oder ein emotionaler Ausdruck in der Sprache tragen ja im realen Dialog zwischen Menschen Information und verändern den Diskurs. Das gilt ebenso für real-manipulative Aktionen wie das physische Einbringen eines Dokumentes. Beispielsweise müssen Merkmale definiert werden, die es ermöglichen, Gestik, Mimik und sprachlichen Emotionen zu erkennen und zu interpretieren. Für’s erste suchen die Wissenschaftler der beteiligten Projektpartner die Anforderungen zu vereinfachen, indem sie sich auf die Entwicklung von Analysatoren für die einzelnen Eingabemodalitäten und eine nachfolgende Modalitätenfusion konzentrieren, ohne ihre Wechselwirkung untereinander zu berücksichtigen. Das folgt erst in einem nächstem Entwicklungsschritt .

„Smartakus“ - als Touristenführer

Zur Erprobung wurden in Zusammenarbeit mit zwölf Forschungszentren, Universitäten und Industrieunternehmen drei Anwendungsszenarien vereinbart, neben dem erwähnten Home/Office ein Informationsstand für Flughäfen und Bahnsteige (SmartKom-Public) und der mobile Kommunikationsassistent (SmartKom-Mobil). Alle Kommunikationssysteme sollen über einen Zugang zum Internet verfügen und in speziellen Fällen, wie bei SmartKom-Mobil soll zusätzlich die Eigenbewegung mittels GPS und Navigationssystem verfolgbar sein. Die Bedienung wird in allen Systemen weitgehend gleich sein - komfortabel und intuitiv im natürlichem Dialog.

Auch für die nächste Generation von Mobilfunkgeräten mit ihren großen Übertragungsbandbreiten sind Spracherkennungskonzepte besonders wichtig, erfordern sie doch ganz neue Anwendungen und Bedienkonzepte. Dazu stellt

klassische Mobiltelefonie

UMTS-Mobilfunkgerät der dritten Generation

Verbmobil (reine Sprache)

SmartKom (Sprache, Gestik, Mimik)

Vom Sprachdialog zum multimodalen Dialog in der Mobilfunktechnik

Professor Wahlster mit Liebling „AIBO“

Prof. Wahlster fest: “Die multimodale Kommunikation mit SmartKom ist für zukünftige UMTS-Anwendungen eine Schlüsseltechnologie.” So hat Sony aus Japan in ihrem europäischen Zentrum in Stuttgart eine Abteilung für Sprachtechnologie eröffnet und Ericsson aus Schweden hat Erlangen als Standort für ihre gesamten weltweiten Forschungs- und Entwicklungsaktivitäten in der Sprachtechnologie gewählt.

“In ferner Zukunft,” so fabuliert Wolfgang Wahlster, “wenn der Computer gelernt hat, Ironie, Sarkasmus, Zustimmung oder Ablehnung, Lob oder Tadel zu unterscheiden, wird er so menschliche Züge bekommen, dass wir uns mit ihm wie mit einem Menschen unterhalten können.” Dass das schon im Kleinen funktioniert, zeigt der kleine japanische sprachgesteuerte Roboterhund AIBO, entwickelt von Sony. ((hier bitte das AIBO-Bild ZAbb3 mit Wahlster einfügen oder?)) Er gehorcht bereits auf Kommando, spielt Fußball und reagiert, dank eines SmartKom-basierten Computerbausteins, auf Streicheleinheiten mit Schwanzwedeln und melodischklingenden Stimmungsäußerungen. Ob es eines Tages Fahrstuhltüren geben wird, die _ wie der englische Science-Fiction-Autor Douglas Adams ironisch schilderte _ dem Benutzer freudig für die Möglichkeit der Pflichterfüllung danken, sei dahin gestellt. Das Ziel von Wolfgang Wahlster aber werden viele unterstreichen: Technik auch ohne Expertenwissen bedienbar machen.