Thema:

Multimodale Interfaces

Verfasser: Nicolas Honegger

99-909-327, nicolashonegger@mac.com

Fabian Rutishauser

02-706-646, fabian.rutishauser@bluewin.ch

Betreuer: Peter Vorburger

Institut: Institut für Informatik der Universität Zürich

Professor: Prof. Dr. A. Bernstein

Fach: Informatik (Seminar Context Aware Computing)

Semester SS 2006

Einführung

Obwohl die Informationstechnologie eine der sich am schnellsten entwickelnden Wissenschaften ist, haben sich die Schnittstellen zwischen Mensch und Computer in den letzten Jahrzehnten faktisch kaum geändert. Solche Eingabearten sind fast nie an den Menschen, sondern meist an die Maschine angepasst. Die Tastatur als extremstes Beispiel ist sogar noch ein Überbleibsel aus der Zeit der Schreibmaschinen, in der gemäss [ADLER1997] Sholes & Glidden im Jahre 1874 die erste „QWERTY"-Tastatur herstellten. In den letzen Jahren geistern immer wieder neue Schnittstellen in der Öffentlichkeit umher, dazu gehören Sprachbefehle, Handschrift- und Gestenerkennung um nur einige zu nennen. Keine dieser teilweise viel versprechenden Technologien hat es bisher geschafft sich über eine punktuelle Benutzung zu erheben. Deshalb wollen wir hier einen der Schlagausdrücke in der Welt der Computer-Mensch Schnittstellen genauer betrachten: „Multimodale Schnittstellen".

Eine Schnittstelle wird als multimodal bezeichnet, wenn sie verschiedene Ein- oder Ausgabearten kombiniert. Am weitesten verbreitet ist momentan die Verbindung von visuellen und gesprochener Kommunikation, wobei aber beliebige verschiedene andere Möglichkeiten genauso denkbar sind. Multimodale Schnittstellen bieten durch die Parallelisierung der Ein-, respektive Ausgabemöglichkeiten nicht nur eine verbesserte Benutzerfreundlichkeit und Robustheit sonder auch eine Verbesserung der Genauigkeit von mehrdeutigen Technologien durch geschickte Kombination verschiedener Modalitäten. Auf diesen Umstand gehen wir in Kapitel 2.1 anhand eines konkreten Projektes genauer ein, in dem die Qualität der Spracheingabe mit Handschrifterkennung verbessert wird. Weitere interessante Anwendungsmöglichkeiten sind die Verbesserung der Computer-Mensch Interaktion für körperlich behinderte Leute, die beispielsweise Hör- oder Sehschädigung haben und dadurch bestimmte Kommunikationskanäle nicht gleich gut nutzen können, was von der Multimodalität aufgefangen werden kann.

Im Folgenden gehen wir auf zwei konkrete Beispiele näher ein, die Gegenstand aktueller Forschung sind. Das Ziel unserer Arbeit liegt darin den momentanen Stand der Wissenschaft auf dem Gebiet der Multimodalen Schnittstellen aufzuzeigen und damit einen realistischen Eindruck von Gegenwart und Zukunft des entsprechenden Forschungszweiges zu vermitteln.

1 Beispiele

1.1 Multimodale Worterkennung

1.1.1 Motivation

Sehr viele Computersysteme benötigen Worte als Eingabesteuerung – geschrieben oder gesprochen. Ein Problem liegt darin, dass Benutzer nicht immer die gleichen Worte einsetzen, um einen bestimmten Inhalt zu komuunizieren. Es ist wünschenswert ein System zu entwickeln, welches ihm nicht bekannte Worte möglichst schnell erkennt; am besten gleich bei der ersten Eingabe. Es ist sehr aufwändig und oft gar unmöglich dem Computer alle möglichen Varianten für eine Spracheingabe mitzugeben, wie das heutzutage üblicherweise gemacht wird. Der Computer soll für neue Befehle und neue Varianten von bereits bestehenden aufnahmefähig gemacht werden. Herkömmliche Systeme sind nicht in der Lage zeitgleich mit dem Input eines nicht bekannten Wortes dieses auch gleich zu erlernen. Die Daten müssen unabhängig davon manuell im System erfasst werden. [KAISER2005] hat ein System entwickelt um mit Hilfe einer multimodalen Schnittstelle die dynamische Worterkennung zu ermöglichen. Diese Schnittstelle nimmt sowohl die Stimme als auch die Handschrift auf um durch die richtige Kombination der beiden Inputs auf das vom Benutzer gewünschte Wort zu kommen, auch wenn das System dieses Wort vorher nicht kannte. Das Ziel des Projektes ist die Erweiterung des Wortschatzes des Computersystems so einfach zu gestalten wie das Lernen eines Menschen. Im vorgestellten Projekt von [KAISER2005] wird dabei auf das Erstellen eines gemeinsamen Wortschatzes von Mensch und Maschine fokussiert. Die Hoffnung liegt darin die Erkennung neuer Wörter durch die Kombination von Sprach- und Handschrifteingabe deutlich zu verbessern.

1.1.2 Lösungsansatz

Das System von Edward C. Kaiser ist so gebaut, dass es ein Zeitplanungs-Meeting unterstützen soll in dem Gantt-Charts benutzt werden. Der Leiter des Meetings arbeitet an einer Weisswandtafel mit Berührungssensoren und benutzt ein Mikrofon. Über die Berührungssensoren an der Tafel nimmt das System die Schrift auf, während durch das Mirkofon die gesprochenen Worte erfasst werden.

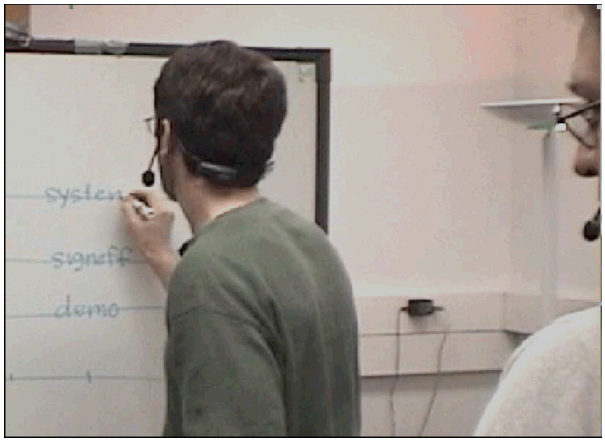

Abbildung 1: Kombination von Handschrift- und Spracherkennung in einem Meeting [KAISER2005]

Der auf der Weisswandtafel von Hand aufgezeichnete Chart wird vom System, wie in Abb. 1 zu sehen, elektronisch erfasst und angezeigt. Dies würde in einem realen Einsatz des Systems unter anderem Meetings ohne Anwesenheit aller Teilnehmer, wie zum Beispiel Videokonferenzen oder auch das Digitalisieren sowie Archivieren von Flip-Charts erleichtern.

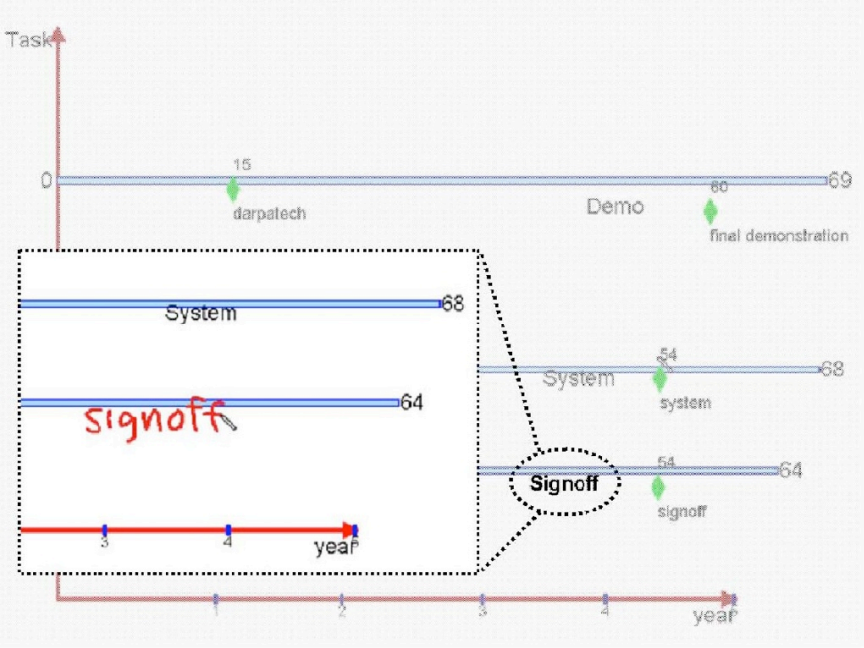

Abbildung 2: Gantt-Chart wie er vom System erfasst wird [KAISER2005]

Wenn ein Wort dem System unbekannt ist, und es gleichzeitig mit der Aussprache vom Meetingleiter auf die Tafel geschrieben wird, analysiert das System die beiden Eingabearten auf ihre Einzelteile um diese danach zu vergleichen. Sowohl die Sprache als auch die Schrift sind dabei aber nicht eindeutig erkennbar, was zu einer gewissen Unsicherheit führt.

Genau hier setzt der multimodale Ansatz an, indem er über die so genannte „mutual Disambiguation“ die erwähnte Unsicherheit verkleinern soll. „Mutual disambiguation“ bedeutet frei übersetzt etwa „gegenseitig eindeutig machen“. Das Konzept soll durch die Kombination von zwei unsicheren, korrelierten Inputs die Wahrscheinlichkeit zum richtigen Ergebnis zu kommen erhöhen. Zum Beispiel kann man, wenn ein gesprochenes Wort nicht richtig verstanden wurde, über ein anderes Medium einen Input bekommen, der die Unsicherheit klärt. Bei einer Konversation von Angesicht zu Angesicht liest der Mensch automatisch auch die Lippenbewegung und erkennt so auch Wörter, die über die Akustik nur undeutlich durchkommen.

Bei der Worterkennung von [KAISER2005] werden die Inputs über das Mikrofon (speech in Abbildung 3) und die Handschrift (handwriting) einzeln analysiert. Spracheingabe wird als

eine Abfolge von Tönen (SP in Abbildung 3, speech-phones) wahrgenommen, welche an Orthografien (SL, speech-letters) zugewiesen werden. Zeitgleich wird die Handschrift in Buchstaben aufgeteilt (HL, handwriting-letters), einzelnen Tönen zugewiesen (HP handwriting-phones). Es werden so genannte OPS-Tupel gebildet, die jeweils einen Wert für eine mögliche Orthografie, Phonetik und Semantik beinhalten. Um eine rudimentäre Semantikwahrnehmung zu ermöglichen wird im Experiment erwartet, dass Menschen in der verbalen Interaktion mit dem Computersystem eine ähnlich vereinfachte Satzstellung verwenden wie das bei der Kommunikation mit Kindern laut [GOGATEet al. 2001] der Fall ist. Orthografie ist das geschriebene Wort, Phonetik steht für die, in phonetischer Schrift dargestellte, Aussprache und die Semantik beinhaltet den grammatikalischen Kontext in dem das Wort vorkommt. Ein Beispiel ist das Wort „handoff“, welches den Wert „handoff“ und „hh ae n d ao f“ bekommt. Die Werte der Tupel entstehen aus den beiden Eingabearten der Sprache und Schrift. Jeder der beiden Inputs ist unsicher und generiert so verschiedene mögliche Werte, aus welchen durch Kombination die Tupel gebildet werden. Jedes der Tupel bekommt einen Wert für die Aussprache und die Orthografie und Semantik. Den einzelnen Werten werden Wahrscheinlichkeiten zugewiesen. Für jedes der Tupel wird eine Punktzahl berechnet, um das richtige ins System aufzunehmen. Die Punktzahl beinhaltet neben den Wahrscheinlichkeiten der Sprach- und Handschrifterkennung noch die so genannte „Edit Distance“ (ED) oder „Levenshtein Abstand“ was gemäss Definition von [BLACK1999] der kleinsten Anzahl an Manipulationen (löschen, einfügen, mutieren) entspricht die notwendig sind um zwei Strings (oder Bäume) gleich zu machen. In unserem Beispiel wird die phonetische Tonabfolge aus der Spracheingabe (SP) mit der aus der Handschrift gewonnenen (HP) verglichen und der Levenshtein Abstand ausgerechnet. Das gleiche geschieht mit den beiden Buchstabenabfolgen (HL und SL). Die Punktzahl eines Tupels ergibt sich durch eine simple Multiplikation der Sprach- und Handschrifterkennung mit der „Edit-distance“ (ED_L,

ED_p). Das Tupel mit der höchsten Punktzahl wird schlussendlich vom System in den Speicher aufgenommen.

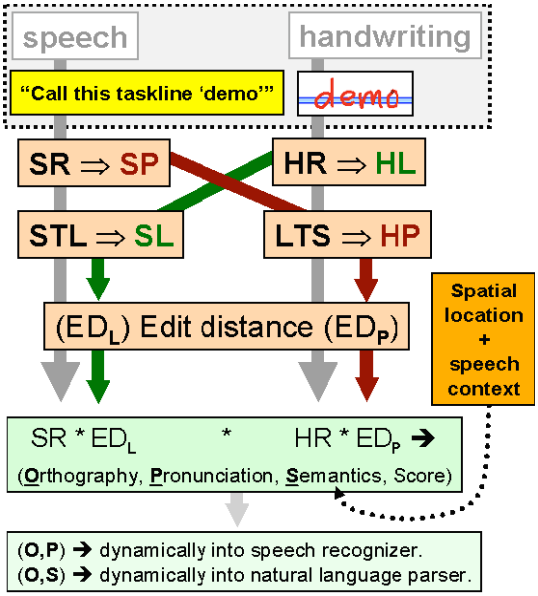

Abbildung 3: Multimodale Worterkennung mit Hilfe von Handschrift und Spracheingabe.

Dieser Algorithmus, der die Gesamtpunktzahl eines Tupels berechnet, ermittelt nicht immer die richtige Lösung. Das System ist so gebaut, dass die als nicht optimal angesehenen Tupel gar nicht aufgenommen werden, was nicht immer dem optimalen Verhalten entspricht, da somit korrekte Tupel gelöscht werden. Eine mögliche Verbesserung wäre es diese passiv aufzunehmen und später, wenn ein ähnlicher Input erfasst wird, mit einzubeziehen.

1.1.3 Performanztest

Es wurden Tests der Funktionalität und des Mehrwerts des multimodalen Systems durchgeführt. Es ging darum die These zu belegen, dass die Kombination von Sprache und Schrift die Erkennung der nicht im System enthaltenen Wörter deutlich verbessert gegenüber einer reinen Sprach- respektive Schrifterkennung und zu zeigen, dass multimodale Schnittstellen im Vergleich zu unimodalen überlegen sein können.

Der Mächtigkeitstest des Systems wurde anhand von 54 Testinstanzen gemacht, das heisst der „Meetingleiter" hat 54 Gantt-Chart-Balken beschriftet und dabei den Namen ins Mikrofon

gesprochen. Die 54 Testeinheiten wurden anhand 18 Schlüsselsätze durchgeführt, in denen insgesamt 30 nicht bekannte Wörter eingebaut wurden. Bei der Durchführung des Versuches wurden die Testsätze gesprochen und gleichzeitig das nicht bekannte Wort auf die Tafel geschrieben. Ein Beispiel für einen Schlüsselsatz ist „Let's call this task-line concur."[1], in welchem „concur" das dem System unbekannte Wort war.

Alle dem System unbekannten Worte wurden als soche erkannt. Es handelt sich dabei um 18.2% aller gesprochenen Worte, wie in Tabelle 1 zu erkennen ist.

[1] Lass und diese Linie „concur" nennen.

Sätze	**54**
Worte insgesamt	297
Nicht bekannte Worte insgesamt	54
Anteil an nicht bekannten Worten	18.2%
Nicht bekannte Worte als solche erkannt	**100.0%**

Tabelle 1: Erkennung der nicht bekannten Worte [KAISER2004]

UM HW Phone-correct OOV words[1]	35.19%
UM HW Phone substitutions	13.67%
UM HW Phone insertions	1.00%
UM HW Phone deletions	4.67%
UM HW Phone accuracy	80.67%
UM HW Phone Error Rate	**19.33%**

Tabelle 2: Phonetikerkennung aufgrund der Handschrift [KAISER2004]

MM SHW Phone-correct OOV words	38.89%
MM SHW Phone substitutions	11.33%
MM SHW Phone insertions	1.33%
MM SHW Phone deletions	3.67%
MM SHW Phone accuracy	83.67%
MM SHW Phone Error Rate	**16.33%**

Tabelle 3: Multimodale Erkennung der Phonetik [KAISER2004]

Der Mehrnutzen der multimodalen Anordnung gegenüber einer einfachen Handschrifterkennung lässt sich gut an der Erkennung der Phonetik der dem System neuen Worte zeigen. Bei der unimodalen Analyse der Handschrift wurden insgesamt 35.19% aller neuen Worte bezüglich Phonetik korrekt interpretiert, während dieser Wert mit der Zuhilfenahme der Sprachanalyse auf 38.89% anstieg. Auch die Rate der falschen Töne insgesamt ist, wie in Tabelle 4 zu sehen, in der multimodalen Anordnung signifikant besser.

[1] „OOV words“ sind die dem System noch unbekannten Worte

Die Verbesserung von 19.33% falscher Töne gegenüber 16.33% bei multimodaler Analyse ist auf den ersten Blick nicht sehr gross. Es konnten allerdings 15% aller falschen Töne eliminiert werden, alleine dadurch, dass nicht nur die Handschrift für isoliert, sondern in Kombination mit der Spracheingabe analysiert wurde. Dies alleine ist schon eine signifikante Verbesserung, wenn man nun noch in Anbetracht zieht, dass die Sprachangabe eine relativ hohe Fehlerquote von 47.33% aufweist sind die Ergebnisse klar als Erfolg zu werten.

1.1.4 Schlussfolgerung

Das Resultat der Tests zeigte eine gewisse Verbesserung der Worterkennung gegenüber einer unimodalen Analyse. Wie in Tabelle 4 zu sehen ist schneidet die Spracherkennung klar am schlechtesten ab. Dies ist ein deutliches Zeichen, dass die Technik in diesem Bereich noch nicht ausgereift ist. Das Interessante an diesem Resultat ist, dass die multimodale Kombination von Sprach- und Schrifteingabe die Fehlerrate um 3 Prozentpunkte mindern kann, obwohl das zusätzliche Medium der Spracherkennung an sich nicht von herausragender Qualität ist. Dies lässt den Schluss zu, dass in diesem Versuchsaufbau die Multimodalität klare Vorteile gegenüber der Unimodalität gebracht hat.

Testanordnung	**Phone Error Rate**[1]
Unimodale Spracherkennung	47.33%
Unimodale Schrifterkennung	19.33%
Multimodale Sprach- und Schrifterkennung	16.33%

Tabelle 4: Zusammenfassung der Versuchsergebnisse. Mit unimodaler und multimodaler Analyse der dem System nicht bekannten Wörter.

Ein Beispiel wie die Multimodalität hier zur Erkennung behilflich sein konnte ist das folgende: Der Satz „Call this line handoff“[2] wurde von der Schriftenanalyse als „Call this line

[1] Phone Error Rate ist die Rate der Töne, die nicht richtig erkannt wurden

[2] Nenne diese Linie „handoff“.

handifi“ interpretiert. „handifi“ ist in der phonetischen Schreibweise: „hh ae n d iy f iy“, was nur zwei Abänderungen von „hh ae n d ao f“ entfernt ist[1]. Dies wurde vom System richtig erkannt und „handoff“ als wahrscheinlichste Variante ausgegeben.

1.2 Affect-Sensitive Multimodal Human Computer Interaction

Was ist das besondere an menschlicher Kommunikation und wie kann man diese Fähigkeit in einem Computer nachbauen? Diese Frage stellen sich zurzeit viele Forscher, die sich mit dem Thema „Emotionale Intelligenz“ beschäftigen. Dabei geht es bei emotionaler Intelligenz darum den affektiven Zustand eines Benutzers zu erkennen, damit die Interaktion mit Maschinen menschlicher und effektiver gestaltet werden kann [PANTIC2003]. Aber was heisst Intelligenz oder intelligentes Verhalten? Und wie findet Kommunikation zwischen Menschen statt?

Wie [BRUCE1992] schreibt, wissen wir vom Erfolg des Telefons, dass Mimik und Gestik nicht entscheidend sind, um eine effektive Kommunikation zu gewährleisten.

Auf der anderen Seite wäre das automatische Erkennen von Langeweile, Unaufmerksamkeit und Stress eine sehr wichtige Fähigkeit in Situationen, in denen der Fehler einer einzelnen Person eine Reihe von schwerwiegenden Konsequenzen nach sich zieht. So zum Beispiel die Unaufmerksamkeit von Piloten, der Stress von Mitarbeitern in einem Atomkraftwerk, aber auch das Einschlafen von Autofahrern am Steuer [PANTIC2003]. Daneben gibt es noch zahlreiche weitere Anwendungen, bei denen das automatische Erkennen von affektiven Zuständen zum Einsatz kommen könnte. Denkbar ist zum Beispiel das automatische Bewerten von Filminhalten für die Altersklassifizierung indem eine Software zum Einsatz kommt, die in der Lage ist Szenen zu identifizieren die Gewaltdarstellungen enthalten.

[1] Der Levenshtein Abstand ist demnach 2.

Auch die Mensch-Roboter Interaktion ist ein mögliches Anwendungsgebiet. Dies ist vor allem der Fall weil klassische Eingabemethoden wie Maus oder Keyboard für die Interaktion mit Robotern ungeeignet sind. Daneben geht es aber auch darum das Vertrauen in Roboter zu steigern, indem man versucht, Roboter in ihrem Kommunikationsverhalten menschlicher zu gestalten. Neben der Autoindustrie, die an Frühwarnsystemen für das Erkennen von Schlaf forscht, interessiert sich aber auch der Staat für diesen Forschungszweig. Ziel ist es ein System zu entwickeln das durch Videoüberwachung auffälliges Verhalten von Personen automatisch erkennt. Ein solches System könnte neben biometrischen Erkennungsmerkmalen auf Flughäfen einen zusätzlichen Nutzen für die Erhöhung der Sicherheit bringen.

1.2.1 Problemstellung

Wie lassen sich affektive Zustände von Benutzern erkennen? Warum braucht es multimodale Interfaces um eine Eindeutigkeit einen affektiven Zustandes zu gewährleisten? Wie [KELTNER2000] argumentiert, kennt die klassische Psychologie folgende sechs menschliche Grundemotionen: Glücklichkeit, Wut, Traurigkeit, Überraschung, Ekel und Angst. Diese Grundemotionen haben die Eigenschaft, dass sie über verschiedene Kulturen und Sprachen konstant sind. Das heisst bei den meisten Menschen treten diese sechs Grundemotionen in ihrem Alltag auf. So hat zum Beispiel [EKMAN2004] festgestellt, dass japanische und amerikanische Probanden diese sechs Grundemotionen gleich häufig gezeigt haben, während sie sich Filme anschauten die stark auf Emotion ausgerichtet waren.

Die Ausdrucksweise von solchen Emotionen ist aber starken kulturellen Unterschieden unterworfen. Das heisst alle Individuen besitzen zwar die Grundemotionen, äussern diese aber auf sehr unterschiedliche Weise.

Einer der Ansätze um Emotionen in der Sprache zu erkennen, beruht auf den prosodischen Merkmalen der Sprache wie Akzent, Tonhöhe und Druckstärke [CHEN1998]. Chen hat in

seinem Experiment versucht, solche Sprachmerkmale einer Emotion zuzuordnen. Die folgenden Grafiken zeigen Beispielhaft wie eine solche Zuordnung funktioniert. Er hat die Tonhöhe der Sprache als Messgrösse genommen um aufgrund der Tonhöhe der Sprache auf die Emotion zu schliessen die eine Testperson empfindet. Dabei stellte er fest, dass es bei solchen Zuordnungen zwei grosse Probleme gibt.

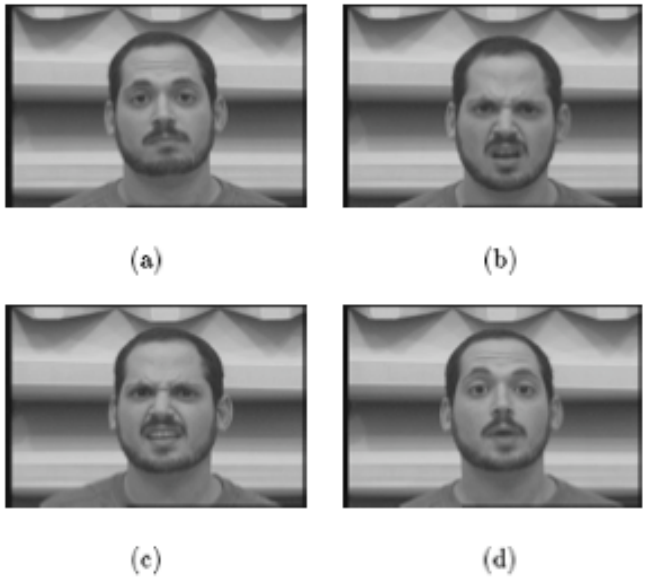

Abbildung 4: Spanisch sprechende Testperson: (a) Traurigkeit, (b) Abneigung, (c) Wut, (d) Überraschung [CHEN1998]

Die Tonhöhe der Sprache ist für sich genommen kein eindeutiges Kriterium für das Erkennen der Emotionen. So sieht man in Abbildung 6, dass die spanisch sprechenden Testpersonen bei einer Tonhöhe um 200 Hz traurig waren oder das Gefühl der Abneigung hatten. Man kann aber mit der Tonhöhe alleine nicht eindeutig entscheiden, welche der beiden Möglichkeiten die richtige ist.

	Happiness	Anger	Fear	Sadness
Pitch	Increase in mean [8], [5], range [41], [5], variability [5]	Increase in mean [27], [5], range [5], variability [5]	Increase in mean [5], range [118], [5]	Decrease in mean [118], [5], range [118], [5]
Intensity	Increased [8], [5]	Increased [8], [5]	Normal [24]	Decreased [27], [5]
Duration (speech rate)	Increased rate [27], [5] Slow tempo [8]	Increased rate [27], [5] Reduced rate [118]	Increased rate [5] Reduced rate [109]	Reduced rate [27], [5]
Pitch contour	Descending line [41]	Descending line [5], stressed syllables ascend frequently & rhythmically [41], irregular up & down inflection [27]	Disintegration in pattern and great number of changes in the direction [24]	Descending line [27], [5]

Abbildung 5: Abbildung von Informationen über multimodale Interfaces. [PANTIC2003]

Auf der anderen Seite hat man festgestellt, dass es zu sehr starken Unterschieden kommt, wenn eine andere Sprache gesprochen wird. Chen hat in seinem Versuch die Sprachen Spanisch und Singhalesisch gewählt, da es sich hier um zwei gegensätzliche Beispiele handelt.

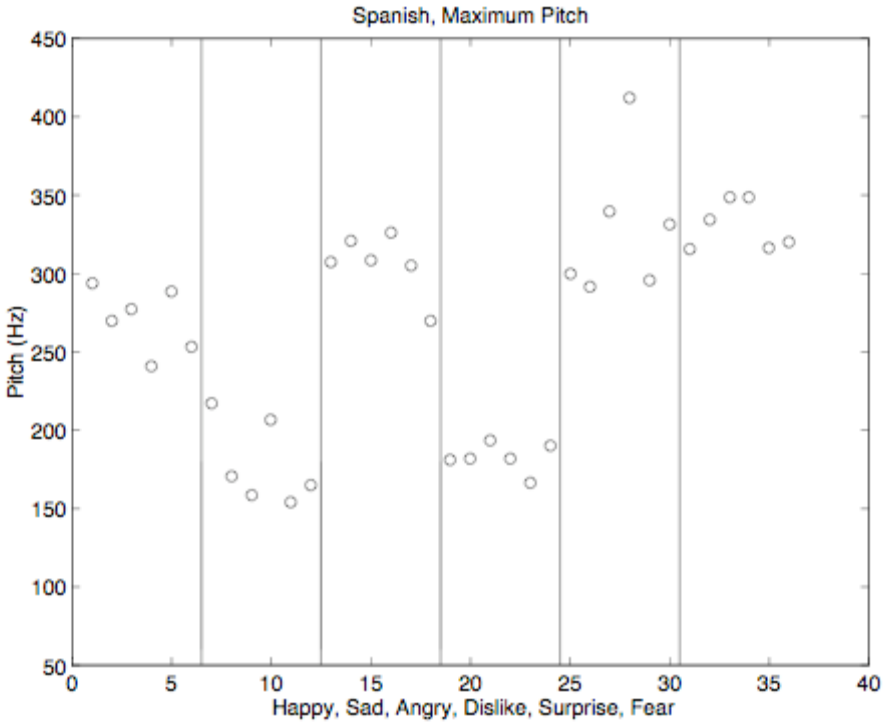

Abbildung 6: Tonhöhenverteilung bei Emotionen (Spanisch) [CHEN1998]

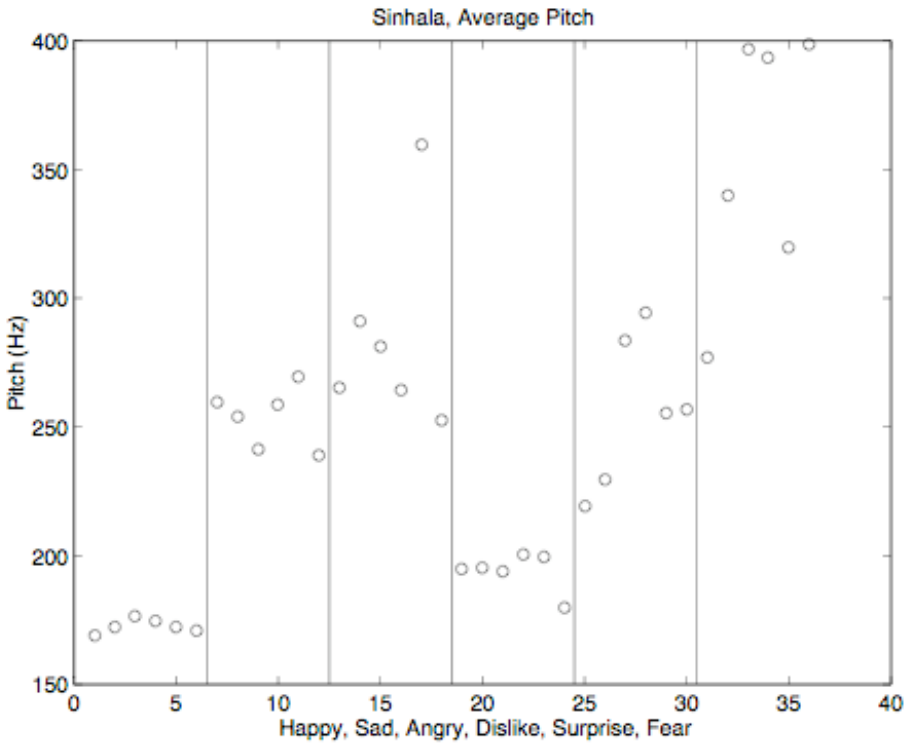

Abbildung 7: Tonhöhenverteilung bei Emotionen (Singhalesisch). [CHEN1998]

Das heisst für das Erkennen von emotionalen Zuständen reicht eine unimodale Erkennungsweise nicht aus. Es braucht ein multimodale Schnittstelle mit dem Benutzer, so dass die Erkennungsrate von emotionalen Zuständen gesteigert werden kann.

So ist die Schwierigkeit des Ansatzes von [PANTIC2003], dass die Informationen, die er auf über Video, Audio und Berührung aufgenommen hat, sinnvoll miteinander kombinieren kann um emotionale Zustände zuverlässig zu erkennen.

1.2.2 Lösungsansatz

In der menschlichen Kommunikation ist es ganz natürlich, dass wir unser Sensorium brauchen um alle Wahrnehmungen die wir machen miteinander zu kombinieren. So fällt es uns oft sehr leicht festzustellen ob zum Beispiel jemand verärgert ist oder nicht. Diese Beobachtung hängt aber nicht alleine von dem ab was unser Kommunikationspartner in einem Gespräch

tatsächlich sagt, Es geht vielmehr darum wie er es sagt. Ist seine Stimme dabei sehr laut und seine Aussprache schnell? Oder benutzt er seine Hände um seine Argumente klar zu untermauern? Vielleicht ist es auch die Art wie sich sein Gesicht verzieht wenn er spricht. Diese Fähigkeit multimodale Inputs miteinander sinnvoll zu kombinieren ist für einen Computer eine ungemein schwierigere Aufgabe als für einen Menschen.

Zu diesem Schluss kommt auch Maya Pantic in ihrem Experiment in dem sie versucht hat, ein System für das automatische Erkennen von affektiven Zuständen zu konstruieren [PANTIC2003]. Bei ihrer Versuchsanordnung ging es darum folgende Fragen zu klären:

1. Welche Informationskanäle, die den menschlichen Kommunikationskanälen entsprechen, sollten für einen Affekt-Analyser verwendet werden?
2. Wie können die Daten, die multimodal gesammelt werden sinnvoll miteinander kombiniert werden um eine menschenähnliche Leistung im Erkennen von affektiven Zuständen zu erreichen?

Modalitäten:

Pantic ist zunächst davon ausgegangen, dass ein System die wichtigsten drei Arten der menschlichen Wahrnehmung - Sehen, Hören und Fühlen – integrieren muss, um ein gutes Resultat zu erzielen. Diese These konnte sie so in Ihrem Versuch jedoch nicht verifizieren. Sie hat vielmehr festgestellt, dass zum Beispiel der tatsächlich gesprochene Inhalt von Worten nur einen sehr kleinen Einfluss auf die Wahrnehmung des affektiven Zustandes hat. Meist spielt es für einen Menschen keine grosse Rolle, was jemand sagt sondern wie er es sagt und wie er sich dabei verhält. Laut Pantics Untersuchung machten sich Versuchspersonen meist eine Meinung über den emotionalen Zustand eines Gesprächspartners, indem sie auf die Mimik und die Intonation ihres Gesprächspartners geschaut haben. Die Körpersprache und das Verhalten spielten meist eine untergeordnete Rolle.

Obwohl in der menschlichen Kommunikation der physische Kontakt eine untergeordnete Rolle spielt, hat Pantic festgestellt, dass physische Messgrössen wie Herzfrequenz oder Schweiss auf der Haut eine sehr wertvolle Datenquelle sind um Zuverlässige Aussagen über den emotionalen Zustand machen zu können.

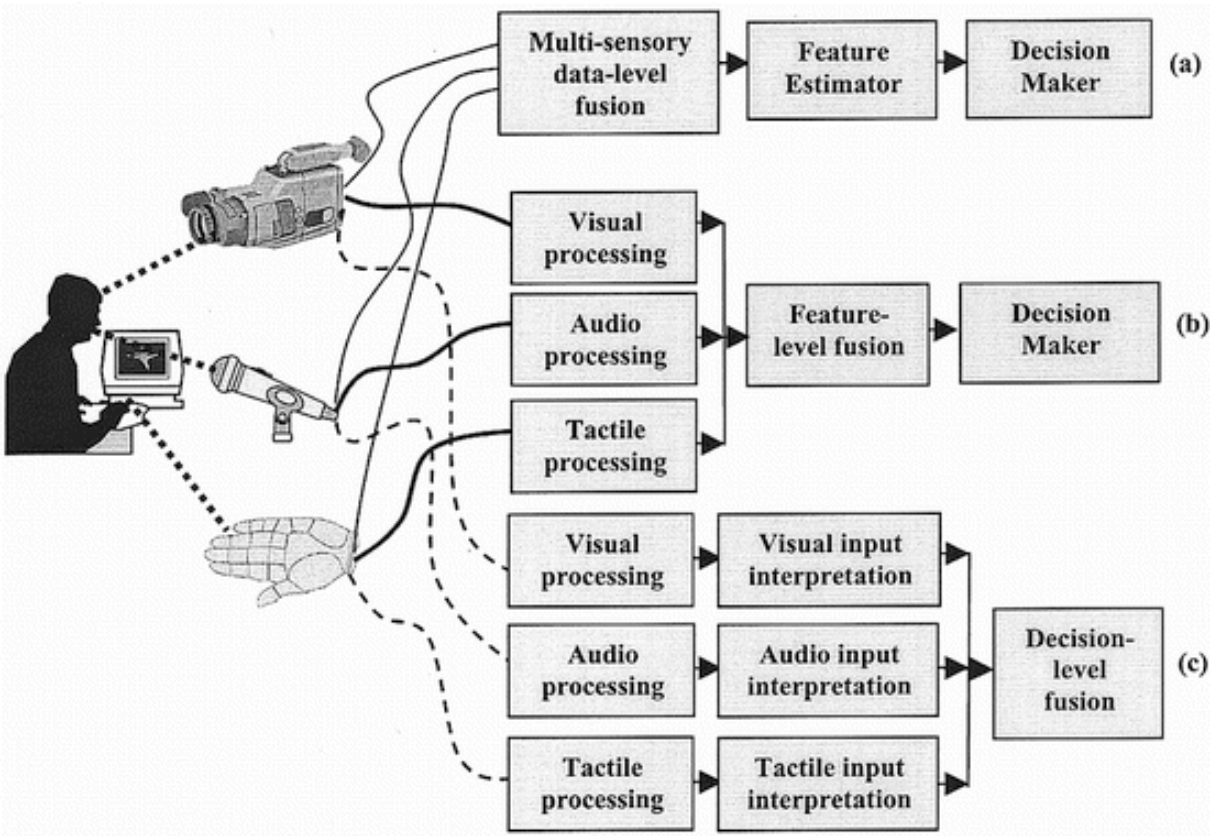

Abbildung 8: Zusammenführung von Informationen über multimodale Interfaces. [PANTIC2003]

(a) Daten-Level Fusion beinhaltet rohe Sensor Daten

(b) Feature-Level Fusion kombiniert Eigenschaften von verschiedenen Interfaces

(c) **Entscheidungs-Level Fusion kombiniert Daten von verschiedenen Interfaces nachdem sie bereits durch das System analysiert sind**

Fusion von Daten über verschiedene Informationsquellen:

Pantic hat festgestellt, dass die Erkennungsleistung ihres multimodalen Affekt-Analysers nicht nur von den verschiedenen Modalitäten abhängig ist, über die Informationen

aufgenommen werden. Vielmehr ist die Erkennungsleistung der Software abhängig von der Technik die eingesetzt wird um die Daten miteinander zu fusionieren. Dabei hatte sie das Ziel, einen die menschliche Kombinationsfähigkeit von verschiedenen Informationen möglichst gut nachzubilden. Um das Problem zu lösen hat sie sich vor allem an der Neurologie orientiert, die zum Schluss gekommen ist, dass Menschen diese konzeptionelle und zusammenführende Fähigkeit entwickeln, indem sie drei Strategien verfolgen:

1. **1+1>2:** Bei Menschen wurde ein stärkerer neuronaler Impuls festgestellt, wenn mehre schwache Reize aufgenommen wurden als bei einem einzelnen starken Reiz. Das heisst Menschen haben eine Art Filter, der es ihnen erlaubt Dingen eher dann zu glauben, wenn sie sie über mehrere Arten wahrgenommen haben.
2. **Kontext Abhängigkeit:** Die Art wie auf Reize reagiert wird ist abhängig davon, welche Kombination von Reizen das menschliche Gehirn bekommt. Nicht jede Kombination von Reizen führt zum gleichen Verhalten.
3. **Umgehen mit sich widersprechenden Informationen:** Je nach wahrgenommenem Kontext werden widersprüchliche Wahrnehmungen im menschlichen Gehirn anders behandelt. Entweder wir die Wahrnehmung einer widersprüchlichen Wahrnehmung ausgeblendet, zum Beispiel wenn es sich um eine Notsituation handelt, oder man schaut noch ein zweites Mal hin um so den Widerspruch in der Wahrnehmung zu lösen.

Panctic implementiert in ihrem System genau diese Mechanismen. Das heisst sie versucht die menschliche Wahrnehmung in Software nachzubilden. Sie kommt in Ihrem Versuch [PANTIC2003] zum Schluss, dass unimodale Interfaces zur affektiven Zustandserkennung nur unbefriedigende Resultate liefern. Die wichtigste Erkennungsart von affektiven Zuständen ist neben der Spracherkennung das Wahrnehmen von Gesichtsgestik. Obwohl festgehalten

wird, dass der Ansatz sehr viel versprechend aussieht weist sie auch darauf hin, dass die Probanden sich durch die enge Interaktion mit dem Computer oft beobachtet vorgekommen sind. Namentlich die dauernde Videobeobachtung oder das konstante Messen von Herz- oder Atemfrequenz wurde von den Testpersonen als unangenehm empfunden.

Trotzdem kommt sie zum Schluss, dass mit einem solchen Ansatz wesentliche intelligentere Computer entwickelt werden können, die ihre Umwelt nicht nur wahrnehmen, sondern sie auch sinnvoll interpretieren können.

2 Fazit

Die Forschung auf dem Gebiet der multimodalen Interfaces hat sehr grosse Fortschritte gemacht. Noch vor wenigen Jahren war Handschriftenerkennung und Spracherkennung eine Spielerei, die für geschäftskritische Anwendungen nicht eingesetzt werden konnte, da die Technik viel zu unzuverlässig war. Dieses Defizit ist in den letzten Jahren, auch dank immer leistungsstärkerer Hardware wesentlich verbessert worden. So bieten zum Beispiel IBM und Philipps Spracherkennungssoftware an, die auf eine Genauigkeit von ungefähr 98% kommt. Das heisst auf 100 Wörter produziert die Software nur zwei Fehler. Auch Handschriftenerkennung durch den Computer ist für sich genommen so gut geworden, dass sie sich produktiv einsetzen lässt.

Problematisch ist das Gebiet nach wie vor dann, wenn versucht wird Informationen, die von multimodalen Interfaces generiert werden, miteinander zu synchronisieren. Allzu oft ist das Problem, dass eine Eingabe oder Erkennungsmethode für sich alleine keine eindeutigen Resultate liefert. Das Problem dabei ist, dass Informationen auch widersprüchlich sein können, was dazu führt dass ein Computer keine Entscheidung mehr treffen kann. Menschen fällt es hingegen oft leicht Informationen zu interpretieren, auch wenn sie noch so widersprüchlich sein mögen. Diese Fähigkeit Informationen zu Filtern, Informationen zu interpretieren und aus widersprüchlichen Informationen richtige Entscheide treffen zu können wird eine grosse Herausforderung von multimodalen Interfaces bleiben, da diese Fähigkeiten in einer Maschine sehr schwierig nachzubauen sind.

Es werden noch Jahre vergehen bis neue Eingabemethoden und Gestenerkennung die traditionellen Interaktionswege mit Computern ersetzen werden. Gelingt es aber diese Informationen miteinander zu vereinen, wird die Bedienung von Computern wesentlich intuitiver und schneller.

3 Literaturverzeichnis

[ADLER1997] Michael Adler, Antique Typewriters - From Creed to QWERTY, 1997, Schiffer Publishing

[BLACK1999] Paul E Black, Algorithms and Theory of Computation Handbook, CRC Press LLC, 1999, "Levenshtein distance": In Dictionary of Algorithms and Data Structures [online], ed., U.S. National Institute of Standards and Technology, 10 November 2005. Verfügbar auf: http://www.nist.gov/dads/HTML/Levenshtein.html

[Bruce1992] V. Bruce, "What the human face tells the human mind: Some challenges for the robot-human interface," in Proc. ROMAN, 1992, pp. 44-51

[CHEN1998] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in Proc. FG, 1998, pp. 396-401.

[EKMAN2004] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in Proc. Nebraska Symp. Motivation, J. Cole, Ed. , 2004, pp. 207-283.

[GOGATE2001] L. J. Gogate, A. S Walker-Andrews, et al., The Intersensory Origins of Word Comprehension: an Ecological-Dynamic Systems View. In: *Development Science* **4**(1), 2001

[KAISER2004] Edward C. Kaiser, "Dynamic New Vocabulary Enrollment through Handwriting and Speech in a Multimodal Scheduling Application," Making Pen-Based

Interaction Intelligent and Natural. In: Papers from the 2004 AAAI Symposium, Technical Report FS-04-06, Arlington, VA., USA, October 21-24, 2004

[KAISER2005] Edward C Kaiser, "Multimodal New Vocabulary Recognition through Speech and Handwriting in a Whiteboard Scheduling Application," Proceedings of the International Conference on Intelligent User Interfaces, San Diego, CA., January 9-12, 2005

[KELTNER2000] D. Keltner and P. Ekman, “Facial expression of emotion,” Handbook of Emotions, M. Lewis and J. M. Havil-Jones, Ed. New York: Guilford, 2000, pp. 236-249.

[PANTIC2003] Pantic, M.; Rothkrantz, L.J. M; Toward an Affect-Sensitive Multimodal Human Computer Interaction. In: Proceedings of the IEEE, Vol. 91, September 2003