K. KROSCHEL (HRSG.), ***ESSV 2003. 14. Konferenz zur Elektronischen Sprachsignalverarbeitung*****, KARLSRUHE, 2003, (S. 114–121).**

EMOTIONSERKENNUNG IN EINEM AUTOMATISCHEN DIALOGSYSTEM: IST DER MENSCH BESSER ALS DIE MASCHINE?

Viktor Zeißler, Johann Adelhardt, Elmar Nöth
Lehrstuhl für Mustererkennung, Friedrich-Alexander Universität, 91058 Erlangen
zeissler@immd5.informatik.uni-erlangen.de

Abstract: Für die Dialogführung in einem modernen Sprachdialogsystem ist es essentiell zu wissen, ob der Dialog wie erwartet verläuft oder vom vorgesehenen Schema abweicht, was eventuell zur Unzufriedenheit des Benutzers führen kann. Die Fähigkeit, solche Situationen frühzeitig zu erkennen, ist für ein gutes Sprachdialogsystem unabdingbar. In diesem Zusammenhang werden in dieser Arbeit das Konzept des emotionalen *Benutzerzustandes* eingeführt und erste Experimente zur automatischen Erkennung dieses Zustandes durch die prosodische Analyse vorgestellt. Da die Erkennungsleistung eines automatisches Systems in der Regel weit hinter der menschlichen Erkennungsfähigkeit liegt, wird der Schwerpunkt der Arbeit insbesondere auf einen Vergleich zwischen Mensch und Maschine gelegt. Es wird gezeigt, dass ihre Stärken und Schwächen in unterschiedlichen Bereichen liegen, wobei in speziellen Fällen die Maschine durchaus eine vergleichbare Leistung erzielt. Mögliche Ursachen dafür werden diskutiert. Desweiteren werden hier eine automatische und eine manuelle Selektionsmethode des Trainingsmaterials verglichen, die es erlauben, eine höhere Qualität der Stichprobe zu erreichen und damit bessere Klassifikatoren trainieren zu können. Durch diese Selektion wurde eine Steigerung der Erkennungsrate von 59 % auf 63 % erzielt.

1 Einleitung

Eine Möglichkeit, den Dialogverlauf bei automatischen Sprachdialogsystemen robuster und benutzerfreundlicher zu gestalten, besteht in der Auswertung zusätzlicher Informationsquellen. So wird im Projekt SmartKom[1] ein multimodaler Ansatz verfolgt, bei dem Sprache, Mimik und Gestik des Benutzers ausgewertet werden [8]. Damit kann der Benutzerzustand automatisch detektiert werden, was dem Dialogmanager ermöglicht, die richtige Dialogstrategie auszuwählen. Zur ausführlichen Beschreibung der Klassifikatoren für die einzelnen Modalitäten sei auf [1] verwiesen worden. Diese Arbeit konzentriert sich dagegen auf die Erkennung des Benutzerzustandes anhand nur einer Modalität, der Sprache.

Der Begriff des *Benutzerzustandes*, der unter anderem in [3] eingeführt wird, umfasst die möglichen Emotionen des Benutzers eines automatischen Dialogsystems bzw. die Vorstufen davon, die für die Mensch-Maschine-Kommunikation relevant sein können. Dies beinhaltet beispielsweise solche Emotionen wie (Un-)Zufriedenheit, Ungeduld, Irritation, Ratlosigkeit und schließt die Emotionen aus, die nur in der zwischenmenschlichen Kommunikation auftreten,

[1]Das dieser Arbeit zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (*BMBF*) im Rahmen des SMARTKOM-Projekts unter dem Förderkennzeichen 01 IL 905 K7 gefördert. Die Verantwortung für den Inhalt liegt bei den Autoren.

wie Hass, Verachtung usw. Um die Emotionenvielfalt zu reduzieren, werden hier die vier folgenden Benutzerzustände unterschieden:

- *Freude* – alle positiven Emotionen, die allgemeine Zufriedenheit mit dem Dialogverlauf bzw. den dargestellten Inhalten ausdrücken,
- *Neutral* – keine Emotionen,
- *Zögern* – Unentschlossenheit, Verwirrung, Hilflosigkeit, Ratlosigkeit: der Benutzer weiß nicht, wie es weitergehen soll,
- *Ärger* – Unzufriedenheit, Mißerfolg, Irritation, Enttäuschung: der Benutzer kommt mit dem System nicht zurecht.

Während beim Auftreten von Freude oder Neutral kein unmittelbarer Handlungsbedarf besteht, weist die Detektion von Zögern bzw. Ärger auf ein mögliches Kommunikationsproblem hin. Die Dialogstrategie muss dann dahingehend geändert werden, dass der Benutzer die notwendige Hilfestellung bekommt und den Dialog ungestört fortsetzen kann.

In einer Mensch-Mensch-Kommunikation werden solche Problemsituationen meist frühzeitig erkannt und behoben, bevor es zu ernsten Schwierigkeiten im Dialogverlauf kommt. Der Mensch benutzt dabei sein umfangreiches Weltwissen und analysiert alle Kommunikationsmodalitäten, wie Sprache (Semantik und Prosodie), Mimik und Gestik des Gesprächspartners. Ein automatisches System dagegen hat ein sehr beschränktes a priori Wissen und wertet die unterschiedlichen Modalitäten einzeln aus. Will man die Benutzerzustände anhand nur einer Modalität erkennen, so stellt sich die Frage, wie repräsentativ in dieser Hinsicht die gewählte Modalität ist, und wie gut der Mensch selber diese Benutzerzustände in der gewählten Modalität erkennen kann. Ist diese Frage beantwortet, kann man die Erkennungsleistung von Mensch und Maschine vergleichen, um auf diese Weise den Schwachstellen des automatischen Klassifikators auf die Spur zu kommen.

Ein weiterer Gesichtspunkt, der in dieser Arbeit untersucht wird, ist die Qualität der Trainingsstichprobe. Bei den aufgenommenen Äußerungen wird der Benutzerzustand nicht immer deutlich ausgedrückt, was negative Auswirkungen auf die Qualität der Stichprobe und folglich auf die Ergebnisse der Klassifikation hat. Eine einfache Methode, dagegen vorzugehen, besteht in der Vorselektion "guter" Sätze (*bootstrapping*-Ansatz), die dann anschließend für das Training verwendet werden.

Die Extraktion der prosodischen Merkmale sowie die eingesetzten Klassifikatoren werden in Abschnitt 2 beschrieben. Es folgt die Beschreibung der Stichprobe in Abschnitt 3. Die Experimente werden in Abschnitt 4 dargestellt, die Diskussion der Ergebnisse findet in Abschnitt 5 statt.

2 Prosodische Analyse

Für die Berechnung der prosodischen Merkmale und ihre Klassifikation wurde in dieser Arbeit das Prosodiemodul verwendet, das seinen Einsatz im aktuellen SmartKom-System findet. Im weiteren folgt eine Übersicht über die Extraktion prosodischer Merkmale und die Klassifikation; die weiteren Details findet man in [2, 5].

2.1 Merkmalsberechnung

Die Berechnung der prosodischen Merkmale erfolgt in zwei Stufen. Zunächst werden die Grundfrequenz und die geglättete Signalenergie (sogenannte "Basismerkmale") frameweise mit einer

Fortschaltzeit von 10 ms berechnet. Für die prosodische Analyse einer Äußerung können diese Merkmale allerdings nicht direkt verwendet werden, da die Information über suprasegmentale Eigenschaften der Äußerung darin nicht enthalten ist. Deswegen werden im nächsten Schritt die sogenannten "strukturellen" Merkmale berechnet, die den Verlauf der F0- und der Energiewerte in einem ausgewählten Segment sowie die Dauereigenschaften dieses Segments beschreiben. Als Segmenteinheit wird im aktuellen System das Wort verwendet. Um die längeren prosodischen Strukturen erfassen zu können, nutzt man die Kontextinformation, indem man die Merkmale für zwei links und rechts liegende Nachbarwörter mitberechnet.

Die Positionen der Wortgrenzen, die man für diesen Ansatz benötigt, werden mit *forced alignment* anhand der tatsächlich gesprochenen Wortkette errechnet. Nachdem sie bekannt sind, werden für jedes Wort diverse Statistiken über die entsprechenden Energie- und F0-Konturen bestimmt: der Mittelwert, die Extrema und ihre Positionen, Regressionskoeffizient, Regressionsfehler und andere. Für die Energie und Dauer werden zusätzlich die absoluten und die intrinsisch normierten Werte berechnet. Für ein Einzelwort ergibt sich auf diese Weise ein Satz von 25 Merkmalen. Bei der Analyse der längeren Abschnitte kommen noch die Merkmale für die Nachbarwörter hinzu. Ein optimierter Merkmalssatz für 1er-Kontext enthält 69 Merkmale, für 2er-Kontext 91 Merkmale.

Zusätzlich zu den prosodischen Merkmalen wurden in einigen Experimenten die linguistischen *Part-Of-Speech*-Merkmale (POS) verwendet. Näheres dazu entnimmt man [4].

2.2 Klassifikation

Für die Klassifikation der Benutzerzustände werden hier Künstliche Neuronale Netze (KNN) in Form eines Mehrschichtperzeptrons (MSP) eingesetzt. Als Trainingsalgorithmus kam der *RPROP*-Algorithmus zum Einsatz [6]. Als Grundtopologie wurde ein MSP mit nur einer verborgenen Schicht gewählt, da die im Vorfeld getesteten komplexeren Topologien schlechter abgeschnitten haben. In der Ausgabeschicht wurde jeder zu klassifizierende Benutzerzustand mit einem Neuron repräsentiert. Auf diese Weise konnten die Aktivierungswerte von Neuronen direkt in die Ausgabewahrscheinlichkeit umgerechnet werden.

Der Erfolg des Trainings hängt bei KNN erfahrungsmäßig sehr stark von unterschiedlichen Konfigurationsparametern wie KNN-Topologie, Initialisierung und Schrittweite ab. Um eine optimale Trainingskonfiguration für jeweils eine Merkmalsmenge oder eine Stichprobe zu erhalten, wurden alle Kombinationen dieser Parameter (mit einer sinnvollen Schrittweite, um den Rechenaufwand zu begrenzen) systematisch ausprobiert. Ausgewählt wurde die Konfiguration, die die besten Ergebnisse auf der separaten Validierungsstichprobe geliefert hat.

Um nicht nur ein einziges Wort, sondern eine ganze Äußerung klassifizieren zu können, müssen die Ausgabewahrscheinlichkeiten des Klassifikators $P(s \mid \omega_i)$, dass das Wort ω_i im Benutzerzustand s produziert wurde, zu einer Benutzerzustandswahrscheinlichkeit für alle Wörter in der Äußerung verknüpft werden. Unter der Annahme, dass die Ausgabewahrscheinlichkeiten für alle Wörter in der Äußerung unabhängig voneinander sind, erfolgt die Berechnung nach der folgenden Vorschrift:

$$P(s \mid \omega_1, \omega_2, \ldots, \omega_n) \approx \prod_{i=1}^{n} P(s \mid \omega_i) \; . \tag{1}$$

3 Stichprobe

3.1 Datenaufnahme

Ein wesentlicher Faktor, der über den Erfolg des Einsatzes der Benutzerzustandsklassifikation in einem Dialogsystem entscheiden kann, ist die richtige Zusammenstellung der Trainingsstichprobe. Sie sollte genügend groß sein, um den Klassifikator damit robust trainieren zu können, und nach Möglichkeit alle Variabilitäten repräsentieren, die im täglichen Einsatz auftreten können. Das ideale Trainingsmaterial sind demnach die spontanen Benutzeräußerungen, die in einem realen Mensch-Maschine-Dialog aufgezeichnet wurden. Die unterschiedlichen Benutzerzustände sollten dabei möglichst gleichverteilt sein, da die statistischen Klassifikationssysteme und neuronale Netze in diesem Fall wesentlich bessere Ergebnisse liefern. Es hat sich jedoch gezeigt, dass die gestellten Anforderungen zum Teil widersprüchlich und in der Realität kaum umsetzbar sind. Bei den sogenannten *Wizard-Of-Oz*-Aufnahmen in SmartKom [7, 4] war es das Ziel, eine umfangreiche Stichprobe mit den spontanen Benutzeräußerungen sowie dem spontanen Benutzerzustandswechsel aufzunehmen. Obwohl es absichtlich durch die Aktionen des Wizards Situationen geschaffen wurden, in denen der Benutzer seine Emotionen zeigen sollte, war es nicht gelungen, emotionale Benutzerzustände in dem Umfang aufzunehmen, dass es zum robusten Training der Klassifikatoren ausgereicht hätte. Deswegen wurde bei dieser Arbeit entschieden, auf die spontanen Äußerungen zu verzichten und statt dessen simulierte Benutzerzustände zu verwenden.

Um eine ausreichende Datenmenge für die Experimente zur Verfügung zu stellen, wurde ein umfangreiches Sprachkorpus von 22 weiblichen und 41 männlichen Sprechern aufgenommen. Die Sprecher wurden angehalten, am Bildschirm gezeigte Sätze so zu sprechen, als ob sie sich in einem der Benutzerzustände befänden und ihre Emotionen ausdrücken wollten. Die Sätze wurden zufällig aus drei folgenden Listen ausgewählt:

- *Kurz-* bzw. *Einzelwortsätze* – sie enthalten die Namen eines TV-Genres, wie z.B. “Krimi” oder “Quizshow”, die je nach Benutzerzustand unterschiedlich geäußert werden.
- *längere* bzw. *Rahmensätze* – sie beinhalten emotionale Äußerungen, die spezifisch für jeden Benutzerzustand sind, beispielsweise “nicht schon wieder Krimi!” für Ärger. In jedem dieser Sätze sind ein oder mehrere TV-Genres aus der ersten Liste enthalten. Die Sätze werden generiert, indem zufällig eine Kombination aus einem der verfügbaren “Rahmen” und einem der TV-Genres erstellt wird.
- *allgemeine Ausrufe* – diese Äußerungen sind im Unterschied zu der zweiten Gruppe unabhängig vom Benutzerzustand, wie z.B. “tolles Programm!”, was auch im Ärger-Zustand sarkastisch gesagt werden kann.

Von jedem Sprecher wurden pro Benutzerzustand 20 unterschiedliche Sätze aufgezeichnet, die alle drei Gruppen repräsentieren. Bei der dritten Gruppe waren statt vier nur drei Benutzerzustände vorhanden, da solche emphatischen Äußerungen beim ‘Zögern’ kaum zu erwarten sind.

3.2 Manuelle Annotation

Bei den Aufnahmen mit simulierten Benutzerzuständen wird jedem Satz ein Label zugewiesen, das dem Zustand entspricht, den der Benutzer simulieren sollte. Damit wird die sogenannte Referenzannotation erstellt, das aber bei weitem nicht perfekt ist. Es tritt beispielsweise oft der Fall auf, dass der Benutzer statt eines emotionalen Zustands eher in eine neutrale Stimmung fällt, da kein echter Grund für die Aufregung existiert. Manchmal werden auch die emotionalen

Zustände miteinander verwechselt, insbesondere, wenn die Testpersonen eine unkonventionelle Sprechweise haben oder sich einfach "verspielen". Um diese "menschlichen" Fehler zu analysieren, wurden die aufgenommenen Sprachdaten nachträglich von mehreren Personen annotiert. Die Annotierer arbeiteten unabhängig voneinander. Sie wurden zudem angehalten, nicht auf den Inhalt der Äußerungen zu achten, sondern ausschließlich die Sprechweise bzw. Sprachmelodie zu bewerten. Die Bewertung erfolgte folgendermaßen: die Sätze wurden in einer zufälligen Reihenfolge vorgespielt, wobei jeder Satz nach Wunsch mehrfach angehört werden konnte. Nach jedem Satz musste sich der Annotierer dann für einen der vorgegebenen Benutzerzustände entscheiden.

Um die ideale Erkennungsfähigkeit des Menschen zu ermitteln, wurden die *Handlabels* anhand einer Mehrheitsentscheidung über aller Annotierer erstellt und dann mit der Referenannotation verglichen. Wenn keine absolute Mehrheit vorlag, erfolge eine Abbildung auf die spezielle Rückweisungsklasse, die in der Referenzannotation nicht enthalten war. Außer der Erkennungsfähigkeit des "idealen" Menschen, ist die Erkennungsfähigkeit des "mittleren" Menschen von großem Interesse. Um sie festzustellen, wurde ein Mittelwert der Erkennungsraten von jedem einzelnen Annotierer berechnet. Dieser Mittelwert kann unter Umständen auch größer als die "ideale" Erkennungsrate sein, was damit zusammenhängt, dass die Rückweisung bei der Mehrheitsentscheidung immer als Fehler interpretiert wird. Der Unterschied der beiden Werte hängt direkt mit der sogenannten *'Interlabeller-Übereinstimmung'* (engl. *inter-labeller agreement*) zusammen, deren Untersuchung für aktuelle Emotionsforschung wichtig ist [4].

4 Experimente

4.1 Mensch-Maschine Vergleich

Die Ergebnisse des Vergleichs für alle drei oben beschriebenen Satzgruppen stehen in den ersten drei Zeilen der Tabelle 1. In der ersten Zeile ist die ideale menschliche Erkennungsrate, in der zweiten Zeile der Mittelwert angegeben. Die dritte Zeile zeigt die Leistung des automatischen Systems. Die Erkennungsraten beziehen sich auf die Sätze und wurden auf einer zufällig ausgewählten Validierungsmenge berechnet, die nicht Teil der Trainingsstichprobe war. Bei den Einzelwort- und Rahmensätzen wurden die Trainings- und Validierungsmenge nicht explizit nach Sprechern und Wörtern bzw. Satzstruktur getrennt. Der gleiche Satz von einem Sprecher kann aber nicht (bis auf wenige Ausnahmen, die auf technische Gründe zurückzuführen sind) gleichzeitig in beiden Mengen vorhanden sein. In der letzten Satzgruppe (allgemeine Ausrufe) ist die Validierungsmenge nach den Sprechern und Sätzen von der Trainingsmenge vollständig getrennt. Da es aber in dieser Satzgruppe zu wenig Material für ein Training gab, wurde die Trainingsmenge in diesem Fall aus allen vorhandenen Satzgruppen zusammengestellt. Dadurch wurden im Vorfeld wesentlich robustere Ergebnisse erzielt gegenüber anderen Varianten der Zusammenstellung.

Für jede der Satzgruppe wurden unterschiedliche, jeweils für diese Gruppe optimierte Merkmalssätze verwendet, wobei sich die Optimierung hauptsächlich auf die Kontextlänge bezog. So tragen bei den Kurzsätzen die Kontext- und Pausenmerkmale keine nützliche Information und können ignoriert werden. Bei den Rahmensätzen hat sich der 1er-Kontext als optimal erwiesen. Zusätzlich wurden in dieser Satzgruppe linguistische POS-Merkmale verwendet. Eine wichtige Randbedingung war dabei, dass der Merkmalssatz keine linguistische Satzstruktur abspeichern soll, was für diese Satzgruppe zu einer 100 %-Erkennungsrate geführt hätte (vgl. Anforderungen an die manuelle Annotation im Abschnitt 3.2). Um dies zu gewährleisten, wurden nur 6 POS-Merkmale für das aktuelle Wort eingesetzt, die POS-Merkmale für die Kontext-Wörter blieben unberücksichtigt. Es konnte jedoch nicht vollständig ausgeschlossen werden, das die prosodische Satzstruktur im Klassifikator abgespeichert wird, was aber im gleichen Maß für

Tabelle 1 - Gesamt- (ER) und klassenweise gemittelte (CL) Erkennungsraten für alle durchgeführten Experimente (zeilenweise) und Satzgruppen (spaltenweise)

Experimente	Menge	Einzelwortsätze		Rahmensätze		Ausrufe	
		CL (%)	ER (%)	CL (%)	ER (%)	CL (%)	ER (%)
Mensch (ideal)	Vali	70	69	93	93	79	78
Mensch (Mittelw.)		71	71	93	93	76	75
Maschine		60	59	94	94	61	60
Maschine / –	Auswahl	65	68	94	94	59	58
Maschine / Masch.		**67**	**69**	89	89	**63**	**62**
Maschine / Mensch		66	65	91	91	57	56

Tabelle 2 - Verwechselungsmatrix zwischen Mensch und Maschine (Anzahl der Fälle)

Mensch		Maschine			
	hat recht:	Freude	Zögern	Ärger	Neutral
Freude	Mensch	**14**	1	2	5
	Maschine		*0*	*0*	*0*
Zögern	Mensch	1	**22**	2	0
	Maschine	*0*		*0*	*0*
Ärger	Mensch	4	5	**6**	2
	Maschine	*0*	*0*		*0*
Neutral	Mensch	6	2	1	**25**
	Maschine	*3*	*2*	*1*	
Rückweisung	Mensch	–	–	–	–
	Maschine	*5*	*2*	*4*	*0*

die menschlichen Annotierer gilt. Für die letzte Satzgruppe, die Ausrufesätze, wurde der 2er-Kontext, allerdings ohne POS-Merkmale verwendet.

Um die Unterschiede in der Erkennung von Benutzerzuständen zwischen Mensch und Maschine zu veranschaulichen wurde eine Verwechselungsmatrix für die Einzelwortsätze (insgesamt 142 Sätze) berechnet, die in der Tabelle 2 angegeben ist. Dabei wurden nur die Fälle berücksichtigt, in denen entweder Mensch oder Maschine das richtige Ergebnis geliefert haben (115 Fälle, 81 %). Jedes Tabellenfeld besteht aus zwei Zeilen: in der oberen steht die Anzahl der Fälle, bei denen der Mensch richtig lag (31 Fälle, 22 %), und in der unteren die Maschine (17 Fälle, 12 %). Auf der Hauptdiagonale hatten beide, Mensch und Maschine eine Übereinstimmung mit der Referenz (67 Fälle, 47 %). Es ist interessant anzumerken, dass in allen Fällen, bei denen sich der Mensch geirrt hat, er sich für Neutral statt eines emotionalen Zustands entschieden hat, oder es hat überhaupt keine Mehrheitsentscheidung gegeben. Dagegen lag die Maschine in vielen dieser Fälle richtig. Dieser Umstand lässt vermuten, dass sie im Vergleich zu Menschen auf feinere Unterschiede reagiert und diese auch richtig interpretieren kann. Besonders deutlich ist dieses Phänomen bei der Freude zu beobachten. Der Grund dafür liegt möglicherweise darin, dass der Mensch dazu tendiert, Emotionen holistisch wahrzunehmen und bei Freude viel mehr auf andere Modalitäten, vor allem auf die Mimik seines Gesprächspartners angewiesen ist und viel weniger auf die Sprache achtet.

4.2 Selektion der Trainingsäußerungen

Das Ziel der im folgenden beschriebenen Experimente war es zu prüfen, ob man durch die Selektion nur guter Sätze in der Trainingsstichprobe die Qualität des Trainings verbessern kann, was durch die besseren Erkennungsraten auf einer fest definierten Validierungsmenge nachgewiesen wird. Als neue Validierungsmenge wurden dabei die Sätze aus der oben beschriebenen Validierungsmenge benutzt, die vom "idealen" Menschen richtig erkannt worden sind. Mit dieser Auswahl sollten die Fehler ausgeschlossen werden, die durch die falsche Referenzannotation zustande kamen.

Es wurden drei Versuchsreihen durchgeführt, deren Ergebnisse in den letzten drei Zeilen der Tabelle 1 angegeben sind. In der ersten Reihe wurden die Klassifikatoren, die auf der gesamten Trainingsmenge trainiert waren (Zeile 3), auf der neuen Validierungsmenge (Auswahl) ausgewertet. Dies stellte eine Vergleichsbasis für die nachfolgenden Experimente dar. In der zweiten und der dritten Experimentenreihe wurden in der Trainingsstichprobe nur solche Sätze beibehalten, die entsprechend von der Maschine (Zeile 5) bzw. dem Menschen (Zeile 6) richtig erkannt werden konnten. Anschließend erfolgte das Training der Klassifikatoren auf den ausgewählten Sätzen und ihre Auswertung auf der neuen Validierungsmenge.

5 Diskussion

Um die Besonderheiten bei der Erkennung des Benutzerzustands von Mensch und Maschine hervorzuheben, vergleichen wir zunächst die Ergebnisse bei den Einzelwortsätzen und Ausrufen. Der Hauptunterschied bei diesen Satzgruppen war das Vorhandensein von Kontext. In beiden Fällen hat die Maschine sehr ähnliche Erkennungsraten geliefert, was darauf hinweist, dass die Kontextinformation bei ihr wenig Einfluss hat. Der Mensch kam mit den Ausrufen wesentlich besser zurecht (79 % gegen 70 %), wobei er anscheinend sein Wissen darüber nutzte, wie die prosodische Struktur des *ganzen* Satzes abhängig vom jeweiligen Benutzerzustand sein soll. Der "mittlere" Mensch war bei den Einzelwortsätzen besser als der "ideale", bei den Ausrufen jedoch um rund 3 Prozentpunkte schlechter. Dies wird plausibel, wenn man bedenkt, dass bei längeren Sätzen viel größere Ausdrucksvariabilitäten auf der Seite des Sprechers und Interpretationsmöglichkeiten auf der Seite des Hörers vorhanden sind — was unweigerlich zu einer schlechteren Interlabeller-Übereinstimmung führt.

Bei den Rahmensätzen hat die Maschine den Menschen sogar überholt. Dies ist allerdings kein ganz fairer Vergleich, da es in beiden Fällen die Möglichkeit zum "Schummeln" gab: der Mensch könnte die Semantik der Sätze verstehen und die Maschine könnte statt der Benutzerzustände einfach die prosodischen Besonderheiten der jeweiligen Satzstruktur klassifizieren. Man hat zwar Vorkehrungen getroffen, die es verhindern sollten, sicherstellen konnte man das aber nicht. Dessen ungeachtet zeigen die Ergebnisse deutlich, dass die Maschine imstande ist, die Emotionen in den Äußerungen mit großer Sicherheit zu erkennen, wenn diese Äußerungen spezifisch für die jeweilige Emotion sind und in der Trainingsstichprobe vorhanden waren.

Die Experimente mit der Selektion des Trainingsmaterials haben ebenfalls bestätigt, dass Mensch und Maschine die Benutzerzustände auf sehr unterschiedliche Weise erkennen. Bei der manuellen Auswahl wurden aus der Trainingsstichprobe alle Grenzfälle eliminiert und nur starke Ausprägungen beibehalten. Das führte nur im Fall von Kurzsätzen zu einer unsignifikanten Verbesserung von 65 % auf 66 %. Bei der automatischen Selektion wurden dagegen die Fehlklassifikationen eliminiert, die vermutlich mit dem verwendeten Merkmalssatz unmöglich zu erkennen waren und beim Training demzufolge nur Störungen verursachen würden. Diese Prozedur führte sowohl bei Kurzsätzen als auch bei Ausrufen zu einer signifikanten Verbesserung von 65 % auf 67 % und von 59 % auf 63 %.

Generell zeigen die in der Tabelle 1 angegebenen Erkennungsraten, dass die Maschine zwar tatsächlich hinter dem Menschen liegt, der Unterschied jedoch nicht sehr groß ist. Die menschlichen Erkennungsraten sind ebenfalls weit von der 100 %-Marke entfernt. Die Stärken des Menschen (und die Schwächen der Maschine) liegen offenbar in der guten Ausnutzung von Kontextinformationen durch die Berücksichtigung der prosodischen Struktur des ganzen Satzes und in einer sehr guten Unterscheidung zwischen emotionalen Benutzerzuständen (siehe in Tabelle 2).

Die Generalisierungsfähigkeit der automatischen Klassifikatoren liegt immer noch weit unter den menschlichen Fähigkeiten. Bei der Erkennung von Benutzerzuständen kann die Maschine ihre starken Seiten erst dann zeigen, wenn im Test keine neuen Satzstrukturen im Vergleich zum Training vorkommen. Ansonsten ist nur eine moderate Leistung zu erwarten.

Für ein automatisches System gibt es zweifelsohne mehrere Verbesserungsmöglichkeiten. Es müssen andere effizientere Formen der Ausnutzung von Kontextinformationen gefunden werden. Die prosodische Analyse auf der Wortbasis muss auf die Phrasen- und Satzprosodie erweitert werden. Eine große Herausforderung in diesem Bereich wäre, eine geeignete Normierung zu finden, die der immensen Phrasenvariabilität entgegenwirken kann. Andererseits müssen neue Merkmale eingesetzt werden, die nicht wie bisher ausschließlich prosodische, sondern auch die spektrale Struktur einer Phrase beschreiben können.

Literatur

[1] J. Adelhardt, R. Shi, C. Frank, V. Zeissler, A. Batliner, E. Nöth, and H. Niemann. Multimodal User State Recognition in a Modern Dialogue System. In *Proc. of the 26th German Conf. on Artificial Intelligence, KI'03 (to appear)*, Hamburg, Germany, 2003.

[2] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The Prosody Module. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 106–121. Springer, New York, Berlin, 2000.

[3] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to Find Trouble in Communication. *Speech Communication*, 40:117–143, 2003.

[4] A. Batliner, V. Zeissler, C. Frank, J. Adelhardt, R. Shi, E. Nöth, and H. Niemann. We are not amused – but how do you know? User states in a multi-modal dialogue system. In *Proc. of the 8th European Conf. on Speech Communication and Technology, EUROSPEECH'03 (to appear)*. Geneva, Switzerland, 2003.

[5] A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Shaker Verlag, Aachen, 1997.

[6] M. Riedmiller and H. Braun. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In *Proc. of the IEEE Intl. Conf. on Neural Networks*, pages 586–591, San Francisco, CA, 1993.

[7] S. Steininger, S. Rabold, O. Dioubina, and F. Schiel. Development of the User-State Conventions for the Multimodal Corpus in SmartKom. In *Proc. of the Workshop "Multimodal Resources and Multimodal Systems Evaluation"*, pages 33–37, Las Palamas, Gran Canaria, Spain, 2002.

[8] W. Wahlster. SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In *Proc. of the Human Computer Interaction Status Conf., HCI2003*, pages 47–62, Berlin, Germany, 2003.