AUTOMATISCHE ERKENNUNG KONTINUIERLICHER SPRACHE IM RAHMEN DES FORSCHUNGSPROJEKTES SPICOS

Hermann Ney

Philips GmbH Forschungslaboratorium Aachen

In diesem Beitrag wird ein Überblick gegeben über die Arbeiten, die innerhalb des durch das BMFT geförderten Verbundprojektes SPICOS von Philips zur automatischen Spracherkennung durchgeführt werden. Der Ansatz zur automatischen Erkennung beruht auf statistischen Verfahren, mit denen die Wortfolge bestimmt wird, die am wahrscheinlichsten gesprochen worden ist. Die Einzelheiten der statistischen Modellierung und die erzielten Ergebnisse werden beschrieben.

1. Einleitung

Die hier beschriebenen Arbeiten sind Teil des Gemeinschaftsprojektes "SPICOS" von Siemens, Philips und des 'Instituut voor Perceptie Onderzoek /IPO)' in Eindhoven, Holland. Das Akronym SPICOS steht für **S**iemens/**P**hilips/**I**PO **CO**ntinuous-**S**peech Recognition and Understanding. Als konkretes Projektziel wurde definiert, ein Datenbank-Anfragesystem für die Kommunikation in gesprochener, natürlicher Sprache aufzubauen. Andere Anwendungen für die im Projekt entwickelten Verfahren sind alle Arten von Dateneingabe in gesprochener Sprache, insbesondere auch die automatisch hörende Diktiermaschine für einen großen Wortschatz. Das Projekt wird im Rahmen eines Verbundprojektes vom BMFT gefördert. Neben den oben genannten Partnern Siemens und Philips sind an diesem BMFT-Projekt beteiligt: AEG, Forschungsinstitut Ulm; TU Berlin, Institut für Fernmeldetechnik; Universität Erlangen-Nürnberg, Lehrstuhl für Informatik 5 (Mustererkennung).

In der ersten Phase des Projektes wurde ein Demonstrationssystem fertiggestellt, welches die kontinuierlich gesprochene Datenbankanfrage erkennt, interpretiert und eine gesprochene Antwort gibt. Es basiert auf einem Vokabular von 917 Wörtern. Die akustischen Referenzdaten sind sprecherabhängig. Die Datenbank enthält Informationen über das Projekt selbst: über Dokumente, Treffen und Personen.

Während das Demonstrationssystem der ersten Phase nur ein einfaches Frage-Antwort-Schema erlaubt, wird in der zweiten Phase des Projektes ein echter Mensch-Maschine-Dialog ermöglicht: Der Benutzer wird in die Lage versetzt, zu einem begonnenen Thema weiterführende Fragen zu stellen und dabei sprachlich auf vorher gemachte Äußerungen Bezug zu nehmen, z.B. durch Verwendung von Personalpronomen. Für die Spracherkennung wurde daher das Vokabular von 917 auf etwa 1200 Wörter erweitert; gleichzeitig wurde ein (fast) Echzeit-System aufgebaut und die Möglichkeit der Sprecheradaption untersucht.

2. Systemüberblick

Das System besteht, wie in Bild 1 dargestellt, aus drei in Serie arbeitenden Modulen, die in verschiedenen Labors entwickelt wurden. Die Aufgabe des ersten Moduls, welcher im Philips Forschungslaboratorium Hamburg erstellt wurde, ist es, die gesprochene Äußerung unter Berücksichtigung aller höheren Wissensquellen zu erkennen. Die dabei verwendete Suchstrategie wird als integrierte Suche bezeichnet. Die Erkennung benötigt je nach Variante 20 bis 100-fache Echtzeit auf einem 1-MIPS-Rechner.

Das Ergebnis des ersten Moduls ist der Text des erkannten Satzes. Er wird an den Dialogmodul übergeben, der ihn interpretiert, die Datenbankanfrage durchführt und eine Antwort formuliert. Dieser Modul wurde bei Siemens und beim IPO entwickelt [1,2]. Der dritte, vom IPO erstellte Modul führt eine Sprachsynthese der Antwort durch [3]. Zur Illustration der Funktion des Gesamtsystems mögen folgende Dialogbeispiele dienen (B = Benutzer, S = System):

1. B: Wer nahm an der letzten Sitzung teil?
   S: Das sind Hamer, Paeseler, Piotrowski und Tomaschewski.

2. B: Zeige das Datum des letzten Briefes von Dr. Schilz an Dr. Lang bezüglich Philips!
   S: Das ist der 3. Mai 1987.
3. B: Wer befaßt sich in Hamburg mit Phonem-Modellen?
   S: Das ist Noll.

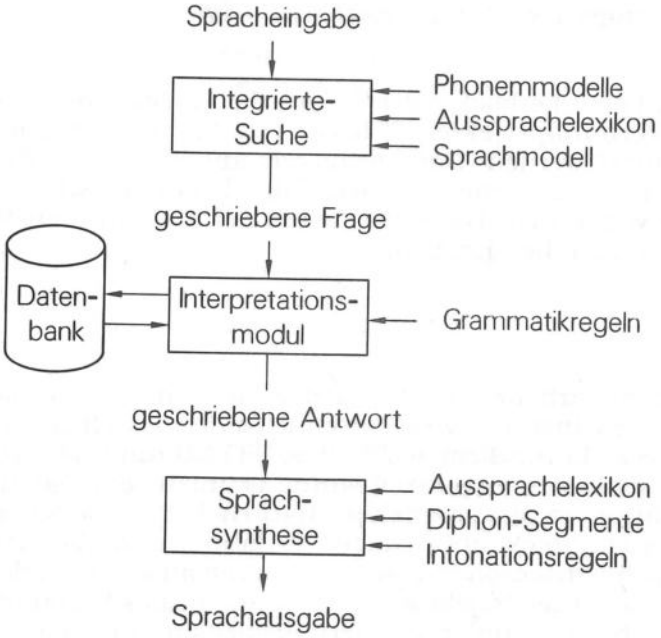

Bild 1. Architektur des Systems.

Die bei Philips durchgeführten Arbeiten konzentrieren sich auf die automatische Erkennung von kontinuierlicher Sprache und die Untersuchung der damit verbundenen Probleme. Diese Probleme der automatischen Spracherkennung sind durch folgende Fakten bedingt:

- Die akustische Realisierung eines Lautes von einer Äußerung zur anderen kann stark variieren, auch wenn es sich um dasselbe Wort und denselben Sprecher handelt.
- Die Sprechgeschwindigkeit kann stark schwanken; es gibt keine feste Zeitskala.
- Die akustische Realisierung eines Lautes hängt häufig von seiner Lautumgebung ab.
- Im akustischen Signal gibt es keine Lautgrenze und bei fließender Sprache keine Wortgrenzen.

Konzeptionell gesehen beruht der Ansatz auf der statistischen Informationstheorie, der bei dem hier gestellten Problem auf eine Verbindung von statistischer und formal-syntaktischer Mustererkennung hinausläuft.

3. Architektur des Erkennungssystems

Die Architektur des Erkennungssystems wird aus der Bayes'schen Entscheidungsregel für minimale Wortfehlerrate abgeleitet und läßt sich vereinfacht durch vier Funktionsblöcke beschreiben:

1. Akustische Analyse und Merkmalsextraktion: Bei gleichzeitiger Datenreduktion werden hier Parameter und Merkmale berechnet, die die zeitabhängigen Eigenschaften des Sprachsignals beschreiben.

2. Akustisch-phonetische Modellierung: Die Aufgabe dieses Funktionsblockes ist es, die Verknüpfung herzustellen zwischen den Parametern und Merkmalen auf der akustischen Ebene und den Symbolen, die die Wörter des Vokabulars repräsentieren. Typischerweise geschieht dies in zwei Stufen: Die Wörter werden anhand der Standard-Aussprache und phonologischer Regeln als Folge von Phonem-Symbolen beschrieben; die Phoneme ihrerseits werden durch stochastische endliche Automaten oder Hidden-Markov-Modelle dargestellt. Diese Modelle werden in der Trainingsphase des Systems individuell für jeden Benutzer aus 200 Sätzen berechnet, die die Phonem-Verteilung der

deutschen Sprache widerspiegeln. Dadurch kann das Vokabular für die Anwendung beliebig gewählt werden; für jedes Wort muß dem System lediglich die Standardaussprache als Folge von Phonemsymbolen bekannt sein.

3. Modellierung der linguistischen Einschränkungen: In diesem Funktionsblock werden rein syntaktisches Wissen über die Sprache und semantisch-pragmatisches Wissen der gegebenen Aufgabenstellung vereinigt, um die Menge der erlaubten Wortfolgen oder Sätze zu beschreiben. Dadurch wird einmal die Verwechselbarkeit auf der akustischen Ebene verringert und so die Erkennungsgenauigkeit erhöht; zum anderen wird der "Entscheidungsraum" verkleinert und so die Rechenzeit für die "Suche" reduziert.

4. Suche und Entscheidungsvorgang: Die Entscheidung über die gesprochene Wortfolge ist so zu treffen, daß sich die beste Erklärung für das beobachtete Sprachsignal ergibt. In der Regel wird das Erkennungsproblem mit statistischen Ansätzen modelliert, so daß die wahrscheinlichste Wortfolge in der Erkennung bestimmt wird. Die akustisch-phonetische Modellierung und die Modellierung der linguistischen Einschränkungen spannen einen strukturierten Entscheidungsraum auf, innerhalb dessen alle erlaubten Sätze liegen. In Abhängigkeit von dem gesprochenen Signal ergibt sich für jede dieser Satzhypothesen eine Bewertung, und die Erkennungsaufgabe besteht nun darin, die Satzhypothese mit der besten Bewertung zu finden. Obwohl dies ein endliches Problem ist, ergeben sich doch erhebliche Schwierigkeiten, weil die Zahl der Satzhypothesen bei 1000 Wörtern und 10 Wörtern per Satz theoretisch bei $1000^{10} = 10^{30}$ liegen kann. Der Suchvorgang muß daher so angelegt sein, daß nur der "relevante" Teil des Entscheidungsraumes ausgewertet wird.

3.1 Akustische Analyse und akustisch-phonetische Modellierung

Die akustische Analyse wird alle 10 ms durchgeführt und berechnet zusätzlich zur Gesamtenergie einen Satz von 30 spektralen Intensitäten, die durch cepstrale Glättung und anschließende Abtastung des Leistungsdichtespektrums ermittelt werden. Um die Dynamik der Sprache zu erfassen, werden zusätzlich die erste und die zweite Ableitung dieser akustischen Parameter berechnet, so daß in den nachfolgenden Schritten insgesamt ein Vektor höherer Dimension verwendet wird. Durch Mittelung benachbarter Spektralbereiche ergibt sich die Dimension zu 63 [4,5].

Die 44 Phoneme werden durch stochastische endliche Automaten oder Hidden-Markov-Modelle modelliert. Die Emissionswahrscheinlichkeiten dieser Phonem-Modelle werden durch Mischungen von kontinuierlichen Verteilungen ('continuous mixture densities') beschrieben [4,5]. Das Training der Phonem-Modelle geschieht anhand der phonetisch balancierten Berliner Sätze in zwei Sitzungen pro Sprecher. Die Überlappungen mit dem SPICOS-Vokabular sind gering (53 Wörter), im wesentlichen handelt es sich dabei um Funktionswörter wie Präpositionen und Artikel.

Die zur Erkennung benutzten Wortmodelle werden durch Verknüpfung der Phonem-Modelle anhand der im Aussprachelexikon festgelegten Regeln gebildet. Das Aussprachelexikon enthält nur die Standardaussprache jedes Lexems. Aussprachevarianten und Koartikulationsphänomene werden durch die Phonem-Modelle erfaßt.

3.2 Sprachmodell (Language Model)

Aufgabe des Sprachmodells ist es, Restriktionen für die Aufeinanderfolge von Wörtern und damit für die erlaubten Sätze zu beschreiben. Diese Restriktionen ergeben sich aus der Syntax, Semantik und Pragmatik. Das Sprachenmodell kann im allgemeinen diese Restriktionen nur approximativ erfassen. Im folgenden werden zwei Varianten betrachtet: ein nicht-stochastisches endliches Netzwerk und ein stochastisches Bigramm-Modell.

Das endliche Netzwerk besteht aus 953 Knoten, 11139 Worttransitionen und 1217 Nulltransitionen. Es wurde im wesentlichen aus Bauplänen für Sätze und Konstituenten unter Berücksichtigung semantischer Gesichtspunkte erstellt [6]. Morphologische Übereinstimmungen wurden dabei häufig nicht erfaßt, weil sie bei dem Verstehen des Satzinhaltes nur eine geringe Rolle spielen.

Die Semantik unseres Diskursbereiches, welche den Inhalt der möglichen Datenbankanfragen definiert, wird in Bild 2 schematisch wiedergegeben. Man findet in dieser semantischen Struktur fünf verschiedene Objektklassen, die auch Datenbankobjekte sind (rechteckig umrahmt) und drei zusätzliche Attributklassen (elliptisch umrahmt). Die Pfeile bezeichnen Relationen und betreffen in erster Linie Datenbankattribute für die Objekte, aber auch Sprachkonstruktionen zur Spezifizierung solcher Objekte. So können z.B. zur Bezeichnung eines Dokumentes Attributwerte für einen Autor und/oder einen Empfänger, ein Datum, ein Thema verwendet werden. Auf diese Art spezifizierte Anweisungen findet man im Dialogbeispiel 2 und in Tabelle 1.

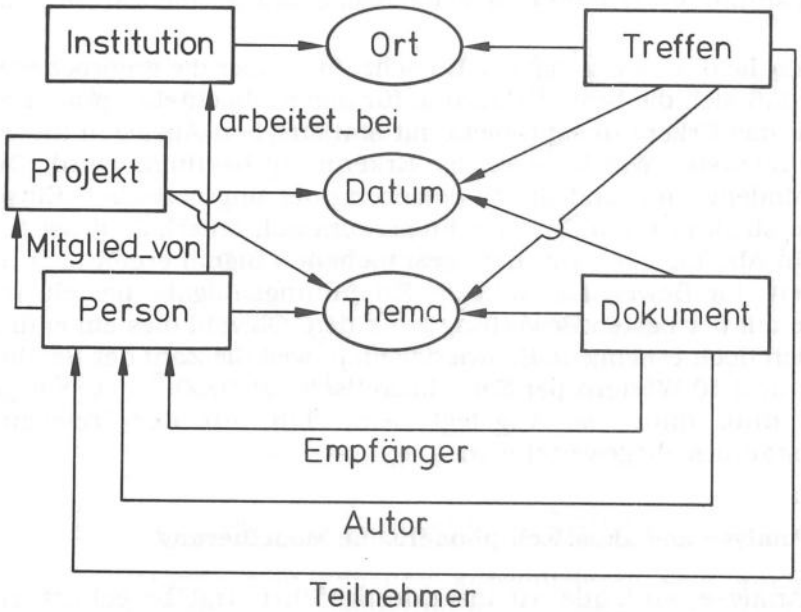

Bild 2. Semantische Struktur der Datenbankeinträge.

Der Grundgedanke des stochastischen Sprachmodells besteht darin, die Auftrittswahrscheinlichkeit eines Wortes in Abhängigkeit von den vorangegangenen Wörtern zu berechnen. Um die Anzahl der zu trainierenden Parameter gering zu halten, beschränkt man die Anzahl der vorangegangenen Wörter meist auf eins oder zwei. Das entstehende stochastische Sprachmodell wird dann Bigramm- oder Trigramm-Modell genannt. Gegenüber nicht-stochastischen Modellen können stochastische Sprachmodelle einige Vorteile bieten: Sie können automatisch durch die Analyse großer Textkorpora trainiert werden; keine Wortfolge wird von vornherein ausgeschlossen, sondern typische Wortfolgen erhalten höhere Wahrscheinlichkeiten.

Da, um die Parameter eines stochastischen Modells zu trainieren, sehr große Textkorpora erforderlich sind, wurde hier ein anderer Weg beschritten: Die Auftrittswahrscheinlichkeiten werden nicht direkt für die einzelnen Wörter des Vokabulars, sondern für Wortkategorien des Vokabulars ermittelt. Das hierfür verwendete Trainingskorpus bestand aus 470 typischen Datenbankanfragen (3761 Wörter), die von mehreren Personen formuliert wurden. Die Wortkategorien basieren auf lexikalischen Klassen wie "Nomen" oder "Adjektiv", die aufgrund der anwendungsspezifischen Semantik weiter unterteilt wurden. Das stochastische Sprachmodell wird also durch die Häufigkeiten von Bi- und Trigrammen bezüglich Wortkategorien und der Abbildung der Kategorien auf die zugehörigen Wörter definiert. Die Folgen von Wortkategorien repräsentieren dabei syntaktische und semantische Aspekte des Anwendungsszenarios. Im SPICOS-System werden 63 bis 100 Wortkategorien verwendet [5].

Der Grad der durch das Sprachmodell erfaßten Restriktionen wird durch die Testsatzperplexität P gemessen. Für einen gegebenen Testkorpus $w_1, \ldots, w_N$ ist P definiert als:

$$P = \sqrt[N]{Pr(w_1 \cdots w_N)} \ ,$$

d.h. als ein geometrischer Mittelwert über die Wahrscheinlichkeiten. Anschaulich bedeutet P die durchschnittliche Zahl von Wörtern, mit der man einen angefangenen Satz fortsetzen kann. Für ein typisches Bigramm-Modell liegt P bei etwa 100. Bei dem nicht-statistischen endlichen Netzwerk nimmt man alle Wörter als gleichwahrscheinlich an und erhält P = 85.

3.3 Suche und Entscheidungsverfahren

Das Sprachmodell, das Aussprachelexikon und die Phonem-Modelle bilden ein vollständiges Modell aller potentiellen Anfragen eines Sprechers. Dadurch wird ein Zustandsraum definiert, der die erlaubten Sätze rigoros einschränkt (nicht-stochastisches Sprachmodell) oder sie mit Wahrscheinlichkeiten versieht (stochastisches Sprachmodell). Das bei der Erkennung eingesetzte Suchverfahren zielt darauf ab, diejenige Zustandsfolge des Zustandsraumes zu bestimmen, die die beobachteten akustischen Daten am besten "erklärt", d.h. die die größte Wahrscheinlichkeit innerhalb der gegebenen Modelle liefert. Konzeptionell ist die Suche eine nichtlineare Zeitanpassung mit dynamischer Programmierung, bei der die Suche auf die wahrscheinlichsten Zustände fokussiert wird (beam search) [7]. Wegen des großen Suchraumes wird eine spezielle Organisation der Suchprozedur verwendet, die in Bild 3 dargestellt ist.

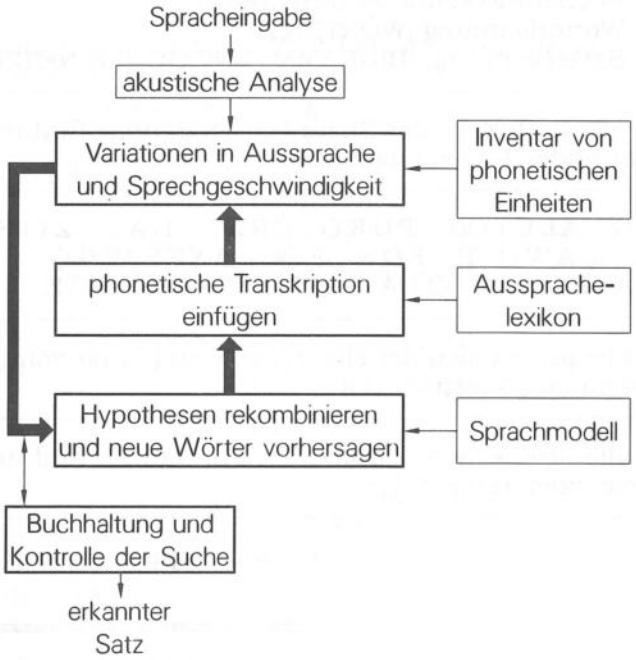

Bild 3. Organisation der Suche.

Die Wissensquellen des Systems werden separat gehalten und während der Suche nach Bedarf ausgewertet. Wird ein Wortende hypothetisiert, so werden mit dem Sprachmodell Wörter ermittelt, die die bisherige Wortfolge fortsetzen. Für diese Wörter werden phonetische Transkriptionen aus dem Aussprachelexikon übernommen und die entsprechenden Phonem-Modelle mit der Spracheingabe verglichen.

Weil alle Wissensquellen während der Erkennung ausgewertet werden, nennt man das Verfahren auch integrierte Suche.

4. Experimentelle Ergebnisse

4.1 Auswirkung der Wissenquellen

Bei den nachfolgend beschriebenen drei Arten von Experimenten werden die oben genannten Wissensquellen nur teilweise ausgewertet, um ihre Funktion bei der Erkennung separat zu erfassen. Die erzielten Ergebnisse sind in Tabelle 1 für zwei Sätze dargestellt.

Bei der *Phonem-Erkennung* wird die am besten mit der aktuellen Äußerung verträgliche Folge von Phonemen ermittelt, wobei jedes Phonem auf jedes folgen kann.

Bei der *Worterkennung* sind einzelne Phoneme für den Suchprozeß unmittelbar nicht mehr sichtbar; es können nur solche Phonemfolgen erkannt werden, die Wörtern des SPICOS-Lexikons entsprechen. Die Aufeinanderfolge von Wörtern unterliegt keinen Beschränkungen, der Suchalgorithmus bestimmt die beste Wortfolge, wobei jedes Wort auf jedes folgen kann.

Bei der *Satzerkennung* werden auch die Wörter nicht mehr als unabhängige Einheiten behandelt, sondern als Bestandteile von Sätzen betrachtet. Es wird die wahrscheinlichste aller gemäß dem Sprachmodell möglichen Wortfolgen bestimmt.

Als Sprachmodell wurden hier die beiden beschriebenen Varianten verwendet: Das Bigramm-Modell und das endliche Netzwerk. An den beiden in der Tabelle 1 gezeigten Sätzen ist klar zu sehen, wie die Fehlerkennungen durch die Verwendung von mehr und mehr Wissensquellen systematisch gesenkt werden.

Tabelle 1: Auswirkungen der Wissensquellen auf die Erkennungsgenauigkeit (Sprecher F-10/2):
- Phonem-Erkennung (PHONEME),
- Worterkennung (WÖRTER),
- Satzerkennung (BIGRAMM, ENDLICHES NETZ).

SATZ 186:	Zeige alle Protokolle der Sitzungen im Januar fünfundachtzig in aufsteigender Reihenfolge!	
PHONEME:	-SAIG ALETOH PUKOLERZT DAE ZITSUNQEENNM - AEM IEAWOER FON FOHDAXTSAEGG - IN AUFSHTAINQDA RAHAIAENFOLBEK -	
WÖRTER:	zeige alle protokollen der sitzungen wem januar von phon acht sechzehn aufsteigender reihenfolge	
BIGRAMM:	zeige alle protokollen der sitzungen im januar fünfundachtzig in aufsteigender reihenfolge	
ENDL. NETZ:	zeige alle protokolle der sitzungen im januar fünfundachtzig in aufsteigender reihenfolge	
SATZ 187:	Gibt es zu der Sitzung am vierzehnten Oktober ein Protokoll ?	
PHONEME:	- IEP - Z EAES - SUH GE ZITSEUNQGK AM FEHRTSAAENPEN OPTOHBAH AN POHBOKOLR -	
WÖRTER:	gibt des zeige sitzungen	am vierzehnten oktober ein protokolle
BIGRAMM:	gibt es zu der sitzung	am vierzehnten oktober an protokolle
ENDL. NETZ:	gibt es zu der sitzungen	am vierzehnten oktober ein protokoll

4.2 Erkennungsergebnisse

Systematische Erkennungsexperimente wurden mit den Sprachprobendaten von vier Sprechern durchgeführt [5]. Die Ergebnisse sind in Tabelle 2 zusammengefaßt. Die Parameter der sprecherabhängigen Referenz-Modelle wurden trainiert an 2 x 100 Sätzen, die anwendungsunabhängig sind und die Phonem-Verteilung der deutschen Sprache widerspiegeln ("Berliner Sätze").

Die Leistungsfähigkeit des Erkennungssystems wird gemessen durch die Wortfehlerrate, die für kontinuierliche Sprache definiert ist als:

$$\text{Wortfehlerrate} = \frac{\text{Auslassungen + Verwechslungen + Einfügungen}}{\text{Zahl der gesprochenen Wörter}}$$

Neben den Auslassungen ("DEL.") und Einfügungen ("INS.") zeigt die Tabelle 2 die Verwechslungen ("SUBSTITUT.") sowohl für die normalen Wörter, d.h. mit voller Endung, als auch für die Wortsstämme ("STEMS"), die man durch Wegstreichen der Endung erhält. Aus der Tabelle geht hervor, daß ungefähr die Hälfte aller Wortverwechslungen durch die Endungen verursacht werden.

Als Maß für den Suchaufwand sind in der Tabelle 2 unter der Spalte "SUCHE" zwei Zahlen angegeben: Die erste Zahl ist die mittlere Anzahl der berechneten Hypothesen pro 10-ms-Sprachsegment; die zweite Zahl ist die Zahl der Fälle, in denen der optimale Pfad für die 200 Testsätze nicht gefunden wurde.

Die Tabelle 2 zeigt, daß die Wortfehlerrate von 22% durch ein Bigramm-Modell (P = 124) auf 9% und durch ein endliches Netz (P = 85) auf 6% verringert wird. Gleichzeitig ergibt sich im Falle des endlichen Netzes eine Reduktion des Suchaufwandes auf ein Drittel.

Tabelle 2: Erkennungsergebnisse für 4 Sprecher und verschiedene Sprachmodelle:
a) kein Sprachmodell P = 917,
b) stoch. Bigramm-Modell: P = 124,
c) nichtstoch. endl. Netz: P = 85.
Für jeden Sprecher wurden Tests an 200 Sätzen = 1391 Wörtern durchgeführt.

	SPRECHER	SUCHE	DEL.	INS.	SUBSTIT.	(STEMS)	SUM.	WORTFEHLERRATE
a)	M-13	6365/4	37	20	162	(91)	219	15.7 %
	M-10	7127/0	61	45	251	(167)	357	25.7 %
	F-01	9274/0	48	34	312	(178)	394	28.3 %
	F-10	7560/4	34	18	193	(116)	245	17.6 %
	MITW:	7582/2						21.8 %
b)	M-03	5746/1	2	4	71	(25)	77	5.5 %
	M-10	7267/1	24	28	138	(62)	190	13.7 %
	F-01	7034/2	13	8	139	(49)	160	11.5 %
	F-10	7110/0	6	6	67	(21)	79	5.7 %
	MITW:	6789/1						9.1 %
c)	M-03	1676/4	8	3	42	(16)	53	3.8 %
	M-10	3909/0	14	14	86	(44)	114	8.2 %
	F-01	2423/3	18	2	85	(22)	105	7.5 %
	F-10	2344/4	11	0	54	(19)	65	4.7 %
	MITW.:	2588/3						6.0 %

5. Diskussion

Die hier beschriebenen Methoden erfassen die Wirklichkeit nur näherungsweise:
- Die Phonem-Modelle sind kontextunabhängig.
- Das Aussprachelexikon enthält nur eine Aussprachevariante.
- Das Sprachmodell läßt Wortfolgen zu, die morphologisch nicht konsistent sind.

Trotzdem bilden die Phonem-Modelle die Grundlage der akustischen Erkennung und leisten die anderen Wissensquellen ihren Beitrag zur Verbesserung der Erkennungsrate, wie die Tabellen 1 und 2 anschaulich zeigen. Das Wesentliche unseres Systems ist dabei, daß Entscheidungen während des Erkennungsprozesses (Weglassen von unwahrscheinlichen Hypothesen, pruning) nicht unter Berücksichtigung nur einer Wissensquelle getroffen werden, sondern erst nach Auswertung aller Wissensquellen.

Der hier beschriebene Ansatz ist in mehrere Richtungen erweitert worden:

- Um eine einfachere Ankopplung der konventionellen linguistischen Wissensquellen an die akustische Erkennung zu ermöglichen, wurde das Suchverfahren so erweitert, daß die n-besten Satzhypothesen (n = 10-100) berechnet werden.[8]. Daneben wurden auch kontextfreie Grammatiken als Alternative zu den regulären Grammatiken (Bigramm-Modell, endliches Netzwerk) untersucht [9].
- In [10] wurde gezeigt, daß das Suchverfahren auch ein Vokabular von 10.000 Wörtern handhaben kann.
- In [11] wurde der Ansatz auf die sprecherunabhängige 'DARPA-Resource-Management'-Aufgabe erweitert. Die erzielten Erkennungsergebnisse sind vergleichbar mit den Ergebnissen, die innerhalb des DARPA-Projektes erreicht werden.

Diese Arbeit ist Teil eines gemeinsamen Siemens-Philips-IPO-Projektes und wurde mit Mitteln des Bundesministers für Forschung und Technologie (Förderungskennzeichen ITM 8801) gefördert. Die Verantwortung für den Inhalt liegt jedoch allein bei dem Autor.

6. Referenzen

[1] DRECKSCHMIDT, G.: The Linguistic Component in the Speech Understanding System SPICOS. Linguistische Datenverarbeitung 9, Hrsg. Tillmann, H.G., Willee, G., Analyse und Synthese gesprochener Sprache. Jahrestagung 1987 der Gesellschaft für Linguistische Datenverarbeitung e.V. Bonn, pp. 96-101, März 1987

[2] van DEEMTER, C.J., BROCKHOFF, G., BUNT, H.C., MEYA, M., de VET, J.H.M.: From TENDUM to SPICOS, in IPO Annual Progress Report 20, Eindhoven, pp. 83-90,1985

[3] van HEMERT, J.P., ADRIAENS-PORZIG, U. ADRIAENS, L.M.H.: Speech Synthesis in the SPICOS-Project, wie [1], pp. 34–39

[4] NEY, H., NOLL, A.: Phoneme Modelling Using Continuous Mixture Densities. 1988 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, New York, NY, pp. 437-440, April 1988

[5] NEY, H., PAESELER, A. : Phoneme-Based Continuous Speech Recognition Results for Different Language Models in the 1.000-Word SPICOS System. Speech Communication, Sonderheft 'Speech Recognition', Vol. 7, pp. 367-373, Dez. 1988

[6] MERGEL, D., PAESELER, A.: Construction of Language Models for Spoken Data Base Queries. 1987 IEEE International Conference on Acoustics, Speech, and Signal Processing, Dallas , TX, pp. 844-847, April 1987

[7] NEY, H. , MERGEL, D., NOLL, A., PAESELER, P.: A Data-Driven Organization of Dynamic Programming Beam Search for Continuous Speech Recognition. 1987 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Dallas, TX, pp. 20.9.1-4, April 1987

[8] STEINBISS, V.: Sentence-Hypotheses Generation in a Continuous-Speech Recognition System. European Conf. on Speech Communication and Technology, Paris, pp. 51-54, Sep. 1989

[9] NEY, H.: Dynamic Programming Speech Recognition Using a Context Free Grammar. 1987 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Dallas, TX, pp.3.2.1.-4, April 1987

[10] STEINBISS, V., NOLL, A., PAESELER, A., NEY, H. et al.: A 10.000-Word Continuous Speech Recognition System. 1990 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Albuquerque, NM, pp. 57-60, April 1990

[11] NEY, H.: Experiments on Mixture Density Phoneme-Modelling for the Speaker-Independent 1.000-Word Speech Recognition DARPA Task. 1990 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Albuquerque, NW, pp. 713-716, April 1990