Spektrum der Wissenschaft Germany March 1994

Maschinen, die aufs Wort gehorchen, sprechende Computer und automatische Simultandolmetscher – Ideen, die bisher zur Science-fiction gehörten, nähern sich nun der Realität. Dafür konzipierte Systeme sind zwar noch weit davon entfernt, einen gesprochenen Satz so zu verstehen wie ein Mensch. Erstaunlicherweise reicht jedoch ein Minimum an Verständnis für beschränkte Anwendungen bereits aus. Die folgenden Beiträge beschreiben Prinzipien der maschinellen Spracherkennung und konkrete Realisierungen für Einzelworterkennungssysteme, automatische Diktiergeräte, telephonische Auskunftssysteme und Übersetzungshilfsgeräte.

Prinzipien, Stand der Technik, sprecherabhängige Einzelworterkennung

Von Klaus Fellbaum

Die sprachliche Kommunikation ist ein sehr vielschichtiger Prozeß. Ein Mensch versteht eine Sprachäußerung nicht nur anhand dessen, was er hört; er setzt vielmehr seine gesamte Spracherfahrung sowie sein Vorwissen über Gesprächsgegenstand und -partner ein; zudem wertet er auch nichtverbale Komponenten wie Gestik, Mimik und die emotionale Klangfärbung der Stimme mit aus. Unter Umständen ist diese Zusatzinformation sogar wichtiger als der Wortlaut. Die natürliche Sprache ist schließlich durch ein hohes Maß an Redundanz (also an eigentlich Überflüssigem) gekennzeichnet, so daß oftmals schon verbale Andeutungen oder Sprachfetzen für eine Verständigung ausreichen. Dadurch erklärt sich, daß ein Gespräch auch in lärmerfüllter Umgebung möglich ist.

Wollte man diese phänomenale Erkennungsleistung durch ein technisches System realisieren, müßte dieses letztlich über das Wissen, die Erfahrungen und die Intelligenz eines Menschen verfügen. Man kann lange darüber philosophieren, ob das ein erreichbares oder sinnvolles Ziel ist. Für den Entwickler eines Spracherkennungssystems, der immer auch den technischen Aufwand (und damit die Kosten) berücksichtigen muß, ist dies sicherlich nicht der Fall; er muß in erster Linie die konkrete Anwendung sehen. Dabei zeigt sich, daß vielfach sehr eingeschränkte Formen der Spracherkennung genügen; es kommt entscheidend darauf an, die für die spezielle Anwendung geeignetste Lösung zu finden.

Kommunikation mit Maschinen über natürliche Sprache – statt wie üblich über Tastatur und Bildschirm, allgemeiner über Schalter und Anzeigegeräte – bietet eine Reihe von Vorteilen: Der Benutzer muß keine neue Technik erlernen, sondern arbeitet mit der ihm vertrautesten Kommunikationsform; er behält Augen und Hände für andere Tätigkeiten frei, ist nicht an einen bestimmten Platz gebunden und kann die Maschine sogar per Telephon fernsteuern. Sprachein- und -ausgaben sind auch in dunklen, schmutzigen und staubigen Räumen möglich und für Behinderte, die keine Tastatur bedienen können, vielleicht die einzige Kommunikationsmöglichkeit mit der Maschine. In der Gegenrichtung erreicht eine sprachliche Äußerung der Maschine auch den abgelenkten oder unaufmerksamen Benutzer.

Gegen die Verwendung dieser Kommunikationsform spricht, daß Leistungsfähigkeit und Zuverlässigkeit für manche Anwendungen noch unbefriedigend sind. Hinzu kommen die Schwächen jeder sprachlichen Kommunikation: Lärmbelästigung Unbeteiligter, unerwünschte

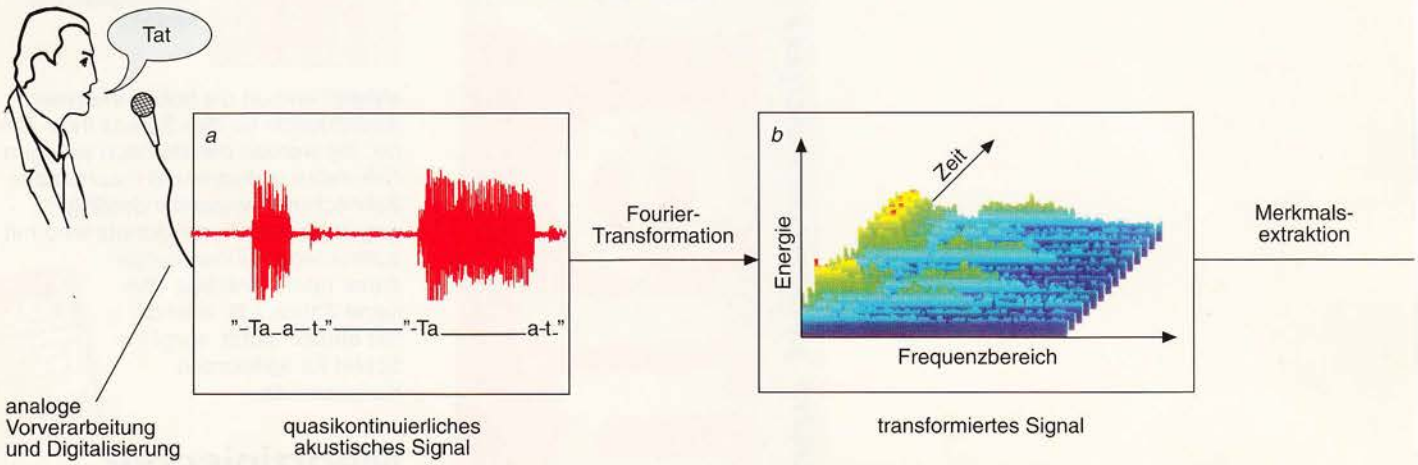


Bild 1: Schema eines Einzelworterkennungssystems. Das Schallsignal (*a*) wird zunächst vorbehandelt; das System bestimmt dann für diskrete Zeitabschnitte die Energien gewisser Frequenzbereiche im Schallsignal (*b*; die Farben von Blau über Grün und Gelb

Mithörmöglichkeiten und Beeinträchtigung durch Störgeräusche, was für die Spracherkennungssysteme gegenwärtig noch ein großes Problem ist.

Die wichtigsten Anwendungen finden sich auf folgenden Feldern:

– *Eingabe von Zahlen und Wortlisten*: Wer lange Zahlen- oder Wortkolonnen von einer Vorlage abzutippen hat, muß immer wieder den Blick zwischen Vorlage, Tastatur und Bildschirm wechseln, was auf die Dauer lästig, ermüdend und eine Fehlerquelle ist. Unmittelbare sprachliche Eingabe ohne Blickabwendung vermeidet dieses Problem. Zum Korrekturlesen kann man Sprachausgabe durch die Maschine einsetzen.

Bereits heute übermitteln manche Qualitätskontrolleure in der Autoproduktion ihre Mängelmeldungen direkt über eine Funkverbindung an einen spracherkennenden Computer, der nicht nur ein Protokoll führt, sondern auch eine entsprechende Meldung an die verursachende Stelle – möglicherweise einen Fertigungscomputer – weitergibt. Der Vorteil gegenüber einer späteren schriftlichen Auswertung ist prompte Reaktion; dadurch sinkt die Ausschußquote.

– *Steuerung von Maschinen und Computern*. Das System erkennt einen eingegebenen Befehl nicht nur, sondern führt ihn auch aus, indem es eine Handlung auslöst. So sind bereits sprachgesteuerte Werkzeugmaschinen auf dem Markt. Es gibt Automobile, in denen Fensteröffner, Scheibenwischer, Radio und Telephon gesprochenen Befehlen folgen. Bestimmte Kontroll- und Korrekturfunktionen in Personal Computern sind auch mit Hilfe von Spracherkennern auslösbar. Für motorisch Schwerbehinderte gibt es sprachgesteuerte Rollstühle. Für sicherheitsrelevante Funktionen wird die Sprachsteuerung bislang nicht eingesetzt, weil die Frage der Haftung bei Systemversagen noch ungeklärt ist.

– *Auskunfts- und Bestellsysteme*. Der Benutzer äußert bestimmte Anforderungen oder Wünsche, und das System gibt eine gesprochene Antwort (siehe den Beitrag von Helmut Mangold auf Seite 97).

– *Diktiersysteme*. Ein beliebiger, fließend gesprochener Text soll in Schrifttext umgesetzt werden. Diese Anwendung erfordert die weitestgehende Form der Spracherkennung (siehe die Beiträge von Marcus Spies auf Seite 90 und Volker Steinbiß auf Seite 94).

Man pflegt Spracherkennungsverfahren in drei Klassen einzuteilen: Erkennung von Einzelwörtern, von Schlüsselwörtern in fließendem Text oder von kontinuierlicher Sprache. In der genannten Reihenfolge steigen Schwierigkeiten und Aufwand drastisch an. Ein weiteres Kriterium ist die Sprecherabhängigkeit. Die üblichen Systeme müssen vor der eigentlichen Nutzung erst an den jeweiligen Sprecher angepaßt werden. Eine Sprecherunabhängigkeit kann man dadurch erreichen, daß man das System vorab mit möglichst vielen Sprechern trainiert. Der Aufwand dafür ist beträchtlich; gleichwohl nimmt die Erkennungssicherheit in der Regel ab.

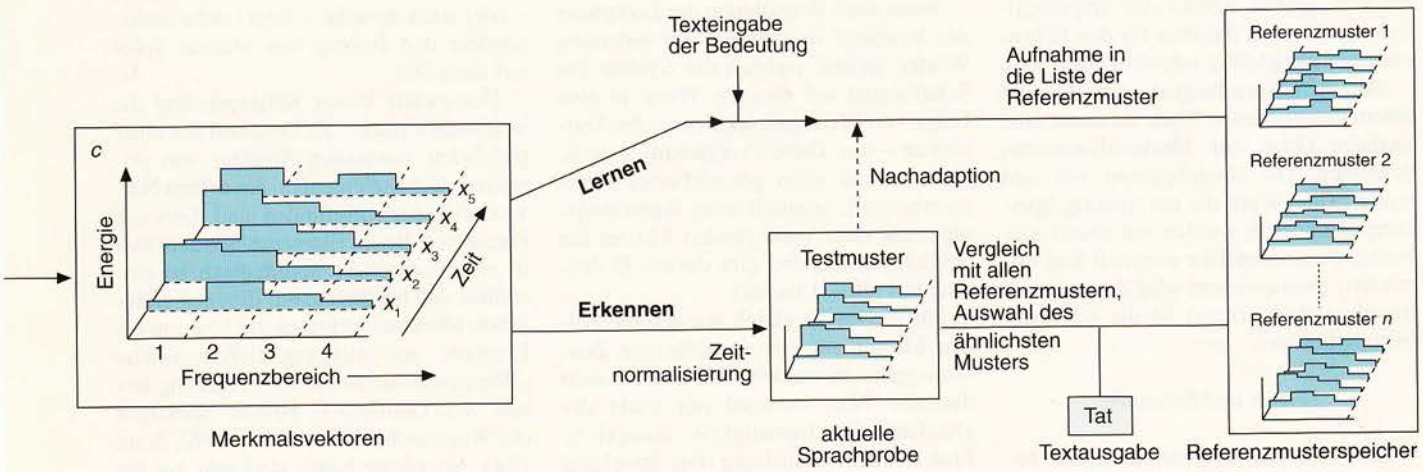
Im folgenden sei die sprecherabhängige Einzelworterkennung genauer dargestellt. Sie ist die derzeit am meisten verwendete und technisch ausgereifteste Form der Spracherkennung.

Der Prozeß besteht im wesentlichen aus zwei Stufen (Bild 1): einer Vorverarbeitung, die aus dem Sprachsignal die für die Erkennung relevanten Parameter extrahiert, und der Klassifikation, die durch Mustervergleiche zwischen Test- und Referenzmustern die zugehörige Bedeutung findet.

Vorverarbeitung

Zur ersten Stufe gehören zunächst die frequenzmäßige Begrenzung (Filterung), die Lautstärkenormierung und eine Analog-Digital-Umsetzung; letztere ist für die (heute ausschließlich digitale) Weiterverarbeitung erforderlich. Das so digitalisierte Sprachsignal besteht zwar bereits aus einer diskreten Folge von Zahlen, die jeweils die Schallenergie zu einem gewissen Zeitpunkt beschreiben. Diese Abtastzeitpunkte liegen jedoch so dicht, daß wesentliche Eigenschaften des ursprünglich kontinuierlichen Signals erhalten bleiben; man spricht von einem quasikontinuierlichen Signal.

Die in dieser Form noch viel zu große Menge an Daten ist nun so zu reduzieren, daß die von überflüssigem Ballast befreite Information die relevanten Eigenschaften des Sprachsignals möglichst präzise charakterisiert. Unter den zahlreichen Möglichkeiten für diesen Schritt betrachten wir im folgenden aus Gründen der Anschaulichkeit die Parameterextraktion aus dem Sprachspektrum. Als



bis Rot stehen für zunehmende Signalenergie bei der zugehörigen Frequenz). Aus diesen Rohdaten extrahiert das System gewisse charakteristische Merkmale (*c*). Lernen und Wiedererkennen von Wörtern finden auf der Basis dieser Merkmalsvektoren statt.

Beispiel diene das Wort *Tat*, einmal kurz und einmal lang gesprochen. Im Zeitsignal (Bild 1 *a*) ist der Plosivlaut *t* an seiner niedrigen Signalenergie und einem regellosen Verlauf zu erkennen; letzterer deutet auf hochfrequente Signalanteile hin. Der Vokal *a* ist durch hohe Signalenergie und den ziemlich regelmäßigen, periodischen Verlauf charakterisiert.

Außerdem erkennt man, daß langsames Sprechen zwar den Vokal verlängert, nicht aber die Plosive. Ein langsam gesprochenes Wort ist also nicht einfach eine Zeitlupenversion eines schnell gesprochenen, was bei der Verarbeitung zu berücksichtigen ist.

Das Schallsignal wird nun nach Frequenzen zerlegt; das entspricht mathematisch einer Fourier-Transformation, angewandt auf Zeitintervalle von etwa 20 bis 30 Millisekunden. In der Frequenzdarstellung (Bild 1 *b*) finden sich wie vorher bei den Vokalen hohe Energien, die im Bereich tiefer Frequenzen (um 1000 Hertz) konzentriert sind; derartige Energiemaxima bezeichnet man als Formanten. Dagegen ist der Frequenzbereich bei den Plosiven sehr breit und reicht bis etwa 10 Kilohertz. Vor dem *t* am Wortende ist nahezu keine Energie vorhanden; solche Pausen charakterisieren die Druckaufbauphasen, die allgemein für das Sprechen von Plosiven erforderlich sind.

Bereits durch die Fourier-Transformation ist die Zeitskala auf das Raster der genannten Intervalle vergröbert worden. Meist treibt man die Datenreduktion noch weiter, indem man auch auf der quasikontinuierlichen Frequenzskala zahlreiche Werte durch Mittelung über 8 bis 20 relativ breite Frequenzbänder zusammenfaßt. Die verbleibenden Zahlen bilden, ihrerseits durch Rundung vergröbert, den sogenannten Merkmalsvektor: ein sehr grobes Abbild des ursprünglichen Spektrums, das aber für den Erkennungsvorgang völlig ausreicht (Bild 1 *c*).

Aus dem Sprachsignal, zum Beispiel einem gesprochenen Wort, ist damit eine zeitliche Folge von Merkmalsvektoren geworden. Die Sprechpausen vor und hinter jedem Wort, die nur unnötig Speicherplatz kosten, werden mit einem speziellen Pausendetektor ermittelt und eliminiert; abgespeichert wird dann nur das jeweilige Wort. Damit ist die Vorverarbeitung beendet.

Lernen und Erkennen

Um nun das System an einen bestimmten Sprecher anzupassen, spricht dieser ein Wort, das in der beschriebenen Weise in eine Folge von Merkmalsvektoren überführt wird, die fortan ein Referenzmuster bildet. Die Bedeutung des gesprochenen Wortes wird dem System üblicherweise über eine Tastatur mitgeteilt. Bedeutung und zugehöriges Sprachsignal werden nun im Referenzmusterspeicher abgelegt. Mehrere Referenzmuster für dasselbe Wort können durch Mittelung zu einem einzigen zusammengefaßt werden, was zufällige Abweichungen bei der Sprachaufnahme kompensiert. In gleicher Weise verfährt man mit allen Wörtern, die das System lernen soll.

Wenn nach Beendigung der Lernphase der Benutzer irgendeines der gelernten Wörter spricht, wandelt das System das Schallsignal auf dieselbe Weise in eine Folge von Merkmalsvektoren – das Testmuster – um. Dieses vergleicht es nacheinander mit allen gespeicherten Referenzmustern, ermittelt unter ihnen dasjenige, das dem vorliegenden Muster am ähnlichsten ist, und gibt dessen Bedeutung aus (Bild 1 rechts).

Für diesen Vergleich müssen die beiden Muster auf eine gemeinsame Zeitskala gebracht werden, weil kein Mensch dasselbe Wort zweimal mit exakt der gleichen Geschwindigkeit ausspricht. Eine schlichte Stauchung oder Streckung der Zeitachse wäre sinnlos, da sich – wie erwähnt – eine Veränderung der Sprechgeschwindigkeit auf verschiedene Laute unterschiedlich auswirkt. Als sehr wirkungsvolles Verfahren hat sich indessen die sogenannte dynamische Zeitanpassung erwiesen: Jeder kleine Zeitabschnitt wird individuell so gedehnt oder gestaucht, daß die Übereinstimmung zwischen Test- und Referenzmuster möglichst groß wird (Bild 2). Die sich dadurch ergebende nichtlineare Anpassungsfunktion wird im Englischen *dynamic time warping function* genannt; das Wort *warp* (sich winden) beschreibt sehr anschaulich, wie sich die Anpassungsfunktion durch das Koordinatensystem windet.

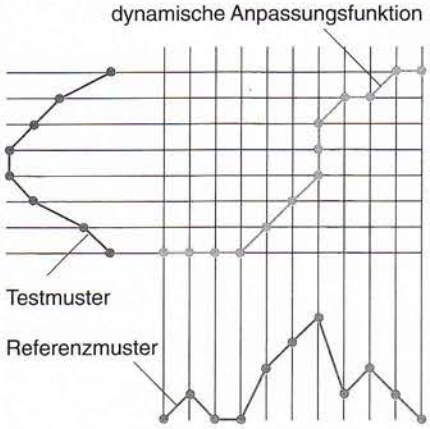


Bild 2: Vergleich zweier Muster mit unterschiedlichen Zeitskalen. Der spitze Gipfel des Referenz- und der flache des Testmusters (entsprechend etwa einem kurz beziehungsweise lang ausgesprochenen Vokal) wirken zunächst nicht sehr ähnlich. Erst eine Zeitverzerrung durch eine dynamische Anpassungsfunktion macht den spitzen Gipfel künstlich flach, indem sie die Zeit an dieser Stelle beliebig langsam vergehen läßt (vertikaler Abschnitt der Anpassungskurve). Über den im Testmuster nicht vertretenen Vorhügel des Referenzmusters huscht die Zeit beliebig schnell hinweg (horizontaler Abschnitt). Dieser Kunstgriff macht die Muster vergleichbar.

Ein letzter Verarbeitungsschritt ist die Nachadaption. Da sich die Stimme eines Sprechers im Laufe der Zeit verändert, benutzt man die Testmuster, die sicher erkannt worden sind, zum Auffrischen der abgespeicherten Referenzmuster, indem man über beide einen geeignet gewichteten Mittelwert bildet. Dadurch arbeitet das System nicht nur mit den während der Lernphase eingespeicherten Wörtern, sondern auch mit kürzlich gesprochenen.

Weitere Verfahren

Zwei andere, sehr erfolgreiche Verfahren arbeiten anstelle des hier dargestellten direkten Vergleichs von Test- und Referenzmustern mit einem eher indirekten Mustervergleich.

Als klarer Favorit gilt zur Zeit die Erkennung mit den sogenannten Hidden-Markov-Modellen. Diese gehen auf den russischen Mathematiker Andrej Andrejewitsch Markow (1856 bis 1922) zurück. Sie arbeiten mit Schätzungen dafür, mit welcher Wahrscheinlichkeit auf einen Zustand eines Systems (zum Beispiel einen Merkmalsvektor) ein anderer – oder auch derselbe – folgt (siehe insbesondere den Beitrag von Marcus Spies auf Seite 90).

Das zweite dieser Konzepte sind die neuronalen Netze. Sie bestehen aus einer parallelen, vernetzten Struktur von primitiven Schaltelementen, die echten Nervenzellen nachempfunden sind. Gewisse Parameter dieser Elemente können sich in einer Lernphase automatisch so einstellen, daß bestimmte am Eingang anliegende Merkmalsvektoren ein bestimmtes Ergebnis am Ausgang liefern (siehe „Wie neuronale Netze aus Erfahrung lernen" von Geoffrey E. Hinton, Spektrum der Wissenschaft, November 1992, Seite 134). Neuronale Netze sind sehr gut für die Spracherkennung geeignet; sie erweisen sich vor allem dann als besonders

erfolgreich, wenn die Testmuster durch Störungen (etwa Umgebungsgeräusche) verfälscht sind.

Welche der genannten Strategien sich langfristig durchsetzen wird, ist noch unklar. So könnten die Hidden-Markov-Modelle ohne weiteres durch neuartige, für die Spracherkennung optimierte neuronale Netze überholt werden.

Um die Leistungsfähigkeit von Spracherkennern weiter zu verbessern, wertet man außer der akustisch-phonetischen Information, die in den Merkmalsvektoren steckt, noch weitere Informationsquellen aus. Eine sehr wichtige ist das aufgabenbezogene Wissen. Bei den meisten Anwendungen ist der Einsatzbereich inhaltlich begrenzt, so daß es nur relativ wenige zulässige Wörter gibt. Deshalb könnte zum Beispiel ein Spracherkennungsprogramm zur Maschinensteuerung einen Befehl, der fehlerhaft als „Maschine Wald" erkannt wurde, problemlos in den gültigen Befehl „Maschine halt" korrigieren.

Eng damit verknüpft ist das pragmatische Wissen: Das Spracherkennungsprogamm erhält Informationen über den Zustand seiner Umgebung und registriert eine phonetisch basierte Erkennung als falsch, wenn sie im Widerspruch zu den Umgebungsbedingungen steht. Wenn etwa das Erkennungssystem zur Maschinensteuerung (durch Meldung von Meßfühlern) weiß, daß die Maschine läuft, würde es den Befehl „Maschine anschalten" als sinnlos erkennen und durch „Maschine anhalten" ersetzen – oder eine Rückfrage auslösen.

Stand der Technik

Sprecherabhängige Einzelworterkenner für einen kleinen Wortschatz (bis zu mehreren hundert Wörtern) lassen sich heute problemlos realisieren. Die meisten der gegenwärtig verfügbaren Systeme sind von diesem Typ.

Vereinzelt werden schon sprecherunabhängige Einzelworterkenner mit bis zu 50 Wörtern angeboten; viele befinden sich aber noch im Forschungs- oder Entwicklungsstadium. Für diese Systeme besteht dringender Bedarf im Telekommunikationsbereich mit seinen immer neuen Benutzern, von denen man naheliegenderweise nicht jedesmal vor einer Benutzung eine Trainingsphase verlangen kann. Eine typische Anwendung sind telephonische Auskunftssysteme.

An der Spitze der Entwicklung liegen wenige Erkenner wie das auf Hidden-Markov-Modellen basierende System „Dictate-30K" der amerikanischen Firma Dragon Systems mit einer Kapazität bis zu 30 000 Wörtern. Das genügt im allgemeinen für das Erstellen üblicher Texte, auch wenn man berücksichtigt, daß ein Wort meist verschiedene Beugungsformen hat und jede Form als eigenständiges Wort zählt. Indem sich das System ohne eine Trainingsphase an ihm unbekannte Sprecher adaptiert, hat es fast die Eigenschaften eines sprecherunabhängigen Erkenners. Die Hardware ist auf einer Karte untergebracht, die in einen Personal Computer eingesteckt werden kann. Die Hauptanwendung liegt im Bürobereich. Ein Nachteil ist sicherlich, daß der Benutzer zwischen je zwei Wörtern eine Pause machen muß (weil es sich eben um einen Einzelwort-Erkenner handelt), was eine abgehackte, unnatürliche Sprechweise erfordert.

Eine besonders interessante Aufgabenstellung ist die Erkennung von Schlüsselwörtern in fließend gesprochener Sprache, das sogenannte *word spotting*. Es bildet den Übergang zur kontinuierlichen Spracherkennung, erfordert aber bei weitem nicht deren Rechenaufwand. Seine Stärken entfaltet es da, wo es nur auf spezielle Informationen – etwa Kommandos, Anfragen, Namen oder Zahlen – ankommt. Beispielsweise hat der Benutzer eines mit *word spotting* arbeitenden Flug-Auskunftssystems beträchtliche Freiheiten, seine Antwort auf die Frage, wohin er fliegen möchte, zu formulieren; das System wird ihn schon dann richtig verstehen, wenn es nur das Wort „Hamburg" im gesprochenen Text korrekt erkennt. Auf diese Weise läßt sich hohe Benutzerakzeptanz erreichen.

Erstaunlicherweise gibt es erst wenige derartige Systeme. Ein Grund mag darin liegen, daß sich *word spotting* in besonderem Maße zum Abhören von Sprachkanälen (insbesondere Telephonleitungen) eignet und die Untersuchungen deshalb der Geheimhaltung unterliegen.

Die komfortabelste, aber auch bei weitem schwierigste Technik ist die zum Erkennen kontinuierlicher Sprache. Die Probleme entstehen vor allem dadurch, daß Wortgrenzen im Sprachfluß häufig nicht erkennbar sind oder gar nicht existieren: „Am Montag" wird gesprochen „amontag". Das macht den Vergleich auf Basis von Wörtern unmöglich, so daß man zu Einzellauten übergehen muß.

Es gibt weltweit erst sehr wenige Systeme, die diese Probleme bewältigen; die meisten befinden sich noch im Labor- oder Prototypen-Stadium. Außer dem „Speech Processing System 6000" der Firma Philips (vergleiche den Beitrag von Volker Steinbiß auf Seite 94) ist vor

allem das System „Sphinx" zu erwähnen, das Kai-Fu Lee und seine Mitarbeiter an der Carnegie-Mellon-Universität in Pittsburgh (Pennsylvania) entwickelt haben. Es basiert im wesentlichen auf Hidden-Markov-Modellen und enthält keine wesentlich neuen Strategien oder Komponenten; seine hohe Leistungsfähigkeit kommt vielmehr dadurch zustande, daß die besten der bekannten Erkennungsalgorithmen in aufwendiger Weise miteinander vereint wurden. „Sphinx" kann kontinuierliche Sprache mit einem Vokabular von rund 1000 Wörtern und einer Treffsicherheit von ungefähr 95 Prozent erkennen. Das System befindet sich derzeit noch im Laborstadium.

Die automatische Spracherkennung gehört zweifellos zu den wichtigsten technischen Innovationen im Bereich der Mensch-Maschine-Kommunikation. Die verfügbaren Systeme sind zwar noch weit von einer Erkennungsleistung entfernt, die der des Menschen vergleichbar wäre, können aber schon jetzt für vielfältige Aufgaben eingesetzt werden. Die meisten technischen Anwendungen haben ohnehin nur sehr eingeschränkte Anforderungen an ein Spracherkennungssystem. Erheblicher Forschungs- und Handlungsbedarf besteht aber noch auf einem nicht-technischen Gebiet: der optimalen Gestaltung des Mensch-Maschine-Dialogs.

Prof. Dr.-Ing. Fellbaum befaßt sich am Institut für Fernmeldetechnik der Technischen Universität Berlin in Forschung und Lehre mit der elektronischen Sprachsignalverarbeitung und mit nachrichtentechnischen Systemen. Sein besonderes Interesse gilt der Entwicklung von Sprachverarbeitungssystemen für Blinde und motorisch Behinderte.

Literaturhinweise

Sprachverarbeitung und Sprachübertragung. Von Klaus Fellbaum. Springer, Heidelberg 1984.

Automatische Spracherkennung. Von G. Ruske. Oldenbourg, München 1988.

Sprachliche Mensch-Maschine-Kommunikation. Herausgegeben von Helmut Mangold. Oldenbourg, München 1992.

Speech Recognition and Understanding. Recent Advances, Trends and Applications. Herausgegeben von P. Laface und R. de Mori. Springer, Heidelberg 1992.

Advances in Speech Signal Processing. Herausgegeben von Sadaoki Furui und M. Mohan Sondhi. Marcel Dekker, New York/Basel/Hong Kong 1992.

Grundzüge der Spracherkennung in einem Diktiersystem

Von Marcus Spies

Das Spracherkennungssystem IBM Speech Server Series (ISSS) setzt gesprochenen Text in Echtzeit und mit extrem hoher Erkennungsgenauigkeit in geschriebenen um und stellt ihn auf dem Bildschirm dar (Bild 1). Es ist in den letzten Jahren in Produktlabors der IBM in Boca Raton (Florida) und Wien sowie in den Wissenschaftlichen Zentren in Rom, Paris, Sevilla, Hursley (Großbritannien) und Heidelberg entwickelt worden. Eine Forschergruppe um Frederic Jelinek, Robert Mercer und Lalit Bahl am Thomas-J.-Watson-Forschungslabor der IBM in Yorktown Heights (New York) hatte durch Grundlagenforschung die wesentlichen Voraussetzungen für diese Entwicklung geschaffen.

Derzeit setzt die Benutzung von ISSS das sogenannte diskrete Sprechen, das heißt Diktieren mit (wenn auch nahezu beliebig kurzen) Pausen zwischen den Wörtern voraus. Diese Einschränkung wurde um der Erkennungsgenauigkeit willen beibehalten; sie ist nicht aus Systemgründen nötig.

Die Erkennung eines Sprachsignals beginnt mit der Vorverarbeitung: Entsprechend der Schallverarbeitung im menschlichen Ohr berechnet das System zunächst über eine Fourier-Transformation, wie intensiv bestimmte feste Frequenzen zu dem jeweiligen Zeitpunkt – genauer: innerhalb eines etwas längeren vorangegangenen Zeitintervalls – im Schallsignal vorhanden sind. Jede Hundertstelsekunde wird diese Information abgegriffen und als sogenanter Merkmalsvektor an die nächste Verarbeitungsstufe weitergereicht. Es hat sich als sinnvoll erwiesen, die Werte benachbarter Zeitpunkte zusammenzufassen und die für die Unterscheidung wichtige Information mit einem statistischen Verfahren, der sogenannten Diskriminanzanalyse, zu extrahieren. Die so gewonnenen verfeinerten Merkmalsvektoren enthalten also die wesentliche Information bereits in verdichteter Form.

Wenn wir einen bestimmten Laut – etwa ein *a* – artikulieren, werden die währenddessen erfaßten Merkmalsvektoren einander ähnlicher sein als einem Merkmalsvektor zu einem anderen Laut. Stellt man sie sich – wie bei Vektoren üblich – als Punkte in einem hochdimensionalen Raum vor, so bilden die zu einem bestimmten Laut gehörenden Merkmalsvektoren beziehungsweise Punkte eine Wolke in diesem Raum. Deren genaue Lage und Form ist für jeden Sprecher unterschiedlich; sie wird in einer sprecherspezifischen Trainingsphase des Systems ermittelt.

Aus Gründen der Rechenökonomie beschreibt man die Wolken angenähert mit Hilfe von rechentechnisch besonders einfachen Standardformen. Man unterstellt gewissermaßen, daß die Wolken kugelförmig oder elliptisch sind und ihre Hauptachsen parallel zu den Achsen des Koordinatensystems liegen (Normalverteilungen mit diagonaler Kovarianzmatrix). Eine Wolke, die stark von der Standardform abweicht, läßt sich durch mehrere Standardwolken annähern.

Aus einer Sprachäußerung wird durch die Vorverarbeitung eine zeitliche Abfolge von Merkmalsvektoren; in dem abstrakten Raum hüpft gleichsam ein Punkt von Wolke zu Wolke. Im Prinzip müßte ein Spracherkennungsprogramm also nur anhand der Merkmalsvektoren die jeweils richtige Wolke identifizieren. Die derart gefundene Abfolge der Wolken ergäbe dann direkt eine Lautschrift des Sprechsignals.

In der Praxis ist die Situation allerdings weitaus komplizierter. Man findet typischerweise nach einem Training erheblich mehr Wolken im Raum der Merkmalsvektoren, als es Laute gibt. Es existiert also keine eindeutige Zuordnung von Lauten (Phonemen) zu Wolken in unserem Merkmalsraum. Vielmehr benutzt jedes Phonem Punkte aus mehreren Wolken; dabei liegt nur grob deren Reihenfolge fest, nicht aber der genaue zeitliche Verlauf. So machen die meisten Sprecher bei einem einzeln ausgesprochenen *w* am Ende einen Abstecher zur Wolke für das *e*; im Verlauf des *w* wird möglicherweise zwischendurch kurz die

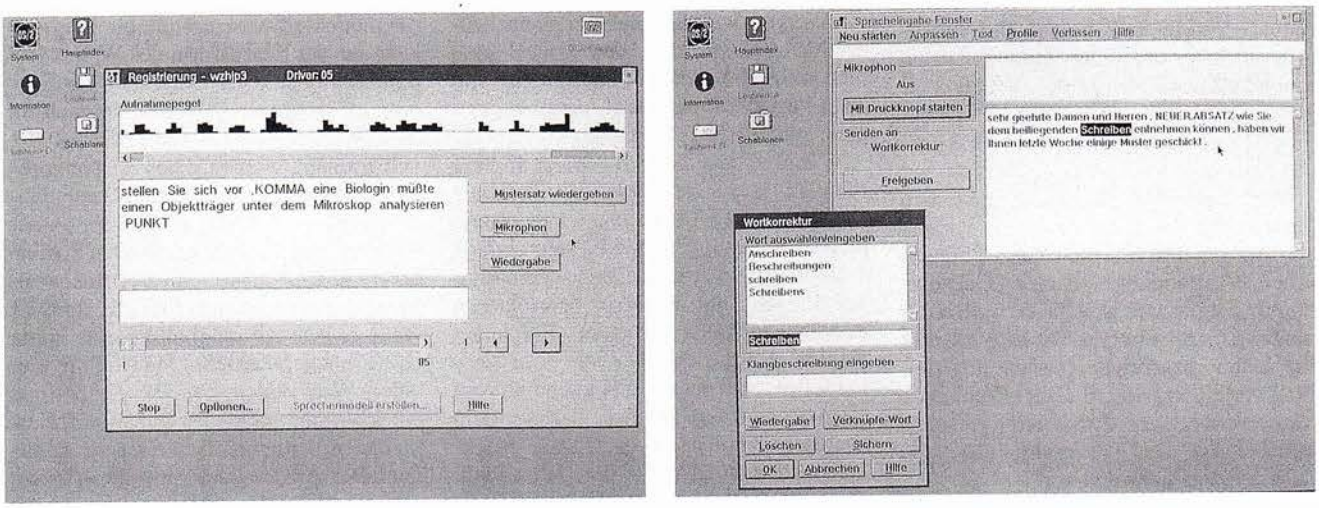


Bild 1: Das automatische Diktiersystem ISSS schreibt gesprochenen Text in Echtzeit auf den Bildschirm. Es erkennt gesprochene Satzzeichen und Kommandos wie „Absatz" und stellt sie vorläufig in Großbuchstaben dar (links). Im Korrekturmodus präsentiert das System dem Benutzer seine bevorzugte Hypothese für ein Wort zusammen mit den nächstplazierten (rechts); ein falsch verstandenes Wort ist durch Anwahl des richtigen (das wahrscheinlich unter den verworfenen ist) bequem zu korrigieren.

u-Wolke benutzt, und so weiter. Schließlich sind die Wolken unscharf begrenzt und überlappen sich gegenseitig.

Markow-Ketten

Gleichwohl ist das Problem nicht unlösbar, wie unsere eigene Fähigkeit zum Sprachverstehen zeigt. Um das menschliche Vorwissen mathematisch zu formalisieren und damit dem Computer verfügbar zu machen, ordnet man einer sprachlichen Äußerung (beispielsweise einem Wort) einen sogenannten Markow-Prozeß zu. Das ist zunächst die Angabe der Wahrscheinlichkeiten, mit denen auf einen Zustand (Merkmalsvektor) aus einer gewissen Menge von erlaubten Zuständen ein anderer folgt. Man stellt das üblicherweise durch eine sogenannte Markow-Kette dar (Bild 2). Eine Realisierung eines Markow-Prozesses besteht darin, daß ausgehend von einem Anfangszustand ein Folgezustand durch Zufall nach Maßgabe der für den Anfangszustand gültigen Übergangswahrscheinlichkeiten bestimmt wird, aus diesem wieder ein Folgezustand und so weiter.

Markow-Ketten sind ein Standardmittel der Statistik zur Modellierung zeitlicher Abläufe. Das Neuartige in der Anwendung auf die Spracherkennung ist, daß die Zustände nicht Merkmalsvektoren, sondern Wolken – genauer: Wahrscheinlichkeitsverteilungen von Merkmalsvektoren – sind. Da ein Merkmalsvektor mehreren Wolken angehören kann, ist aus einer beobachteten Folge von Merkmalsvektoren nicht ohne weiteres auf die zugehörige Folge der Zustände (Wolken) zu schließen (vergleiche Kasten Seite 92); diese bleibt – zunächst – verborgen, weshalb sich die Bezeichnung *hidden Markov models* eingebürgert hat. Was wie eine unnötige Komplizierung anmutet, ist deswegen so erfolgreich, weil die Hidden-Markov-Modelle mit ihrer eingebauten Unschärfe die natürliche Ungenauigkeit der Artikulation sehr gut wiedergeben können.

Für die Lernphase muß ein Sprecher einen Text von etwa einer Dreiviertelstunde Dauer verlesen. Aus den zahlreichen Realisierungen jedes gesprochenen Lautes berechnet das System sodann Schätzwerte für die Verweil- und Übergangswahrscheinlichkeiten eines zugehörigen Markow-Prozesses. Zugleich werden die Parameter der Wolken geschätzt. In diesen Zahlen ist also gewissermaßen das Wissen gespeichert, wie dieser Sprecher einen Laut auszusprechen pflegt.

Für die große Mehrheit der Wörter, die im Trainingstext nicht vorgekommen sind, gleichwohl aber erkannt werden sollen, ist der zugehörige Markow-Prozeß aus den verfügbaren Daten zu konstruieren. Mit Hilfe eines wissensbasierten Systems, das Klaus Wothke und weitere Computer-Linguisten bei der IBM in Heidelberg geschrieben haben, gewinnt man aus der geschriebenen Form eines Wortes die Abfolge der Phoneme, aus denen sich die gesprochene Form zusammensetzt; zu jedem Phonem gehört eine Markow-Kette, aus deren Verkettung man die Markow-Kette für das ganze Wort erhält.

Soll nun das System im Betrieb den richtigen Laut im Kontext einer Äußerung erkennen, muß es entscheiden, welche unter einer großen Zahl denkbarer Markow-Ketten mit der größten Wahrscheinlichkeit die vorliegende Beobachtung (Merkmalsvektor-Folge) erzeugt hat. Hier wird ein Verfahren verwendet, das der amerikanische Nachrichtentechniker Andrew J. Viterbi 1967 in einem ganz anderen Zusammenhang vorgeschlagen hat (siehe Kasten Seite 92). Erst mit diesen Prinzipien und ihrer Umsetzung in sehr effiziente Algorithmen wurde es möglich, im Computer Phoneme der menschlichen Sprache treffsicher unterscheidbar zu machen.

Indem das System mit Markow-Ketten nicht nur für einzelne Phoneme, sondern auch für Phonemfolgen – insbesondere Wörter – arbeitet, stellt es sich gewissermaßen nicht einfach die Frage, welches unter mehreren in Frage kommenden Phonemen soeben am wahrscheinlichsten gesprochen wurde, sondern welches unter Berücksichtigung des bereits gehörten Wortfragments das wahrscheinlichste ist.

In diesem Sinne gleicht seine Arbeitsweise der Wahrnehmung des Menschen: Wir registrieren nicht unvoreingenommen, sondern versuchen das Wahrgenommene in bereits teilweise vorgefaßte Hypothesen einzupassen. Durch diese Rekonstruktionsleistung (für die das Schlagwort „Analyse durch Synthese" geprägt worden ist) sind wir imstande, Unvollständigkeiten und Schwankungen in der Gestalt der Objekte zu korrigieren.

Dieses Vorgehen läßt sich in der maschinellen Spracherkennung weiter verallgemeinern, indem man phonetische Wortmodelle ihrerseits zu größeren Datenstrukturen zusammenfaßt, aus denen

durch Viterbi-Aufreihung die wahrscheinlichste aus einer großen Anzahl von Worthypothesen zu ermitteln ist.

Verfeinerte Sprachmodelle

Die regelhafte Beziehung zwischen Aussprache und Schreibweise – Voraussetzung für das Erstellen eines phonetischen Wortmodells – ist im Deutschen und im Englischen weitaus komplizierter als etwa im Spanischen. Deshalb hat man die bisher beschriebene Modelldarstellung verfeinert, was zu einer erheblichen Verbesserung der Erkennungsgenauigkeit beiträgt. Das Programm ISSS geht dabei in zwei Schritten vor.

Das bisher beschriebene Prinzip war, Wörter in Phoneme zu zerlegen und dann für jedes Phonem eine Markow-Kette zu formulieren. Ist es nicht sinnvoll, das Phonem als Beschreibungsebene der Spracherkennung überhaupt auszulassen? Würde es demnach nicht genügen, für die phonetische Darstellung eines Wortes einfach eine Markow-Kette durch die Wolken des Merkmalsraums nachzuzeichnen? Dies hätte den Vorteil, daß man die Feinheiten der Aussprache wie etwa Verschleifungen und Verkürzungen ganz genau erfassen könnte.

Während eine Markow-Kette für ein Phonem typischerweise 7 Zustände (Wolken) und 13 Übergänge hat, könnte man recht lange, gleichwohl sehr einfach strukturierte Markow-Ketten formulieren, die sich nur auf Merkmalsvektoren und ihre Abfolge beziehen: Für jeden Knoten gibt es den Übergang zum Knoten selbst, zum nächsten Knoten und zum übernächsten für den Fall, daß der dem nächsten Knoten entsprechende Merkmalsvektor in der Kette der Daten fehlt. Das so erhaltene Wortmodell orientiert sich nur an den unmittelbar beobachteten akustischen Phänomenen und hat deshalb den Namen *fenonic base form* erhalten, zu übersetzen etwa als „phänonische Grundform".

In dieser Form ist das Wortmodell allerdings noch unpraktikabel. Weil jedes Wort unmittelbar durch Merkmalsvektoren modelliert wird, entfällt die Möglichkeit, aus der Schriftform des Wortes Hypothesen über seine akustische Realisierung zu gewinnen. Also müßte auch für jedes Wort ein eigenes phänonisches Modell trainiert werden. Dies ist unökonomisch und nicht für die Praxis geeignet.

An dieser Stelle setzt nun der zweite entscheidende Schritt an, der den Spracherkenner ISSS auszeichnet. Offenbar ist ja ein phänonisches Modell sinnvoll, wenn es sich auf möglichst kleine lautliche Einheiten bezieht. Man müßte also einen Weg finden, aus der Schriftform eines Wortes möglichst genau auf diese kleinsten lautlichen Einheiten zu schließen. Der Ansatz dazu sind die phonetischen Entscheidungsbäume.

Die Idee besteht darin, an die Stelle der traditionellen, eher groben Aussprachegregeln solche zu setzen, die empirisch aus umfangreichem Datenmaterial gewonnen werden. Die Aussprache eines geschriebenen Buchstabens – genauer: einer Buchstabenfolge, die ein Phonem vertritt, wie etwa *sch* – ist vom Kontext

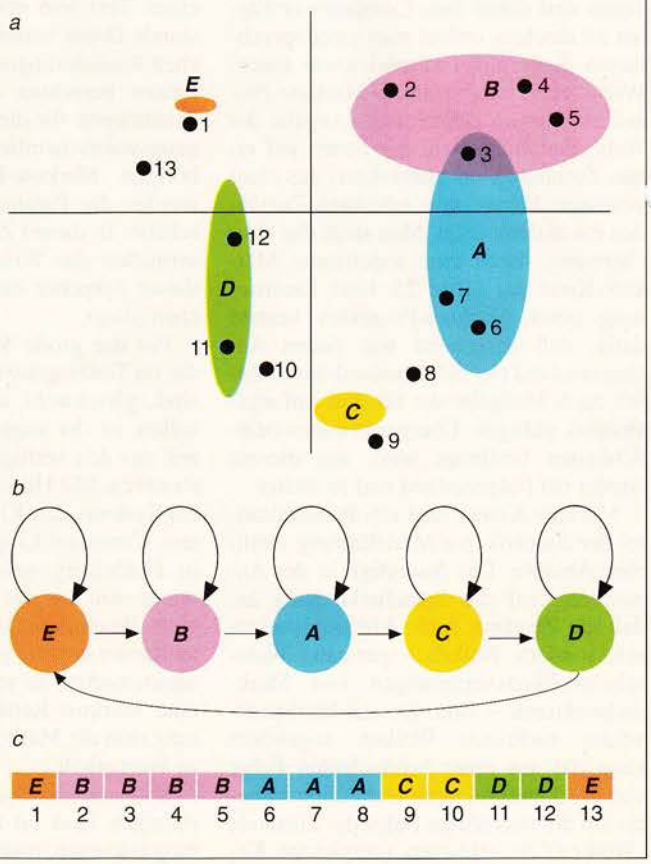
Wie findet ein System die wahrscheinlichste Markow-Kette zu einer Folge von Merkmalsvektoren?

In dem abstrakten, hier zweidimensional dargestellten Raum der Merkmalsvektoren (*a*) sind fünf Wolken eingezeichnet (*A* bis *E*), die für verschiedene Phoneme stehen mögen. Genauer: Je näher ein Merkmalsvektor dem Mittelpunkt einer der Ellipsen liegt, desto sicherer ist es, daß er zu dem entsprechenden Phonem gehört. Auf die Fläche der Ellipse fallen im Mittel 50 Prozent der Realisierungen des entsprechenden Phonems. Wenn also ein Punkt außerhalb einer Ellipse liegt, ist dadurch nicht ausgeschlossen, daß er zu deren Phonem gehört, nur relativ unwahrscheinlich.

Dem Spracherkenner liege die Folge der Merkmalsvektoren 1 bis 13 in der Reihenfolge ihrer Numerierung vor. Es stellt sich die Frage, ob es sich um das Wort mit der in *b* gezeichneten Markow-Kette handelt. Die Pfeile entsprechen den von 0 verschiedenen Verweil- beziehungsweise Übergangswahrscheinlichkeiten. Beispielsweise könnten diese Wahrscheinlichkeiten sämtlich gleich einhalb sein.

Die Viterbi-Aufreihung (*Viterbi alignment*) ordnet nun – unter dieser Voraussetzung – die beobachteten Punkte jeweils der Wolke zu, die unter Berücksichtigung der bisher beobachteten Punktfolge die plausibelste ist (*c*). Unmittelbar leuchtet die Zuordnung der Punkte 1, 2, 4, 5 ein. Punkt 3 wird als *B* erkannt, weil bei einer Zuordnung zu *A* die Punkte 4 und 5 ebenfalls als *A* erkannt werden müßten, obwohl sie unter *B* viel wahrscheinlicher sind. Wenn die Verweilwahrscheinlichkeit bei *C* erheblich größer ist als die Übergangswahrscheinlichkeit von *C* nach *D*, wird Punkt 10 zu *C* geordnet, obwohl er für sich genommen unter *D* wahrscheinlicher wäre.

Würde man die Pfeilrichtungen der Zustandsübergänge umdrehen, ließe sich nur eine sehr wenig wahrscheinliche Beschreibung der beobachteten Abfolge geben. Im Erkennungsprozeß entspricht dann das unter *b* dargestellte Modell einem plausiblen, das Modell mit verdrehten Pfeilen einem sehr wenig plausiblen Wortkandidaten.



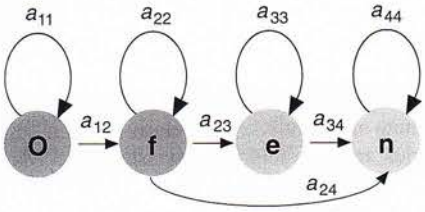


Bild 2: Eine hypothetische Markow-Kette für das Wort *Ofen*. Die farbigen Kreise geben die erlaubten Zustände an, die Pfeile die Übergangswahrscheinlichkeiten zwischen Zuständen. Wenn sich beispielsweise der Prozeß im Zustand 2 (*f*) befindet, wird er mit der Wahrscheinlichkeit a_{22} in diesem Zustand verbleiben, mit der Wahrscheinlichkeit a_{23} in den Folgezustand 3 (*e*) und mit der Wahrscheinlichkeit a_{24} direkt in den Zustand 4 (*n*) übergehen, denn viele Sprecher verschlucken das *e* in *Ofen*.

abhängig; *s* wird in *Amsel* stimmhaft, in *Emsland* stimmlos ausgesprochen. Auch kontextabhängige Verschleifungen können hier erfaßt werden.

Das Programm strukturiert nun die Menge der akustischen Realisierungen des Phonems *s* (oder irgendeines Phonems), indem es die Gesamtmenge abhängig vom Kontext in zwei jeweils möglichst homogene (in sich einheitliche) Teilmengen aufteilt. (Um zu bestimmen, ob zwei Folgen von Merkmalsvektoren sich ähnlich sind, kann man für sehr kurze Abschnitte des Sprachsignals – wie die Erfahrung gezeigt hat – sogar deren Reihenfolge ignorieren; es genügt der wesentlich weniger aufwendige Vergleich ihrer Häufigkeiten.) Die Kriterien dieser Aufteilung muß niemand vorgeben; sie werden automatisch vom Programm errechnet. Dabei werden Kontexte von bis zu fünf Phonemen vor und hinter dem jeweils zu beschreibenden berücksichtigt.

Jede Teilmenge wird wiederum in zwei möglichst homogene Teilmengen aufgeteilt, und so weiter. Dabei gilt das statistische Kriterium des maximalen Informationsgewinns. Mit der Aufteilung fährt man fort, bis sich kein nennenswerter Informationsgewinn mehr ergibt.

Insgesamt erhält man so eine hierarchische (baumartige) Struktur, an deren Enden (den sogenannten Blättern) jeweils eine Menge von Kontexten mit nahezu derselben Aussprache für das Phonem versammelt ist. Eine solche statistisch hergeleitete Phonemaussprache heißt im englischen Jargon *leafeme*, was etwa mit Blattlaut zu übersetzen wäre. Das stimmhafte *s* beispielsweise könnte ein Blattlaut sein, oder auch der Laut, der kurzfristig bei der Verschleifung von *a* und *u* auftritt.

Am Ende dieser Prozedur kann man jedes Wort als eine Folge von Blattlauten modellieren. Und nun wird es sinnvoll, die Blattlaute durch ein phänonisches, also nur die Merkmalsvektoren berücksichtigendes, Modell zu beschreiben. Damit hat man die Ökomonie des phonetischen Wortmodells gewahrt, ist aber in der Feinanpassung an Aussprachedetails und Kontextvarianten von Phonemen beträchtlich weitergekommen. Das Ergebnis ist eine äußerst zufriedenstellende Qualität des Spracherkenners in Echtzeit.

Sprachmodelle

Mit Hilfe der bisher beschriebenen Verfahren gewinnt der Spracherkenner genaugenommen eine Zahl, die angibt, wie wahrscheinlich das vorliegende Sprachsignal ist, wenn ein bestimmtes Wort beziehungsweise eine Wortfolge vorausgesetzt ist. Nennen wir dies die Synthesewahrscheinlichkeit. Eigentlich wollen wir aber umgekehrt wissen, welche Wortfolge die wahrscheinlichste ist, wenn man das Sprachsignal als gegeben ansieht.

Für diesen Umkehrschluß von einer bedingten Wahrscheinlichkeit auf die andere ist das Bayessche Theorem aus der Statistik anzuwenden, das auf den englischen Mathematiker Thomas Bayes (1702 bis 1761) zurückgeht (vergleiche Kapitel 1 und 5 aus meinem Buch „Unsicheres Wissen", Heidelberg 1993). Es besagt, daß bei gegebenem Sprachsignal die Wahrscheinlichkeit für eine Wortfolge proportional zum Produkt aus Synthese- und Grundwahrscheinlichkeit der Wortfolge ist. Wenn etwa die Synthesewahrscheinlichkeit für das Wort „Kant" größer als die für „Hand" ist, das Spracherkennungssystem aber im medizinischen Bereich eingesetzt ist, wo erheblich öfter von Händen als von Philosophen die Rede ist, dann sollte es gleichwohl auf „Hand" schließen.

Wie aber findet man – außerhalb aller akustischen Überlegungen – die Grundwahrscheinlichkeit eines Wortes oder einer Wortfolge? Alle heute gängigen Systeme verwenden nicht etwa wissensbasierte, am Textverstehen orientierte Ansätze, sondern statistische Sprachmodelle. Diese erfassen Wortzusammenhänge anhand beobachteter Wortfolgen.

In unserem Spracherkenner verwenden wir Häufigkeitsbeobachtungen auf Dreiwortfolgen (Trigrammen) in großen Textsammlungen, wie sie auch von Kunden geliefert werden. Daß es hierbei mit dem Auszählen allein nicht getan ist, zeigt sich, wenn man überlegt, daß be-

reits bei 20 000 Wörtern mehr als eine Billion Trigramme denkbar sind. Selbst umfangreiche Textkorpora erreichen selten diese Größe. Man muß also Häufigkeiten nicht beobachteter Trigramme schätzen und auch die beobachteten Häufigkeiten durch Schätzungen korrigieren. Dafür verwenden wir Verfahren, die ursprünglich aus der Biostatistik stammen.

Im Bereich der Sprachmodelle sind bei uns einige Forschungsarbeiten im Gange. So versuchen wir, die Häufigkeitsmodelle für Trigramme durch Einbeziehung der Wortart zu verfeinern. Dadurch wird zum Beispiel das Wissen, daß die Folge Artikel – Substantiv wesentlich häufiger ist als die umgekehrte, für das System verfügbar.

Im Deutschen kommen zahlreiche Wortkomposita (etwa „Gelenkarthrose") in Texten vor, die das Vokabular unnötig belasten, weil ihre Bestandteile („Gelenk" und „Arthrose") meistens als eigenständige Wörter bereits im Vokabular verzeichnet sind. Dafür entwickeln wir zur Zeit einen neuen Ansatz zur Schätzung von Häufigkeiten der Kompositabestandteile.

Andererseits scheint es sinnvoll, die Berücksichtigung eines Kontextes von fester Länge, wie bei den Trigrammen, durch die Beobachtung auch weiter entfernter Wörter zu ergänzen. Gerade im Deutschen kommen weitgespannte Abhängigkeiten sehr oft vor („Er *kam* erst am späten Abend *an*"). Hier scheint es vielversprechend, mit regelähnlichen statistischen Strukturen relevante Kontexte herauszufinden und auf ihnen Sprachmodelle zu formulieren.

Es gibt also noch zahlreiche und praxisrelevante Forschungsaufgaben. Dennoch lohnt es sich schon jetzt, den Einsatz der Spracherkennung am Arbeitsplatz zu erwägen.

Privatdozent Dr. Spies arbeitet in der Forschungsgruppe Spracherkennung im Wissenschaftlichen Zentrum der Firma IBM Informationssysteme GmbH in Heidelberg.

Pausenlos diktieren – kontinuierliche Spracherkennung in der Radiologie

Von Volker Steinbiß

Mit einem Tastendruck beendet der Radiologe das Diktieren des Befundes. Während er sich dem nächsten Patienten zuwendet, wird der Text schon geschrieben – nichts Ungewöhnliches, würde nicht ein Computer statt einer Sekretärin die Sprachaufnahme in Text umsetzen.

Sprache zu erkennen gehört zu den ersten Dingen überhaupt, die ein Mensch lernt. Was macht diese scheinbar elementare Leistung für Maschinen so schwierig? Der Hauptgrund ist die hohe Variabilität des Sprachsignals: Auch wenn wir es kaum wahrnehmen, unterscheiden sich die akustischen Realisierungen sehr stark, wenn verschiedene Leute denselben Text sprechen, und selbst dann, wenn ein Sprecher seine eigene Äußerung exakt wiederholt. Beim natürlichen fließenden Sprechen kommt hinzu, daß Wortgrenzen im Sprachsignal meist nicht unmittelbar erkennbar sind und Laute verschliffen werden.

Ein Mensch erkennt das richtige Wort, weil er ein reichhaltiges Vorwissen einsetzt: Aus der Situation, den Regeln der Grammatik und dem Inhalt des bisher Gesagten kann er häufig schon erschließen, welches Wort als nächstes kommt. Außerdem vermag er dank seiner Hörerfahrung mühelos sehr verschiedene Lautäußerungen mit demselben Wort zu identifizieren.

In zwei Forschungsprogrammen Anfang der siebziger beziehungsweise Ende der achtziger Jahre hat man versucht, dieses Vorwissen mit Methoden der Künstlichen Intelligenz zu formalisieren und dadurch für die maschinelle Spracherkennung nutzbar zu machen – mit insgesamt unbefriedigendem Erfolg. Durchgesetzt hat sich dagegen der sogenannte statistische Ansatz.

An die Stelle des semantischen und pragmatischen Vorwissens tritt dabei ein stochastisches Sprachmodell, das – unabhängig von akustischer Information – nur Auskunft darüber gibt, mit welcher Wahrscheinlichkeit ein Wort in diesem Kontext auftritt; an die Stelle der Hörerfahrung tritt ein akustisches Modell für die Wahrscheinlichkeit, daß dieses Wort, würde es ausgesprochen, so klingt wie das Gehörte. Aus der statistischen Entscheidungstheorie ergibt sich, daß das Spracherkennungssystem sich für diejenige Wortfolge entscheiden sollte, für die das Produkt der beiden genannten Wahrscheinlichkeiten maximal ist (vergleiche den vorstehenden Beitrag von Marcus Spies).

Beim statistischen Ansatz begnügt man sich damit, einen relativ kleinen Teil des menschlichen Vorwissens der Maschine verfügbar zu machen. Der Lohn dieser Selbstbescheidung besteht darin, daß man – im Gegensatz zu der Formalisierung von Wissen – brauchbare Schätzwerte für die genannten Wahrscheinlichkeiten nahezu automatisch gewinnen kann, indem man ein Computerprogramm große Mengen an gesprochenen und geschriebenen Texten analysieren läßt. Zugleich kombiniert eine solche quantitative Beschreibung so verschiedene Wissensquellen wie Linguistik, Akustik und Phonetik in einem einheitlichen Formalismus.

Entscheidend für die Entwicklung eines leistungsfähigen Spracherkenners ist die Vorgabe mathematischer Strukturen oder Modelle für die Schätzung dieser Wahrscheinlichkeiten. Beispiele sind die Hidden-Markov-Modelle für Folgen akustischer Zustände und die – den Gaußschen Glockenkurven ähnlichen – Verteilungen, mit denen man beschreibt, wie weit und mit welcher Wahrscheinlichkeit die konkrete Realisierung eines Phonems – einer kleineren, bedeutungsunterscheidenden Lauteinheit – von einem Mittelwert abzuweichen pflegt.

Solche Modelle enthalten noch freie, nicht von vornherein bekannte Parameter, etwa die genannten Mittelwerte. Insbesondere die letzteren sind von Mensch zu Mensch verschieden. Das System muß sie gewissermaßen lernen, indem es Sprachaufnahmen bekannter Texte verarbeitet. Nach dieser vollautomatisch ablaufenden Trainingsphase kann es nun neu gesprochenen Text erkennen.

Ein Maß für die Güte der Modellierung ist die Wortfehlerrate bei der Erkennung unter wohldefinierten Testbedingungen. Sie hängt sehr stark von der Person des Sprechers, seiner Sprechweise, dem zugrundeliegenden Vokabular, der Redundanz des gesprochenen Textes und den akustischen Bedingungen ab; typische Werte für die zur Zeit weltweit besten Systeme liegen bei 0,1 bis 1 Prozent für Ziffernketten (zum Beispiel Telefonnummern) und im Bereich von 10 Prozent für Diktate bei sehr großem Wortschatz. Es werden also im letzten Fall neun von zehn Wörtern korrekt erkannt.

Ein Spracherkennungssystem muß zunächst das Sprachsignal in eine für die Analyse geeignete Form bringen (akustische Vorverarbeitung), dann für eine große Anzahl von Hypothesen über den ge-

sprochenen Text die oben genannten Wahrscheinlichkeiten finden (akustische beziehungsweise Sprachmodellierung) und unter diesen die wahrscheinlichste ausfindig machen (Suche). Am Philips-Forschungslaboratorium in Aachen beschäftigen wir uns mit diesen vier Problemkreisen.

Akustische Vorverarbeitung

Ähnlich wie bei einer Aufnahme für eine Compact Disk wird das analoge Sprachsignal zunächst digitalisiert. Aus den Abtastwerten wird in einem 10-Millisekunden-Zeitraster eine Folge von Merkmalsvektoren gewonnen, die (bei gleichzeitiger Datenreduktion) noch möglichst relevante Informationen über das Gesprochene enthalten sollen. Wir verwenden das logarithmierte Leistungsdichtespektrum, das Auskunft über die spektrale Verteilung innerhalb eines kurzen Zeitfensters (bei uns 25 Millisekunden) gibt; außerdem werden noch zeitliche Änderungen dieses Spektrums berücksichtigt. Weitere Verarbeitungsschritte erhöhen die Robustheit gegenüber unterschiedlichen Aufnahmebedingungen.

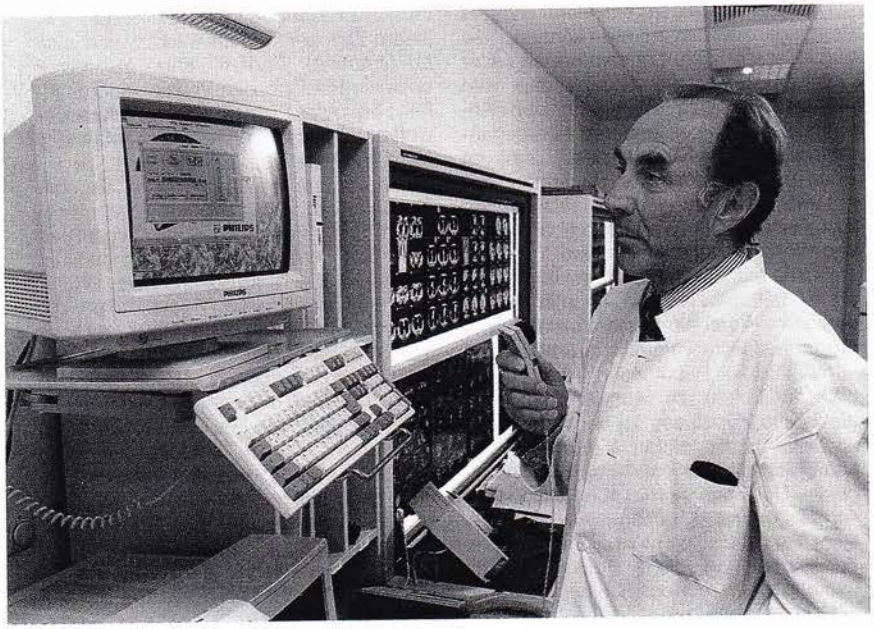
Akustische Modellierung

Stimme und Sprechweise des Sprechers werden durch das akustische Modell beschrieben. Es benutzt etwa eine halbe Million freier Parameter, die anhand gesprochener Daten in der Trainingsphase geschätzt (gelernt) werden. Bei großem Erkennungsvokabular werden typischerweise nicht direkt Wortmodelle gelernt, sondern Phonemmodelle. Wie sich Phoneme als lautliche Bausteine zu Wörtern zusammensetzen lassen, so werden Phonemmodelle anhand eines Aussprachelexikons zu Wortmodellen kombiniert.

Die Topologie unserer Modelle ist einfach: Sie bestehen aus Ketten von Zuständen, die in etwa einer zeitlichen Abfolge von Lauten entsprechen. Alle akustischen Phänomene werden über die Parameter akustischer Wahrscheinlichkeitsverteilungen für die Zustände beschrieben. Es handelt sich um gewichtete Summen von glockenförmigen Verteilungen.

Wie spricht man in ein automatisches Diktiersystem?

Um von der Maschine zuverlässig verstanden zu werden, genügt es, einige einfache Regeln zu beachten:
- Sprechen Sie deutlich (aber nicht übertrieben oder unnatürlich);
- verschleifen Sie Wörter und insbesondere Wortenden nicht zu sehr: „können wir noch einen" statt „könnwanochein";
- sprechen Sie im Training genauso wie später im Betrieb. Wer stets schnell und undeutlich spricht, sollte dies auch in der Trainingsphase tun, damit die Maschine sich daran gewöhnen kann.



„Speech Processing System 6000" im Einsatz im Klinikum rechts der Isar in München.

Sprachmodellierung

Wir gehen, wie heute bei Spracherkennungssystemen üblich, von einem beliebig vorgegebenen Vokabular von Wörtern aus, die überhaupt erkannt werden sollen. Beispielsweise können das alle Wörter sein, die in einer bestimmten Textsammlung vorkommen.

Das Sprachmodell gibt zu jedem potentiell möglichen Satz eine Wahrscheinlichkeit dafür an, daß diese Folge von Wörtern vorkommen könnte. Im Idealfall enthält das Modell also Wissen über Grammatik, Bedeutung und die Äußerungssituation, wie es den linguistischen Beschreibungsebenen Syntax, Semantik und Pragmatik entspricht. Mit großen Textmengen trainierte stochastische Sprachmodelle, die auf der Häufigkeitsanalyse von Wortpaaren (Bigrammen) oder Worttripeln (Trigrammen) basieren, haben sich in vielen Experimenten als die erfolgreichsten erwiesen.

Suchverfahren

Bei der Suche nach der optimalen Wortfolge geht es darum, ein Optimierungsproblem in einem enorm großen Suchraum so geschickt zu lösen, daß es auf einem heutigen Rechner in akzeptabler Zeit abläuft. Die erste Schwierigkeit ist die astronomische Anzahl möglicher Wortfolgen: Mit 10^4 Wörtern etwa lassen sich 10^{80} Sätze von 20 Wörtern Länge bilden. Zudem ist über die unbekannten Wortgrenzen zu optimieren; und auch innerhalb von Wörtern läßt die variierende Sprechgeschwindigkeit viele Möglichkeiten zu, das sprachliche Signal mit den Phonemmodellen zu synchronisieren.

All diese Probleme werden zurückgeführt auf die Aufgabe, den optimalen Pfad in einem bewerteten Graphen zu finden (siehe Kasten Seite 96). Da künftige Beobachtungen für die Wahl der optimalen Wortfolge ausschlaggebend sein können, betrachtet das System zu jedem Zeitpunkt viele verschiedene Teilsatzhypothesen; es trifft nur dann eine endgültige Entscheidung, wenn künftige Beobachtungen sie nicht mehr beeinflussen können. Der gesamte Erkennungsvorgang läuft auf einem Personal Computer mit einem heute gängigen Prozessor des Typs 486 sowie einer Beschleunigerkarte in 0,8- bis 3-facher Echtzeit ab.

Von der Technologie zum Produkt

Die Ergebnisse unserer Arbeit sind Grundlage des von Philips Dictation Systems in Wien entwickelten Spracherken-

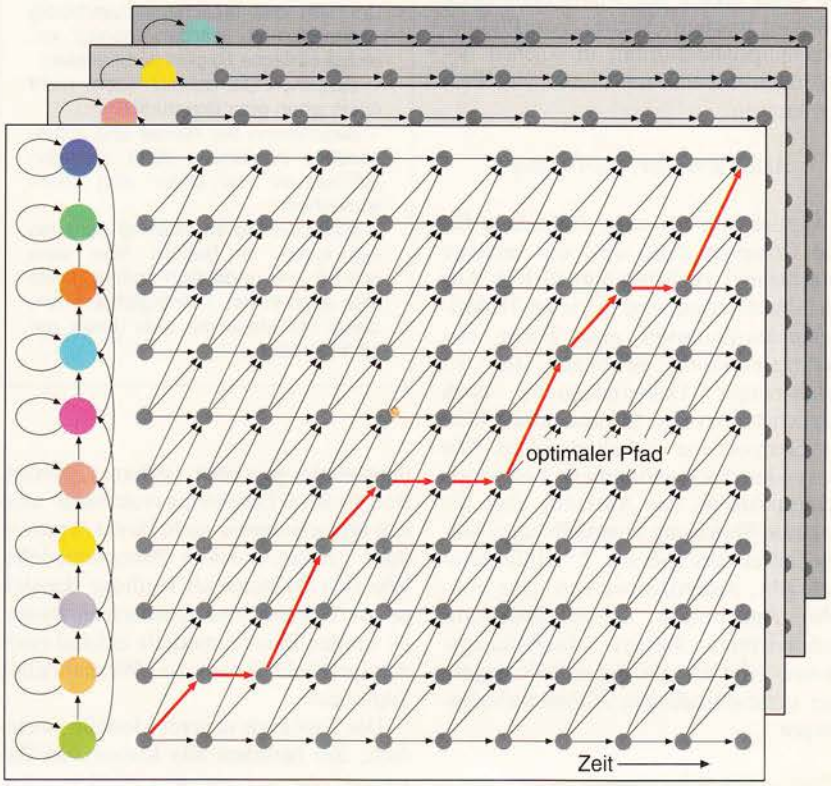
Auf der Suche nach der optimalen Wortfolge

Bei der Erkennung geht es darum, zu einer vorliegenden Folge von Merkmalsvektoren (einer „akustischen Beobachtung" im Fachjargon) die wahrscheinlichste Zustandsfolge zu bestimmen. Ein Zustand ist die Abstraktion eines Lautereignisses, genauer: von Anfang, Mitte oder Ende eines Phonems. Zu jedem Wort gehört eine Markow-Kette, die aus Zuständen und Übergängen zwischen diesen besteht. Jede von ihnen kann sehr viele Zustandsfolgen erzeugen, und unter all diesen besteht die Auswahl. Man hat es sozusagen mit einem gigantischen Puzzlespiel zu tun, bei dem – Hauptschwierigkeit der akustischen Erkennung – die einzelnen Puzzleteilchen nur ungefähr zueinander passen.

Beschränken wir uns zunächst auf ein Teilproblem, die Synchronisierung eines isoliert geprochenen Wortes mit dem akustischen Zustandsmodell, auch nichtlineare Zeitanpassung genannt. Es ist eine Folge von Beobachtungen zu diskreten Zeitpunkten mit den Zuständen eines Modells in Einklang zu bringen. Zur Veranschaulichung trägt man zweckmäßig in einem Diagramm über jedem Zeitpunkt die Kette der Zustände auf (Bild). Im Zustandsmodell (links) entspricht jeder Pfeil einem Übergang von einem Zustand zu einem Folgezustand (oder auch dem Verbleib in diesem Zustand). In der Zeitdarstellung verlaufen diese Pfeile außerdem jeweils von einem Zeitpunkt zum nächsten. Man erhält ein Gitter, dessen Knotenpunkte Zuständen zu bestimmten Zeitpunkten entsprechen; zu einer zeitlichen Zustandsfolge gehört ein erlaubter Pfad durch dieses Gitter. Unsere Modelle berücksichtigen Übergänge von einem Zustand zu sich selbst, zum nächsten und zum übernächsten in der Kette; im zugehörigen Gitter gehen entsprechend von jedem Punkt drei Teilpfade aus, und in jedem mit Ausnahme der Randpunkte treffen drei Pfeile vom vorigen Zeitpunkt zusammen.

Auf der Suche nach dem wahrscheinlichsten Pfad zwischen Wortanfang und -ende (im Bild links unten und rechts oben) geht das Spracherkennungssystem in der gleichen zeitlichen Reihenfolge wie bei der Aufnahme vor. Jeder Teilpfad hat eine gewisse Wahrscheinlichkeit für sich und bekommt deren Logarithmus als sogenannte Bewertung zugewiesen. Die Bewertung eines Gesamtpfades ist die Summe der Einzelbewertungen seiner Teilpfade. Deshalb kann bei ihrer Berechnung auf die Bewertungen der Pfade zum vorhergehenden Zeitpunkt zurückgegriffen werden.

Laufen mehrere Pfade in demselben Gitterpunkt zusammen, wird nur der beste weiterverfolgt, denn er kann auch durch spätere Beobachtungen von seinen Konkurrenten nie mehr eingeholt werden. Diese Rekombination von Pfaden sorgt dafür, daß die Gesamtzahl der bearbeiteten Pfade nur linear, nicht aber exponentiell mit der Zeit wächst.



Der Suchvorgang läuft gleichzeitig für alle Wörter im Vokabular ab; das System verfolgt also stets Pfade in mehreren gleichsam hintereinanderliegenden Gittern. Der beste Pfad kennzeichnet das Wort, das vermutlich gesprochen wurde.

Um anstelle eines einzelnen Wortes Wortfolgen erkennen zu können, die durch Pausen getrennt sein dürfen (aber nicht müssen), wird erlaubt, daß vom Endzustand eines Wortes in den Anfangszustand eines anderen übergegangen wird. Der Suchvorgang bleibt im Prinzip gleich, doch der Suchraum vergrößert sich, da jetzt zusätzlich zur Zeitanpassung im Wort noch über Wörter und Wortgrenzen optimiert wird.

Würde man Rekombination von Pfaden auf Wortebene zulassen, könnten Fehler auftreten. Wenn ein System bislang zwei Wörter des Sprachsignals verarbeitet und daraufhin die Hypothesen „die Lunge" und „der Lunge" zur Auswahl hat, wäre die zweite Hypothese zu verwerfen, weil nach der vorliegenden Bigramm-(Wortpaar-)Statistik das erste Paar weit häufiger ist. Der Sprecher hat aber „Der Lunge vorgelagert ist..." gesagt, was das System dann nicht mehr erkennen könnte. Zur Vermeidung solcher Fehler führt das System mehrere Kopien eines Wortmodells im Speicher mit.

Der Suchaufwand läßt sich stark verringern, indem die Suche auf Pfade mit relativ guter Bewertung eingeschränkt wird. Dazu läßt man zu jedem Zeitpunkt nur noch diejenigen Pfade weiterverfolgen, deren Bewertung sich von dem zur Zeit besten Pfad um nicht mehr als einen vorgegebenen Schwellenwert unterscheidet: Man schneidet gewissermaßen von dem in Entstehung befindlichen Baum zu jedem Zeitpunkt die kümmerlichsten Äste ab (*pruning*). Dadurch kann der global optimale Pfad zwar im Prinzip verlorengehen; bei guter Einstellung des Schwellenwertes fallen Suchfehler dieser Art jedoch nicht ins Gewicht.

Der Suchraum läßt sich weiter dadurch einschränken, daß man das Vokabular in einem phonetischen Baum organisiert (der mit obigem Entscheidungsbaum nichts zu tun hat), denn viele Wörter beginnen mit derselben Phonemfolge. Zusätzlich wird durch eine kurze zeitliche Vorausschau geprüft, ob das nächste Phonemmodell auch in 60 Millisekunden noch gut genug bewertet sein wird.

Durch alle diese Maßnahmen läßt sich die Anzahl der zu berücksichtigenden Pfade in jedem Moment in der Größenordnung von einigen tausend halten. Dies sind wenig genug, daß die Erkennung mit einem Wortschatz von 20 000 Wörtern auf einem Spracherkennungs-PC nur ein- bis dreimal so lange dauert wie die Sprachaufnahme selbst.

nungssystems „Speech Processing System 6000" (SP 6000), mit dem beispielsweise Radiologen ihre Befunde diktieren können (Bild). Gerade in den radiologischen Abteilungen großer Krankenhäuser fällt eine Flut von Schreibarbeiten an: häufig sind es mehr als 100 000 Befunde pro Jahr. Das Erkennungsvokabular ist mit 25 000 Wortformen genügend groß, um dem Arzt auch inhaltlich freies Diktieren zu ermöglichen. Man darf fließend, das heißt ohne künstlich eingefügte Pausen zwischen den Wörtern, und mit der gewohnten Sprechgeschwindigkeit diktieren. Die Beachtung einiger einfacher Regeln ist hilfreich (siehe Kasten Seite 95).

Die drei Phasen Diktieren, Spracherkennung und Korrektur laufen räumlich und zeitlich getrennt ab. Den erkannten Text muß die Schreibkraft nur noch redigieren. Sie wird dabei von einem neuartigen Korrektur-Editor unterstützt, der beim Abspielen der Aufnahme das jeweils gesprochene Wort auf dem Bildschirm hervorhebt und auf Anforderung zu einer Textstelle den zugehörigen Aufnahmeabschnitt abspielt. Zur endgültigen Formatierung wird der Text schließlich in ein übliches Textverarbeitungsprogramm übernommen.

Die hier beschriebenen Methoden sind nicht auf das Diktieren deutschsprachiger Radiologiebefunde beschränkt; auch amerikanischen Radiologen wurde schon ein Prototyp vorgestellt. Es ist lediglich eine Frage der Zeit, wann andere Anwendungsgebiete außerhalb der Medizin erschlossen werden. In der Radiologie hat die Zukunft nur etwas früher begonnen.

Dr. Steinbiß ist Projektleiter für Spracherkennung in den Forschungslaboratorien der Philips GmbH in Aachen.

Das Telephon als intelligenter Gesprächspartner

Von Helmut Mangold

Jeder Mitspieler der Nordwestdeutschen Klassenlotterie kann nach der Ziehung eine Telephonnummer wählen und der sich meldenden Stimme seine Losnummer vorsprechen. Daraufhin erhält er in natürlicher Sprache Auskunft, ob – und wenn, wieviel – er gewonnen hat.

Hinter der Telephonstimme steckt ein Computer, der in diesem Falle von der Firma Dornier realisiert worden ist. Er kann nur die für einen solchen Dialog erforderlichen Wörter erkennen: die zehn Ziffern und einige Kommandowörter. Die Eingabe „einundzwanzig siebenundvierzig" statt „zwei eins vier sieben" ist also nicht vorgesehen. Doch diese Leistung muß das System erbringen, einerlei ob eine männliche oder eine weibliche, eine junge oder eine alte Stimme, hochdeutsch oder dialektgefärbt, die Wörter spricht.

Sprechende Auskunftssysteme sind nicht neu; schon Anfang der achtziger Jahre konnte man innerhalb des Ortsnetzes Frankfurt am Main Fahrplanauskünfte der Bundesbahn über das von Dornier und AEG gemeinsam realisierte System „Karlchen" einholen. Bei diesem zeitweise weltweit größten System mußte man allerdings alle Daten mühsam über die Wählscheibe beziehungsweise Tastatur des Telephons eingeben; die Auskunft wurde dann in natürlicher Sprache erteilt. Spracherkennende Systeme dagegen sind erst in jüngster Zeit zum praktischen Einsatz gekommen.

Heute gelingt die automatische Erkennung kleinerer Vokabularien von einigen zehn bis zu einigen hundert Wörtern auch für die Stimmen ungeübter Benutzer mit großer Zuverlässigkeit. Enorm gesteigerte Computerleistungen erlauben den Einsatz neuer Verfahren, die hauptsächlich auf den statistischen Eigenschaften unserer Sprache basieren. In der Trainingsphase sind gigantische Datenmengen auszuwerten: Hunderte oder gar Tausende sehr unterschiedlicher Sprecher müssen einen hinreichend repräsentativen Querschnitt aller Stimm- und Sprachvarianten liefern. Nur dann kann das System später fast jeden Sprecher erkennen.

Erst seit die Systeme solche Datenbasen nahezu automatisch analysieren können, ist diese Massenauswertung überhaupt möglich. Bis dahin mußte man mühsam sämtliche Sprachdaten – Laute, Wörter und Sätze – einzeln abhören und mit entsprechenden Marken versehen, die dem Computer signalisierten, mit was er gerade trainiert wurde. Inzwischen reicht es aus, dem Computer parallel zum gesprochenen Sprachsignal dessen geschriebene Darstellung zu liefern.

Aus dieser erzeugt man mit Hilfe von Regelsystemen eine phonetische Beschreibung für die Normaussprache sowie die üblichen Aussprachevarianten und aus dieser die entsprechenden Markow-Modelle für den Spracherkenner. Damit dient also die geschriebene Form als Basis für das Training. Dafür ist besondere Sorgfalt erforderlich; Schreib- oder Sprechfehler in diesem Stadium würden sich verheerend auswirken.

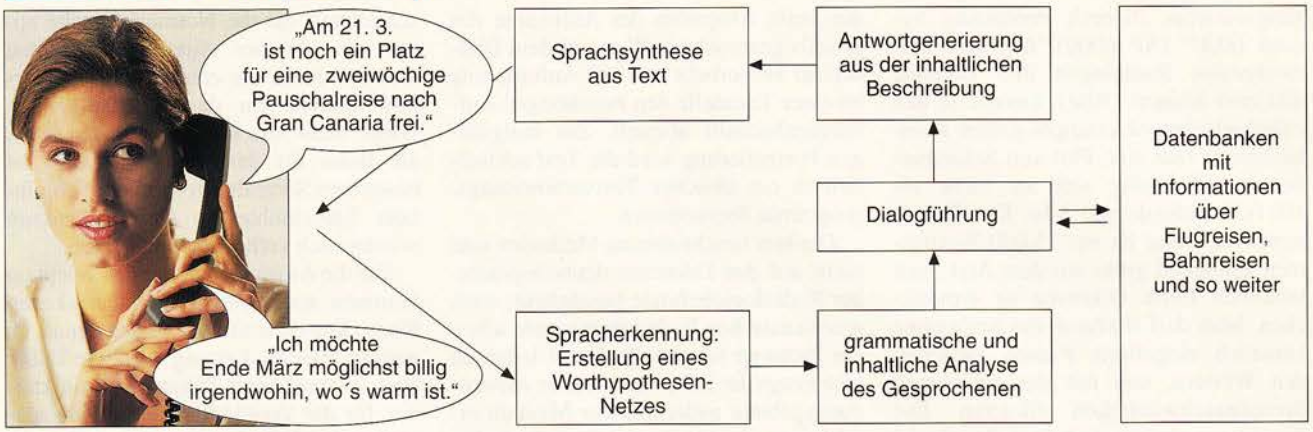
Für die Anwendung über das Telephon kommen noch weitere Schwierigkeiten hinzu: Die Qualität des Sprachsignals ist gering; über die Leitung wird wie bisher auch in absehbarer Zukunft nur ein kleiner, für die Verständlichkeit gerade ausreichender Teil des Frequenzspektrums übertragen (ungefähr 3 Kilohertz). Die Lautstärke schwankt stark, und wenn der Benutzer den Hörer unter das Kinn klemmt, wird das Sprachsignal verzerrt. Meistens erreicht aber ein Spracherkenner auch dann noch Erkennungsraten, die für kleine Vokabularien nur wenig unter 100 Prozent liegen.

Hin und wieder stößt das System allerdings auf Sprecher, deren stimmliche Charakteristika anscheinend im Trainingsmaterial nicht genügend berücksichtigt worden waren. Bei ihnen sinkt dann die Erkennungsrate möglicherweise auf 90 Prozent oder weniger ab. Dieses Problem soll schon bald durch adaptive Systeme gelöst werden, die sich sehr schnell – nach wenigen Wörtern – den spezifischen Eigenschaften der jeweiligen Stimme anpassen.

Dazu müssen im wesentlichen alle Eigenschaften der Markow-Modelle der einzelnen Wörter an die neue Stimme adaptiert werden. Wenn der Sprecher einigermaßen hochdeutsch spricht, kann sich die Anpassung auf die besonderen Frequenzeigenschaften seiner Stimme beschränken. Wenn er jedoch zum Beispiel regelmäßig „eens" oder „oans" statt „eins" sagt, sind weitergehende Maßnahmen erforderlich.

In jedem Fall muß für den Anpassungsprozeß, insbesondere für die Berechnung der Abweichungen von der Norm, bekannt sein, welches Wort gemeint war. Wenn das – gerade wegen der schlechten Aussprache – unsicher ist, muß das System unter mehreren möglichen Alternativen für das Wort versuchsweise diejenige auswählen, welche die beste Anpassung liefert, und sich notfalls sogar durch Rückfrage beim Sprecher vergewissern, ob es das Wort richtig erkannt hat. Wenn nicht, unternimmt es einen weiteren Optimierungsschritt.

Es ist bei allen solchen automatischen Adaptionsverfahren wichtig, die Parameter nicht zu weit in Richtung des soeben



Schema eines Dialogs zwischen Mensch und Maschine.

Gehörten zu verändern. Dadurch würde das System gleichsam diesen Einzelfall zur Regel erheben und beim nächsten Satz desselben Sprechers mit dieser – möglicherweise falschen – Hypothese scheitern.

Sehr stark durch Dialekt verfärbte Wörter sind auf diese Weise nicht mehr zu erfassen, sondern müssen als gänzlich neue Wörter aufgefaßt und gelernt werden. In ähnlicher Form versagt in der Regel auch unser menschliches Adaptionsvermögen.

Das Dialogsystem kann im Extremfall rückfragen oder sogar den Benutzer bitten, das zuletzt gesprochene Wort zu buchstabieren; damit erfährt es zumindest die Schreibweise eines neuen Wortes. Sofern dieses nach den üblichen Regeln ausgesprochen wird, läßt sich mit Hilfe eines entsprechenden Regelwerks seine Aussprache und damit das entsprechende Markow-Modell leicht ermitteln, und fortan gehört das Wort zum aktiven Wortschatz des Systems. Das ist allerdings – wie bei einem Menschen, der ein neues Wort lernt – nur dann sinnvoll, wenn auch die grammatische Funktion und die Bedeutung des neuen Wortes klar sind. Beim heutigen Stand würde das für Eigennamen gelten sowie für Synonyme, die ein Benutzer anstelle der im Dialog vorgesehenen und zulässigen Wörter verwendet.

Die Kunst der Dialoggestaltung

Ein geschickt aufgebauter Dialog trägt unter Umständen mehr zur Zufriedenheit des Benutzers bei als eine perfekt funktionierende Spracherkennung. In vielen Fällen kann das gesprächsführende Programm sogar einzelne Schwächen der Erkennung oder auch der Sprachausgabe kompensieren.

Es verwundert deshalb nicht, daß gerade die telephonischen Informationssysteme, deren Dialogablauf in enger Zusammenarbeit mit potentiellen Benutzern optimiert worden ist, zu den erfolgreichsten zählen. Sie enthalten in der Regel auch mehrere Dialogvarianten für mehr oder weniger erfahrene Systembenutzer.

Bei einfachen Systemen wie der erwähnten Lotterie- oder auch einer Fahrplanauskunft geht die Dialogführung immer vom System aus. Der Benutzer hat keine Chance, den Ablauf selbst zu bestimmen, sondern muß sich ebenso strikt an die Vorgaben halten wie der Benutzer heutiger menügesteuerter Computerprogramme.

Weiterentwickelte Systeme werden jedoch eine viel natürlichere Spracheingabe bis hin zu ganzen Sätzen verarbeiten können. Dann wird sich auch der Dialog zwischen Mensch und Maschine dem unter Menschen annähern.

Die wesentliche Aufgabe eines solchen Systems ist die Abbildung der menschlichen Dialogabläufe auf die maschinellen Möglichkeiten (Bild). Dazu gehören nicht nur die automatische Erkennung der menschlichen Sprache und das Synthetisieren einer sprachlichen Antwort, sondern auch das inhaltliche Verstehen und damit die Einbindung in den entsprechenden Wissenshintergrund von Mensch und Maschine.

Das System SUNDIAL (*Speech Understanding and Dialogue*), das wir im Rahmen des europäischen Forschungsprogramms ESPRIT in Zusammenarbeit mit Partnern aus Deutschland, Großbritannien, Frankreich und Italien entwikkelt haben, realisiert erstmals einen solchen Ansatz. Es erteilt Verkehrsauskünfte in natürlicher Sprache und erkennt natürlich gesprochene Sätze. Der Benutzer hat damit viel Freiheit bei der Formulierung seiner Wünsche; er kann insbesondere einen Satz mit diversen Informationen befrachten: „Bitte geben Sie mir eine Zugverbindung erster Klasse mit Speisewagen morgen früh von München nach Dortmund." Bei heute noch üblichen Systemen würde der Computer völlig schematisch Begriffe wie „Abfahrtsort", „Zielort", „gewünschte Abfahrtszeit", „Wagenklasse", „Speisewagen" und so weiter nacheinander abfragen. SUNDIAL ist zwar von ungeübten Benutzern ausgiebig getestet, bisher jedoch noch nicht praktisch eingesetzt worden. Doch werden schon in Kürze wesentliche Elemente in neue telephonische Auskunftssysteme eingehen.

Sprachverstehen

Für die entscheidende neue Komponente, das inhaltliche Verstehen des Gesprochenen, hat die moderne Computerlinguistik in den letzten Jahren wesentliche Voraussetzungen geschaffen. Ein Mensch spricht nur in sehr seltenen Fällen schriftreif, und die klassischen Methoden der Computerlinguistik versagen bei unvollständigen oder grammatisch falschen Sätzen. Außerdem gelingt die maschinelle Erkennung der einzelnen Wörter eines flüssig gesprochenen Satzes weit schlechter als diejenige isolierter Wörter; das Programm muß dann zahlreiche Hypothesen für die Wörter des Satzes berücksichtigen. Erst die grammatische und inhaltliche Analyse engt die Auswahl mehr und mehr ein.

Wir gehen dazu von sogenannten Wortinseln aus. Das sind spezielle Wörter, die besonders sicher erkannt werden und gewissermaßen als Ankerpunkte für die weitere linguistische Verarbeitung dienen. Typischerweise handelt es sich um den Namen eines Reiseziels oder ein

sonstiges bedeutungstragendes Element eines Informationswunsches, das ein Benutzer besonders deutlich auszusprechen pflegt. In einem Satz können sich dabei durchaus mehrere solcher Inseln finden, die letztlich zu einem inhaltlichen Ganzen zu verbinden sind. Die Technik dieser linguistischen Verfahren, die wir unter dem Begriff „Insel-Parser" zusammenfassen, hat die sprachverstehenden Systeme in den letzten Jahren deutlich vorangebracht.

Ergebnis der grammatischen und inhaltlichen Analyse ist eine Beschreibung des Informationswunsches in einer abstrakten Form. Sie ist das Material, mit dem die nächste, für den Dialog verantwortliche, Komponente des Systems arbeitet. Diese stellt einerseits die Verbindung zum Wissenshintergrund des Systems, also der Datenbank, her. Andererseits versucht sie, möglichst den gesamten bisherigen Gesprächsverlauf in seinen wesentlichen inhaltlichen Teilen zu speichern. Dieses sogenannte Dialogwissen ist letztlich entscheidend dafür, ob das System intelligent wirkt – und damit im Sinne des berühmten Turing-Tests (vergleiche Spektrum der Wissenschaft, März 1990, Seite 47) als intelligent anzusehen ist – oder nur ohne Rücksicht auf das bisher Gesprochene die jeweils letzte Frage beantwortet.

Von Dialogsystemen mit völlig natürlicher Sprache sind wir allerdings noch weit entfernt. Ein aktueller Forschungsgegenstand ist die Problematik der Spontansprache. Über die Unvollständigkeit und die grammatische Fehlerhaftigkeit von Sätzen hinaus geht es vor allem um die vielgebrauchten „äh" und „mh" sowie abgebrochene Wörter (vergleiche den folgenden Beitrag von Wolfgang Wahlster). Daß schließlich auch noch Hintergrundgeräusche das Sprachverstehen erschweren, sei hier nur am Rande erwähnt. Für die meisten dieser Probleme gibt es bereits erste Lösungen.

Intelligente Informationsnetze

Mit dem Schlagwort vom Telephon als einem intelligenten Gesprächspartner meinen wir die Gesamtheit aus Mensch-Maschine-Dialog und Hintergrundwissen. Erst durch eine umfangreiche Datenbasis und das darin gespeicherte Wissen wird der Dialog nützlich und attraktiv.

Durchsetzen werden sich allerdings nicht Informationssysteme, die nur mit einer einzigen Datenbank im Hintergrund arbeiten, sondern vernetzte Systeme. Oftmals kann der Benutzer vorab gar nicht genau wissen, welche Datenbank er in Anspruch nehmen will. Beispielsweise klärt sich erst im Verlauf eines Dialogs über eine geplante Urlaubsreise, ob er eine Bahn- oder eine Flugauskunft wünscht und von welcher Region ihn das Hotelverzeichnis oder eine Liste der Freizeitangebote interessiert. Ein entsprechendes System muß also intelligent genug sein, uns entsprechend unseren gesprochenen Wünschen fast automatisch durch all diese Informationsbereiche zu führen. Es wird dabei selbst entscheiden, aus welcher Quelle es die jeweilige Auskunft bezieht, und möglicherweise auch, welche mit Priorität auszugeben ist.

Mit heutigen Experimentalsystemen lassen sich bereits Dialoge verwirklichen, die zwar ein Vokabular von einigen tausend Wörtern benutzen, aber einstweilen noch grammatisch weitgehend richtige Sätze als Eingaben verlangen. Schon in naher Zukunft wird man auch manche Unkorrektheiten beherrschen. Damit werden sich die maschinellen Dialogsysteme immer mehr dem menschlichen Benutzer und seinem Verhalten anpassen.

Diplom-Ingenieur Mangold
ist Leiter des Arbeitsbereichs
Sprachverstehende Systeme
im Institut für Informationstechnik
der Daimler-Benz AG
in Ulm.

Verbmobil – Übersetzungshilfe für Verhandlungsdialoge

Von Wolfgang Wahlster

Verbmobil ist ein Verbundprojekt des Bundesministeriums für Forschung und Technologie (BMFT) mit dem langfristigen Ziel, ein System zu entwickeln, das die Übersetzung eines Dialogs mit fremdsprachlichen Gesprächspartnern zu unterstützen vermag. Es geht nicht um neue Hardware oder die Ersetzung des klassischen Dolmetschers durch eine Maschine; die zu entwickelnde Software soll vielmehr auf transportablen Computern (daher der Name Verbmobil) nutzbar und während eines Gesprächs hilfsweise aktivierbar sein. Solange die erforderlichen Rechenleistungen und Datenmengen so nicht verfügbar sind, greift Verbmobil beispielsweise über drahtlose Telekommunikation auf stationäre Hochleistungsrechner zurück.

Das Projekt führt erstmals bislang getrennte Entwicklungslinien der Sprachtechnologie zusammen. Dies und der Zusammenschluß möglichst vieler Experten sollen Deutschland in den nächsten Jahrzehnten eine Spitzenposition in der Sprachtechnologie verschaffen.

Das Projekt ist auf acht bis zehn Jahre angelegt. Die erste vierjährige Phase ist in zwei Stadien gegliedert: Der Demonstrator, eine Frühversion des Systems, soll nach zwei, ein Forschungsprototyp nach vier Jahren verfügbar sein. Für die Zeit danach ist die Entwicklung zur Marktreife vorgesehen. Die wissenschaftliche Leitung liegt beim Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) in Saarbrücken.

Vorausgegangen waren zwei umfangreiche, vom BMFT in Auftrag gegebene Machbarkeitsstudien, eine öffentliche Ausschreibung und die Begutachtung aller eingegangenen Projektanträge. Eine der Studien wurde von einem deutschen Konsortium erstellt, an dem unter anderem die Universitäten Hamburg, Karlsruhe und Stuttgart, die Technische Universität Berlin, das DFKI sowie die Firmen Siemens und IBM beteiligt waren, die andere vom Center for the Study of Language and Information (CSLI) an der Universität Stanford (Kalifornien). Die Hauptphase des Projekts hat 1993 begonnen. Beteiligt sind die Unternehmen Alcatel SEL, CAP Debis, Daimler-Benz, IBM, Philips, Siemens und die Deutsche Aerospace sowie 20 deutsche Universitäten und Forschungsinstitute.

Das anspruchsvolle Ziel wird in einer Folge von wohldefinierten Schritten angestrebt. Der Forschungsprototyp wird den Dialog beispielsweise zwischen einem Deutschen und einem Japaner unterstützen, die beide Englisch verstehen, aber nicht perfekt sprechen können. Es wird angenommen, daß große Teile solcher Dialoge auf Englisch stattfinden, wie es für internationale Diskussionen im Bereich von Technik und Wirtschaft zutrifft. Aber bei wenig gebräuchlichen Wörtern oder Formulierungen, schwierigen Satzkonstruktionen und für den Gesprächserfolg wichtigen Diskussionsab-

schnitten möchten die Gesprächspartner auf ihre Muttersprache zurückgreifen. In derartigen Fällen soll jeder seine Version von Verbmobil (Deutsch–Englisch beziehungsweise Japanisch–Englisch) aktivieren können und es die nun folgenden Worte aus seiner Muttersprache ins Englische übersetzen lassen (Bild 1).

Demnach muß das System über drei verschiedene Arbeitsweisen verfügen:

– Solange beide Partner sich in Englisch als Fremdsprache unterhalten, muß Verbmobil offensichtlich nichts übersetzen, aber den Dialog verfolgen und Kontextinformation für nachfolgende Übersetzungsaufgaben extrahieren – gleichsam zuhören, damit es weiß, worum es geht. Dies ist ein extrem schwieriges Problem, weil die Gesprächspartner mit ihren für Deutsche beziehungsweise Japaner typischen Unzulänglichkeiten und Fehlern sich in Aussprache, Wortwahl und Grammatik viel weniger als ein Muttersprachler an die üblichen Regeln halten werden. Das System wird daher in den meisten Fällen nur ein sehr flaches (oberflächliches) und stark vereinfachtes Diskursmodell aufbauen können. Für die Realisierung will man zunächst auf Schlüsselworterkennung (*word spotting*) und andere robuste, aber nicht erschöpfende Analysetechniken zurückgreifen.

– Wenn einer der Dialogpartner innerhalb einer Äußerung in seine Muttersprache wechselt, da er spontan nicht imstande ist, seine Intention in Englisch weiterzuformulieren, soll Verbmobil eine englischsprachige Äußerung synthetisieren, die das bereits ausgesprochene englische Satzfragment korrekt fortsetzt.

– Der Gesprächspartner formuliert eine vollständige Äußerung in seiner Muttersprache, und Verbmobil übersetzt sie ins Englische. In diesem Falle muß es versuchen, möglichst ohne lästigen Zeitverzug eine angemessene Näherung an das zu finden, was der Sprecher sagen wollte. Auch ein menschlicher Dolmetscher muß in dieser Situation vielfältige Kompromisse eingehen. Entsprechend ist nicht zu erwarten, daß Verbmobil den Gehalt einer Äußerung nach Semantik (Bedeutung) und Pragmatik (Situationsangemessenheit) verlustlos in die Zielsprache überführt.

Auf absehbare Zeit werden seine Verstehens- und Übersetzungsfähigkeiten nicht ausreichen, die im Szenario unterstellten Kenntnislücken der Gesprächspartner perfekt zu überbrücken. Deswegen spielen Dialoge zur Behebung von Unklarheiten und Mißverständnissen eine wichtige Rolle. Im Projekt werden zwei Arten von Klärungsdialogen untersucht: solche zwischen den Gesprächspartnern, die Verbmobil ebenso unterstützt wie jeden anderen Dialog, und solche zwischen dem System und einem Benutzer, worin es diesen – der Zuverlässigkeit zuliebe in dessen Muttersprache – um zusätzliche, für die Übersetzung erforderliche Information ersucht.

Die Projektziele

Verbmobil baut auf den Ergebnissen mehrerer Vorgängerprojekte auf. Ein japanisches Zentrum für Sprachübersetzung, die ATR Interpreting Telecommunications Research Laboratories in Kioto, konnte Anfang 1993 die erste Phase des Projektes ASURA (*Advanced Speech Understanding and Rendering System of ATR*) zur Übersetzung von Telephongesprächen mit einer erfolgreichen Demonstration abschließen. JANUS, ein Projekt der Carnegie-Mellon-Universität (CMU) in Pittsburgh (Pennsylvania) und C-STAR, ein Gemeinschaftsprojekt von ATR, der CMU und Siemens, haben sich die Übersetzung von telephonisch geführten Auskunftsdialogen – Gespräche zum Zweck der Informationsbeschaffung, in denen eine Seite allein die Initiative hat – zum Ziel gesetzt.

Im Gegensatz dazu geht es bei Verbmobil nicht um Gespräche über das Telephon, sondern in kleinen Räumen von Angesicht zu Angesicht. In einer solchen Situation können Mimik und Gestik zusätzliche Information transportieren. Im Forschungsprogramm sind Untersuchungen darüber vorgesehen, wie menschliche Übersetzer und Dolmetscher sich in ähnlichen Situationen verhalten.

Es geht – ebenfalls im Gegensatz zu den Vorgängerprojekten – von Beginn an nicht um die Verarbeitung abgelesener Texte, sondern um das schwierigere Verstehen inkrementell – das heißt nicht vorgeplant, sondern von einem Augenblick zum nächsten – erzeugter Spontansprache. Solche Äußerungen sind selten grammatisch korrekt. Verbmobil muß deshalb mit abgebrochenen Sätzen, Einschüben, Selbstkorrekturen und ähnlichem umgehen können.

An der Universität Karlsruhe sind sogenannte Müllmodelle entwickelt worden. Nicht bedeutungstragende Geräusche wie Räuspern, Schmatzen, *äh* und *ehm* werden bei der Spracherkennung zunächst wie spezielle Wörter behandelt und für die weitere Analyse aus der Eingabe entfernt. Ein bei Siemens entwickeltes neuartiges Verfahren zur robusten Analyse fehlerhafter Äußerungen korrigiert beispielsweise den Satz „Die Teilnehmer der Sitzung sind heute ehm kommen morgen" automatisch in „Die Teilnehmer der Sitzung kommen morgen".

Der Demonstrator soll eine Diskurssituation beherrschen, in der die Partner ihr nächstes Treffen vereinbaren, wobei sie ihre Terminkalender benutzen (Bild 2). Inzwischen wurden schon mehr als 200 Terminabsprachen mit Versuchspersonen aufgenommen, niedergeschrieben und analysiert. Die gewonnenen Daten werden auf einer eigens gepreßten CD-ROM an die Forschungsgruppen verteilt, die damit ihre Spracherkennungsprogramme trainieren. Bisherige Erfahrungen bei der Datensammlung zeigen, daß man sich für dieses Szenario auf ein Vokabular von etwa 1500 Wörtern beschränken kann.

Für den Forschungsprototyp soll der Anwendungsbereich weiter ausgedehnt werden, beispielsweise auf die Planung einer gemeinsamen Geschäftsreise, wobei die Partner auf verschiedene Termin- und Verkehrspläne zurückgreifen. Man will grundsätzlich voraussetzen, daß das Gesprächsthema eingeschränkt ist, die Dialogziele der Partner vorher bekannt sind, beide das Gespräch kooperativ führen und sehr an dessen erfolgreichem Abschluß interessiert sind.

Auch unter diesen Einschränkungen ist die Aufgabe noch schwer genug. Verbmobil muß eine Reihe von Teilproblemen lösen: das Sprachsignal analysie-

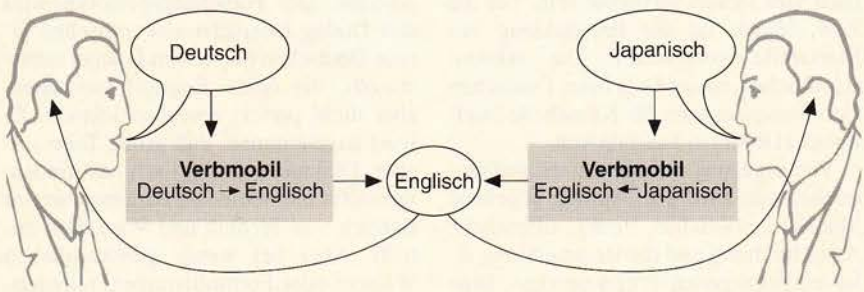


Bild 1: Verhandlungsszenario mit Verbmobil. Ein Deutscher und ein Japaner unterhalten sich auf englisch; sie greifen bei Bedarf auf die Übersetzungshilfe zurück.

Brown: Hello, Mister Schmidtke, how are you?
Schmidtke: Good day, Doctor Brown.
Brown: (Schmatzen) You know (Räuspern), we have to go to (ähm) Switzerland to talk to our partners there.
Schmidtke: ↑Oh ja, gut, nach meinem Terminkalender (Pause), wie wär's im Oktober?↓
Verbmobil: Just looking at my diary, I would suggest October.
Schmidtke: (Pause) I propose from Tuesday the fifth (Pause) no, Tuesday the fourth to Saturday the eighth (Pause), those five days?
Brown: Oh, that's too bad, I'm not free right then. (Pause) I could fit it into my schedule (Schmatzen) the week after, from Saturday to Thursday, the thirteenth.
Schmidtke:↑(ähm) Das paßt echt schlecht bei mir.↓
Verbmobil: That doesn't suit me at all.
Brown: Oh.
Schmidtke: ↑Als Ausweichmöglichkeit (Pause) bei mir kommt wieder in Frage (Pause) zwischen dem fünfzehnten und dem neunzehnten.↓
Verbmobil: Alternatively, I would be available again between the fifteenth and the nineteenth.
Brown: I'm already booked then.
Schmidtke: ↑Zur Not können wir es auf den ersten verlegen.↓
Verbmobil: If necessary, we could move it to the first.
Brown: I can't make it at the beginning; I'm on vacation then.
Schmidtke:↑Ach, dann können wir den Oktober ja komplett vergessen, aber nicht den November.↓
Verbmobil: Oh, so we can totally forget about October, but not November.
Brown: (Räuspern) Could we do it in the first half?
Schmidtke: Die erste Hälfte, (Räuspern) das ist schlecht.
Verbmobil: The first half doesn't suit me at all.
Brown: How is the week of October thirty-first to the fourth of November?
Schmidtke: ↑Ist Allerheiligen nicht ein Feiertag bei Ihnen?↓
Verbmobil: Isn't All Saints' Day a holiday with you?
Brown: No, not for us.
Schmidtke: OK, let's do that.

Bild 2: Referenzdialog für den Demonstrator. Bei ↑ wird die Übersetzungskomponente von Verbmobil ein-, bei ↓ ausgeschaltet.

ren, daraus Hypothesen für Wörter gewinnen, die Satzstruktur erkennen, daraus unter Einbeziehung von Wissen über Gesprächsthema und -kontext eine Darstellung der Bedeutung erzeugen, diese in einem Übersetzungsprozeß in die Zielsprache Englisch überführen, eine Satzstruktur und daraus schließlich einen gesprochenen Text erzeugen. Anerkannte Theorien aus den Bereichen Künstliche Intelligenz, Computerlinguistik, Spracherkennung, Neuroinformatik und Übersetzungswissenschaft tragen zu einem interdisziplinären Ansatz bei. Für alle Teilprobleme gibt es bereits erste Software-Lösungen, deren komplexes, mehrfach rückgekoppeltes Zusammenspiel jedoch noch intensiver Forschung bedarf.

Offensichtlich muß man gerade für ein System wie Verbmobil besonders auf das Verhältnis von Verarbeitungsgeschwindigkeit und Qualität der Übersetzung achten. Um einer schritthaltenden Übersetzung von Äußerungen möglichst nahe zu kommen, ohne den natürlichen Dialogfluß durch lange Wartezeiten zu stören, sollten Spracherkennung und -analyse nicht tiefer als nötig gehen sowie die Übersetzung so flach wie möglich und der Generierungsprozeß so früh wie möglich erfolgen. Das bedeutet, daß die Hauptkomponenten des Systems nicht erst abwarten dürfen, bis alle relevanten Informationen beisammen sind, sondern jeden Teil des Eingabestroms berarbeiten müssen, sowie er eintrifft (inkrementelle Arbeitsweise).

Bei der Übersetzung ist dieses Konzept eng verknüpft mit der Idee der variablen Verarbeitungstiefe. Zur Analyse der Bedeutung einer Äußerung muß Verbmobil eine Hierarchie von Repräsentationen benutzen, die in den oberflächennahen Ebenen in verschiedener Weise unvollständig sein dürfen. Zu jeder Ebene muß es eine spezielle Inferenzkomponente geben: ein Programmsegment, das die vorliegende Aussage nach den Regeln der Logik umformen, insbesondere aus ihr Schlüsse ziehen kann.

Andererseits ist bereits eine oberflächliche Repräsentation für die nachgeschaltete Übersetzungskomponente häufig eine brauchbare Arbeitsgrundlage. Beispielsweise ist die Zweideutigkeit eines Satzes wie „Der Mann sieht die Frau mit dem Teleskop" nur mit beträchtlichem Vorwissen oder überhaupt nicht aufzulösen. Häufig hat jedoch die Zielsprache eine genau gleichartige Mehrdeutigkeit, so daß deren Auflösung dem menschlichen Dialogpartner überlassen bleiben kann. Für Verbmobil bedeutet dies, daß der Sprachgenerator auch aus unvollständig spezifizierten Eingaben – wenn etwa wegen schlechter Aussprache Singular und Plural nicht zu unterscheiden sind – und sogar disjunktiven semantischen Strukturen (entsprechend Aussagen der Form „*A* oder *B*", weil die Analysekomponente – bislang – keine der beiden ausschließen konnte) Dialogbeiträge in der Zielsprache zustande bringen muß.

Durch Kopplung verschiedener Satzbaupläne im Deutschen und im Englischen wurde in Verbmobil bereits ein Modul zum schnellen Transfer von Redewendungen entwickelt. Aus dem Satz „Lassen Sie uns doch solch einen Termin ausmachen" wird in 0,8 Sekunden „Let us just fix such a date".

Der Erfolg eines so anspruchsvollen Übersetzungsprojekts hängt offensichtlich von internationaler Kooperation ab. Deshalb ist unter anderem eine intensive Zusammenarbeit mit ATR in Kioto geplant. Im März 1993 hat dieses Zentrum ein neues Projekt gestartet, das bis März 2000 dauern soll. Wie in Verbmobil wird dabei die Übersetzung spontansprachlicher Dialoge angestrebt. Während in ASURA jeder neue Sprecher vor dem Dialogbeginn noch etwa zehn vordefinierte Wörter für die Sprecheradaption vorlesen muß, soll in dem Folgeprojekt wie in Verbmobil eine dynamische Sprecheradaption während des Dialoges stattfinden. Hauptgebiete der geplanten Kooperation sind die Datensammlung, Spracherkennungsmodule und linguistische Wissensquellen für die japanische Sprache.

In Korea wird am Center for Artificial Intelligence Research (CAIR) des Korea Advanced Institute of Science and Technology (KAIST) seit 1991 das auf 15 Jahre angelegte Dialogübersetzungsprojekt ATI (*Automatic Telephony Interpretation*) durchgeführt, bei dem in der ersten, auf sieben Jahre geplanten Phase – sehr ähnlich zu Verbmobil – die Übersetzung koreanischer Spontansprache ins Englische im Vordergrund steht.

Für Arbeitspakete zum Englischen sind Kooperationen mit dem erwähnten CSLI sowie mit einer Forschergruppe an der Carnegie-Mellon-Universität in Pittsburgh (Pennsylvania) vorgesehen.

Dr. Wahlster ist Professor für Informatik an der Universität des Saarlandes und Wissenschaftlicher Direktor am Deutschen Forschungszentrum für Künstliche Intelligenz Saarbrücken/Kaiserslautern.