WIRTSCHAFTSINFORATIK, Bd. 41, Nr. 6, 1999, S. 328–547

Spracherkennung - Stand der Technik, Einsatzmöglichkeiten und Perspektiven

Florian Gallwitz, Heinrich Niemann, Elmar Nöth

Universität Erlangen-Nürnberg
Lehrstuhl für Mustererkennung (Informatik 5)
Martensstr. 3, 91058 Erlangen
email: {gallwitz,niemann,noeth}@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de

Zusammenfassung

Das automatische Erkennen und Verstehen von gesprochener Sprache ist schon seit Jahrzehnten Gegenstand intensiver Forschungsarbeit. Seit wenigen Jahren sind nun erste Anwendungen auf Grundlage der hieraus hervorgegangen Technologie kommerziell verfügbar, und es zeichnet sich ab, dass spracherkennende Systeme schon in wenigen Jahren auch aus dem betrieblichen Alltag nicht mehr wegzudenken sind. In diesem Beitrag werden neben den technologischen Grundlagen von Spracherkennungssystemen die Möglichkeiten und auch die Grenzen dieser Technologie dargestellt. Heutige und zukünftige Anwendungsmöglichkeiten, von einfachen Kommandosteuerungen über Diktiersysteme bis hin zu natürlichsprachlichen Dialog- und Übersetzungssystemen, werden diskutiert.

1 Einleitung

Maschinen, die sich mit Menschen unterhalten können, sind fester Bestandteil jedes Science-Fiction-Films. Schon vor über dreißig Jahren entwarf Arthur C. Clarke in seinem von Stanley Kubrick verfilmten Roman „2001 - Odyssee im Weltraum“ die Vision des Computers HAL, der wie selbstverständlich mit den Menschen an Bord des Raumschiffes sprachlich kommuniziert.

Das im Roman erwähnte Datum seiner Inbetriebnahme, der 12. Januar 1997, gab weltweit Anlass für eine Bestandsaufnahme der Forschung im Bereich der künstlichen Intelligenz. Wenn auch einzelne Fähigkeiten von HAL, beispielsweise die des Schachspiels oder die der Steuerung und Navigation eines Raumfahrzeugs, bereits heute Realität geworden sind, so erscheint das „(un)menschliche“ Verhalten von HAL heute utopischer denn je. Auch von der Möglichkeit einer verbalen Kommunikation, wie sie im Roman beschrieben ist, sind wir heute noch weit entfernt.

Dennoch hat es gerade im Bereich des automatischen Verstehens von gesprochener Sprache in den vergangenen Jahren erhebliche Fortschritte gegeben, die in Form von ersten Produkten auch dem Endverbraucher zugänglich gemacht werden. Hierzu gehören

EVAR:	*Guten Morgen! Hier ist die automatische Intercityauskunft. Welche Information benötigen Sie?*
Anrufer:	*Ja guten Morgen, ich bräuchte morgen abend so gegen sechs eine Verbindung nach Hannover bitte.*
EVAR:	*Sie wollen am Freitag dem 27. Februar gegen 18 Uhr nach Hannover fahren. Wo wollen Sie abfahren?*
Anrufer:	*In Nürnberg.*
EVAR:	*Sie wollen von Nürnberg abfahren?*
Anrufer:	*Ja, genau.*
EVAR:	*Nach Hannover fahren Sie ab Nürnberg Hauptbahnhof um 18:32. Ankunft in Hannover Hauptbahnhof um 21:30. Ich wiederhole: Abfahrtszeit 18:32, Ankunftszeit 21:30. Soll ich die Verbindung wiederholen?*
Anrufer:	*Nein, vielen Dank. Auf Wiederhören!*
EVAR:	*Auf Wiederhören.*

Abbildung 1: Beispieldialog mit dem automatischen Zugauskunftssystem EVAR

beispielsweise Mobiltelefone, die bei der Nennung eines Namens automatisch die entsprechende Nummer wählen, und Diktiersysteme, die das Gesprochene mit durchaus überzeugenden Ergebnissen in geschriebenen Text umwandeln. Sogar natürlichsprachliche Dialoge mit Computern sind heute bereits möglich. So existieren beispielsweise telefonische Fahrplanauskunftssysteme, die das Verhalten einer menschlichen Auskunftsperson zu simulieren versuchen. Ein typischer Dialog mit dem an unserem Lehrstuhl entwickelten Auskunftssystem EVAR für Intercity-Zugverbindungen [9] ist in Abbildung 1 wiedergegeben.

Bevor wir die wesentlichen technologischen Grundlagen eines solchen Sprachdialogsystems erläutern, werden wir in Abschnitt 2 zunächst auf einige Leistungskriterien eingehen, mit deren Hilfe sich die Komplexität einer von einem Spracherkenner zu bewältigenden Aufgabe in verschiedenen Anwendungsfeldern beurteilen lässt. Anschließend werden in Abschnitt 3 die technologischen Grundlagen des eigentlichen Spracherkennungsvorgangs erläutert, also die Umsetzung des Gesprochenen in eine Folge von Wörtern. In Abschnitt 4 beschreiben wir einige weitere Teilprobleme, die neben der eigentlichen Spracherkennung noch zu lösen sind, bevor es zu einem quasi-natürlichen Dialogverhalten kommen kann. Schließlich werden wir in Abschnitt 5 eine Reihe von heute bereits existierenden und zukünftigen Anwendungsmöglichkeiten von Spracherkennungstechnologie diskutieren.

2 Leistungskriterien für die automatische Spracherkennung

Forscher im Bereich der automatischen Spracherkennung sehen sich häufig mit Aussagen konfrontiert, wie: „*Spracherkennung, wieso? Das gibt's doch schon. Hab' ich mir neulich bei ALDI gekauft*". Der verbreitete Eindruck, dass dieses Problem mehr oder weniger gelöst sei, hängt damit zusammen, dass die Leistungsfähigkeit von Diktiersystemen bei der Eingabe von Texten unter bestimmten Voraussetzungen durchaus mit der von geübten Computerbenutzern vergleichbar sein kann. Die Anwendungsmöglichkeiten dieser Technologie sind jedoch weitaus vielfältiger, und die Anforderungen an die eigentliche Spracherkennungs-Komponente können sich je nach Anwendungssituation stark unterscheiden.

Während Diktiersysteme in ruhiger Umgebung beispielsweise mit Vokabulargrößen von über 60.000 Wörtern *Erkennungsraten* von bis zu 95 Prozent erzielen (d. h. im Mittel ein falsch erkanntes Wort alle 20 Wörter; man spricht auch von einer *Fehlerrate*, hier 5 Prozent), kann bereits die Erkennung von einfachen Ziffernfolgen in einem fahrenden Auto wegen der Fahrgeräusche große Probleme bereiten.

Betrachtet man eine Reihe von unterschiedlichen Anwendungen für Spracherkenner, so kann man allgemeine Leistungsmerkmale erkennen, die die Komplexität einer von einem Spracherkenner zu bewältigenden Aufgabe bestimmen. In Anlehnung an [21] lassen sich fünf Leistungsachsen definieren:

- **Sprecherabhängigkeit:** Spracherkennung kann *sprecherunabhängig* oder *sprecherabhängig* erfolgen, wobei sprecherabhängige Erkennung mit einer erheblich höheren Genauigkeit möglich ist. *Sprecheradaptive* Systeme bilden hier einen Mittelweg, indem sie sich allmählich an die Stimme ihres Benutzers anpassen. Es hängt stark von der Applikation ab, inwieweit eine sprecherabhängige Erkennung realisierbar ist. Während dem Benutzer eines Diktiersystems (heute noch) das Vorlesen einiger Übungssätze zugemutet werden kann, ist dies zum Beispiel bei einem Fahrplanauskunftssystem oder gar bei einem sprachgesteuerten Getränkeautomaten nicht praktikabel.

- **Sprechart:** Die Unterscheidung zwischen *diskreter* und *kontinuierlicher Sprache* verliert, zumindest im Zusammenhang mit Diktiersystemen, zunehmend an Bedeutung: Die kurzen Sprechpausen zwischen den Wörtern, die bis vor etwa zwei Jahren noch von den Benutzern von „diskreten" Diktiersystemen verlangt wurden, werden von den „kontinuierlichen" Systemen nicht mehr gefordert. Diese Pausen erleichtern die Bestimmung der Wortgrenzen und verbessern damit das Erkennungsergebnis, erfordern aber eine sehr unnatürliche Sprechweise des Benutzers. Einen Spezialfall stellen die *Einzelworterkenner* dar, die voraussetzen, dass nur ein einzelnes Wort gesprochen wird.

  Noch erheblich schwieriger als der Umgang mit kontinuierlicher Sprache ist dagegen die Erkennung von *spontaner Sprache*. Darunter versteht man Äußerungen,

die nicht abgelesen sind, und die sich der Sprecher nicht — wie im Falle einer Diktieranwendung — vor dem Sprechen zurechtgelegt hat. Typisch für spontane Sprache sind ungrammatische Sätze, äh-s und ähm-s, Pausen, Abbrüche, Versprecher, Verschleifungen und Wiederholungen, die von menschlichen Hörern normalerweise sehr gut verabeitet werden, die jedoch die automatische Verarbeitung drastisch erschweren [8]. So muss zum Beispiel in einem Fahrplanauskunftssystem mit der folgenden Anfrage gerechnet werden: „*äh ja hallo also ähm nach Hamburg wollt' ich fahr'n ab München Pasing so gegen acht gegen zwanzig Uhr heut' a'md.*“ Zusätzliche Schwierigkeiten ergeben sich noch durch Sprecher mit regionalem Dialekt oder ausländischem Akzent.

Anzumerken wäre hier, dass auch der Mensch Spontansprache nur aufgrund von jahrelangem Training so scheinbar leicht verarbeiten kann. Man kann das für sich selbst verifizieren, wenn man zuerst leise und dann laut die Verschriftung einer spontanen Äußerung liest.

- **Wortschatz:** Der Einfluss der Vokabulargröße auf die Schwierigkeit des Spracherkennungsproblems ist offensichtlich, allerdings wirkt sich diese in aller Regel mehr auf die erforderliche Rechenleistung aus, als auf die zu erwartende Fehlerrate (sehr große Wortschätze erfordern zudem hochkomplexe und ausgefeilte Suchalgorithmen). Von wesentlich größerer Bedeutung für die Fehlerrate ist jedoch die grammatische Komplexität (s.u.). Das Problem, 500 Eigennamen ohne Kontextinformation zu unterscheiden, kann in dieser Hinsicht wesentlich schwieriger sein, als einen grammatisch korrekten Text mit einem Vokabular von 60.000 Wörtern zu erkennen.

  In einigen kommerziellen Systemen wird zwischen *aktivem Vokabular* und dem *Gesamtvokabular* unterschieden; in diesem Falle wird z. B. in einem bestimmten Dialogschritt nur ein Teil der Wörter erlaubt, oder ein spezielles Inventar an Fachbegriffen wird in einem Diktiersystem nur auf Wunsch in den Erkennungswortschatz aufgenommen.

- **Grammatische Komplexität** oder **Perplexität**: Nicht jedes Wort aus dem Wortschatz tritt an jeder Position einer Äußerung mit der gleichen Wahrscheinlichkeit auf. So ist es z. B. sehr wahrscheinlich, dass nach den Wörtern „*Guten Tag, mein Name*“ das Wort „*ist*“ folgen wird, und auf dieses Wort wiederum ein Eigenname. Je besser sich die Wörter selbst ohne Kenntnis des akustischen Signals bereits aus der Anwendung und aus dem Kontext vorhersagen lassen, desto einfacher ist naturgemäß die Aufgabe des Spracherkenners. Ein entscheidendes Maß für die Schwierigkeit eines Spracherkennungsproblems ist daher die sogenannte *Perplexität*, die angibt, wieviele Wörter im Mittel in Frage kommen, wenn die Vorgängerwörter bereits bekannt sind.

  Mittels einer statistischen Grammatik lässt sich die Wahrscheinlichkeit für eine gegebene Wortfolge berechnen. Eine solche Grammatik kann entweder explizit vorge-

geben werden, beispielsweise in einer Anwendung, in der nur Ziffernfolgen erkannt werden sollen, oder sie kann aus einer großen Menge geschriebenen Textes automatisch erlernt werden, wie dies z. B. im Falle von Diktiersystemen geschieht.

Eine Grammatik reduziert die Zahl der Erkennungsfehler drastisch, solange der Benutzer sich innerhalb der vorgesehenen Anwendung bewegt. Ein Spracherkenner in einem Fahrplanauskunftssystem wird allerdings i.d.R. auch in der Frage nach dem Wetter des folgenden Tages eine Fahrplananfrage erkennen, und ein Diktiersystem für Juristen wird einen romantischen Liebesbrief mit juristischen Floskeln und Fachtermini anreichern (siehe Abschnitt 3.3).

- **Eingabemedium**: Von großer Bedeutung für automatische Spracherkenner ist der sogenannte „Eingabekanal"; hierzu gehören das Mikrophon und, z.B. im Falle einer Telefonanwendung, auch die Art der Übertragung des Signals (Festnetztelefon vs. Handy). Beispielsweise lassen sich aufgrund der geringen Bandbreite des Telefonkanals die Konsonanten 'f' und 's' in einem Telefongespräch praktisch nicht unterscheiden (was zum Beispiel beim Buchstabieren über Telefon offenbar wird). Als optimales Aufnahmemedium gelten hochwertige Nahbesprechungsmikrophone oder *Headsets*, bei denen das Mikrophon in der Nähe des Mundwinkels positioniert wird. Dennoch wird der Einfluss der Qualität des Mikrophons oft überschätzt; viel wichtiger ist es, dass der Spracherkenner mit Daten *trainiert* wurde (siehe Abschnitt 3.2), die nach Möglichkeit mit dem gleichen Mikrophon unter möglichst ähnlichen akustischen Bedingungen aufgenommen wurden.

  Besonders schwierig wird es, wenn das Mikrophon sich nicht mehr direkt am Mund des Sprechers befindet, z. B. bei Anwendungen im Auto oder bei der Bedienung von mobilen Robotern. Hintergrundgeräusche (z. B. Fahrgeräuche im Auto oder Geräusche in einer Bahnhofshalle), oder gar mehrere Sprecher, die gleichzeitig reden („Cocktailparty-Effekt", oder auch das typische Verhalten der Teilnehmer von Fernseh–Diskussionsrunden) erschweren die Spracherkennung weiter oder machen sie nahezu unmöglich.

Grundsätzlich ist es so, dass Spracherkennungssysteme, die sich in einem oder mehreren der genannten Leistungsmerkmale im „schwierigen" Bereich bewegen, dies dadurch kompensieren, dass der Anwender in Bezug auf die anderen Merkmale Abstriche machen muss. Dies wird in Abschnitt 5 an Hand einiger konkreter Applikationen für Spracherkennungstechnologie verdeutlicht.

3 Wie funktioniert automatische Spracherkennung?

In diesem Abschnitt wird die prinzipielle Funktionsweise von automatischen Spracherkennungssystemen erläutert. In den vergangenen Jahren haben sich einige wenige Verfahren herauskristallisiert, auf deren Grundlage nahezu alle heutigen Spracherkennungssysteme basieren. Ihnen gemein ist, dass der Spracherkenner zunächst an Hand von Beispielen

„trainiert“ wird, d. h. die Aussprache bestimmter Laute oder Wörter wird automatisch erlernt.

Zunächst wird die Vorverarbeitungsphase beschrieben, in der das Sprachsignal für den eigentlichen Erkennungsvorgang aufbereitet wird. Anschließend werden die beiden verbreitetsten Verfahren zur Wortmodellierung und -erkennung vorgestellt. Die Einschränkung der möglichen Wortfolgen mit Hilfe von Sprachmodellen wird im darauffolgenden Abschnitt erörtert. Zuletzt wird kurz auf das Problem eingegangen, zu einer vorliegenden Äußerung an Hand der gegebenen Wort- und Sprachmodelle die wahrscheinlichste Wortfolge möglichst effizient zu suchen.

Für eine wesentlich ausführlichere Darstellung der Funktionsweise automatischer Spracherkennungssysteme und Hinweise auf weiterführende Literatur verweisen wir auf [19, 17, 11].

3.1 Vorverarbeitung und Merkmalberechnung

Am Beginn der Verarbeitungskette werden die in Form von Luftdruckschwankungen vorliegenden Schallwellen mittels eines Mikrophons in ein elektrisches Signal umgewandelt. Um dieses Signal einer Verarbeitung durch den Computer zugänglich zu machen, wird es zunächst digitalisiert, d. h. in eine Folge von digitalen *Abtastwerten* verwandelt. Dabei ensteht eine enorme Datenflut; so entprechen eine Minute Sprache z. B. bei einer üblichen *Abtastrate* von 16 kHz und 16 bit pro Abtastwert bereits knapp 2 Megabyte.

Aus dem digital vorliegenden Signal gilt es nun, Informationen über die jeweils gesprochenen Laute (bzw. *Phoneme*, die kleinsten bedeutungsunterscheidenden Lauteinheiten) zu gewinnen. Dies gelingt am besten im *Spektralbereich*: Phoneme, von denen es im Deutschen etwa 40 gibt, zeichnen sich i.d.R. durch ein charakteristisches Spektrum aus (eine Zerlegung eines Klanges in seine verschiedenen spektralen Bereiche kann man z. B. bei einer hochwertigeren Stereoanlage beobachten, bei der die Energie in verschiedenen Frequenzbändern mit Leuchtdioden angezeigt wird). Als günstig hat es sich herausgestellt, dieses Spektrum etwa alle 10 Millisekunden über ein kurzes *Zeitfenster* von ca. 25 Millisekunden zu berechnen (die mittlere Lautdauer liegt bei etwa 70 Millisekunden). Mittels geeigneter Transformationen, die sich an der Funktionsweise des menschlichen Gehörs orientieren, wird das so errechnete Spektrum, zusammen mit anderen geeigneten Kennzahlen des Signalabschnitts wie etwa der Signalenergie, in einen sogenannten *Merkmalsvektor* mit etwa 20 bis 50 Komponenten umgesetzt (bei den Komponenten des Vektors handelt es sich meist um die *Mel-Cepstrum-Koeffizienten* und die komponentenweise 1. und 2. Ableitung).

Jeder solche Merkmalsvektor repräsentiert charakteristische Eigenschaften des zugehörigen kurzen Signalabschnitts: Ähnliche Laute führen zu ähnlichen Merkmalsvektoren, und deutlich unterscheidbare Laute führen zu deutlich verschiedenen Merkmalsvektoren. An die Stelle des Signals tritt also nun eine Folge von Merkmalsvektoren im Zeittakt von etwa 10 Millisekunden. Die Datenmenge reduziert sich dabei gegenüber dem ursprünglichen Signal, je nach Größe des Merkmalsvektors, etwa um den Faktor zwei bis

vier.

3.2 Wortmodellierung und -klassifikation

Die Aufgabe besteht nun darin, in der Folge von Merkmalvektoren das gesprochene Wort bzw. die gesprochene Wortfolge zu bestimmen (zu *klassifizieren*). Hierfür kommen im wesentlichen zwei Ansätze in Betracht, die unter den Begriffen *DTW (Dynamic Time Warping)* und *HMM (Hidden-Markov-Modell)* bekannt sind. Der ältere DTW-Ansatz eignet sich vor allem für sehr einfache, sprecherabhängige Einzelworterkennung, beispielsweise zum Abrufen von gespeicherten Telefonnummern in einem Mobiltelefon. Der leistungsfähigere HMM-Ansatz wird in nahezu allen komplexeren Spracherkennungssystemen verwendet.

Im Falle der sprecherabhängigen Einzelworterkennung besteht das Erkennungsproblem darin, aus einer Menge von zuvor prototypisch gesprochenen Wörtern (*Referenzmustern*) das ähnlichste Wort auszuwählen. Es ist also lediglich ein Vergleich jeweils zweier (nicht notwendigerweise gleich langer) Folgen von Merkmalsvektoren vorzunehmen, und ein geeignetes *Abstandsmaß* zu bestimmen. Auf der Ebene einzelner Merkmalsvektoren kann beispielsweise der Euklidische Abstand verwendet werden. Bei der Zuordnung einzelner Merkmalsvektoren der gesprochenen und der prototypischen Merkmalsvektorfolge müssen *nichtlineare zeitliche Verzerrungen*, wie z. B. Dehnungen einzelner Laute, zugelassen werden. Beim *DTW-Algorithmus* wird nun unter geeigneten Nebenbedingungen auf sehr effiziente Weise eine optimale Zuordnung zwischen zwei Merkmalsvektorfolgen ermittelt. Auf diese Weise wird ein Abstandsmaß zwischen dem gesprochenen Wort und jedem zuvor gespeicherten Wort berechnet, und die für das Wort mit dem geringsten Abstand vorgesehene Aktion (beispielsweise das Wählen einer Telefonnummer) wird eingeleitet.

Hidden-Markov-Modelle ermöglichen weitaus leistungsfähigere Spracherkennungssysteme. Sie bilden die Grundlage für eine *stochastische*, wahrscheinlichkeitstheoretisch fundierte Herangehensweise an die Spracherkennung. Die Produktion eines Wortes durch einen menschlichen Sprecher wird als ein zweistufiger Zufallsprozess aufgefasst, der sowohl die zeitlichen Verzerrungen, als auch die lautlichen Variationsmöglichkeiten umfasst. Die hierbei involvierten Wahrscheinlichkeiten können an Hand von i.d.R. umfangreichen *Trainingsdaten* automatisch erlernt werden. Der HMM-Formalismus bietet hierfür die mathematische Grundlage, und er ermöglicht es in der Erkennungsphase, die Wahrscheinlichkeit für das Vorliegen eines bestimmten Wortes (oder einer bestimmten Wortfolge) zu einer Folge von Merkmalsvektoren zu berechnen. Das *wahrscheinlichste Wort* (bzw. die *wahrscheinlichste Wortfolge*) wird i.d.R. als Erkennungsergebnis ausgegeben.

Hidden-Markov-Modelle erlauben es, phonetisches Wissen über die Aussprache von Wörtern, wie es etwa in Wörterbüchern zu finden ist, mit empirischem Wissen aus dem Sprachmaterial zu kombinieren, das zum *Training* des Spracherkenners herangezogen wird. Auf diese Weise kann ein Spracherkenner zum Beispiel automatisch erlernen, dass der Buchstabe 'n' im Wort „*haben*“ häufig als 'm' gesprochen wird.

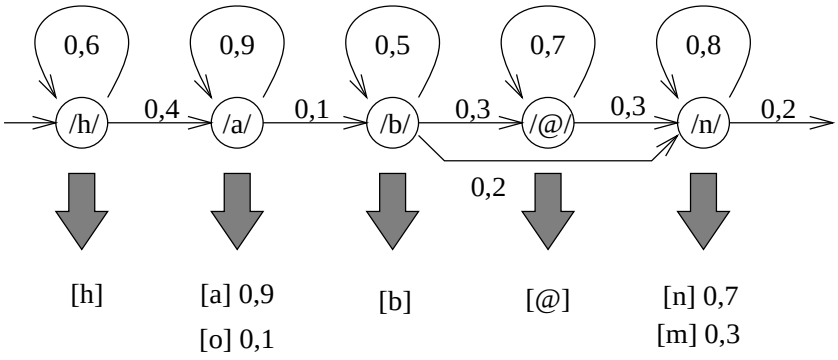

Abbildung 2: Einfaches HMM für das Wort „*haben*“. Das Modell enthält *Emissionswahrscheinlichkeiten*, die die Wahrscheinlichkeit für das Auftreten bestimmter Laute (z.B. [m] vs. [n]) in bestimmten Modellzuständen bestimmen, und *Übergangswahrscheinlichkeiten*, die die zeitlichen Variationsmöglichkeiten erfassen.

In Abbildung 2 ist ein einfaches HMM für das Wort „*haben*“ schematisch dargestellt. Die Struktur des Modells wird an Hand der *phonetischen Umschrift* des Wortes festgelegt, die Modellparameter werden dagegen automatisch erlernt. Auch die Erkennung von nicht in den Trainingsdaten enthaltenen Wörtern ist möglich, indem Sie nach dem Baukastenprinzip aus sogenannten *Wortuntereinheiten* zusammengesetzt werden. Hierfür kommen beispielsweise einzelne Laute in Betracht. Durch Einschränkung der Nachbarlaute während des Trainings können sogenannte *Koartikulationseffekte*, die gegenseitige Beeinflussung von Lauten durch die benachbarten Laute, berücksichtigt werden. So kann man z.B. eine spezialisierte Lauteinheit bereitstellen, die nur auf denjenigen ‘b’s in den Trainingsdaten trainiert wird, deren linker Nachbarlaut ein ‘a’ ist. Diese Einheit kann dann beispielsweise innerhalb der Wortmodelle für „*haben*“ und „*Tabak*“ zur Modellierung des Lauts ‘b’ herangezogen werden, nicht jedoch im Modell für das Wort „*Rebe*“; hier verwendet man ein ‘b’ mit linkem Nebenlaut ‘e’.

3.3 Sprachmodellierung

Unter dem Begriff *Sprachmodellierung* (engl. *language modeling*) fasst man Verfahren zusammen, die dem Spracherkenner Wissen über die Wahrscheinlichkeit von bestimmten Wortfolgen vermitteln, und ohne die eine befriedigende Erkennungsleistung in den meisten Fällen nicht möglich wäre. Unabhängig von dem tatsächlich Gesprochenen erhält jede Wortfolge *a priori* eine Wahrscheinlichkeit zugewiesen, die dann mit den auf Basis der Hidden-Markov-Modelle errechneten akustischen Wahrscheinlichkeiten kombiniert wird.

In einer Diktieranwendung ist es beispielsweise sehr unwahrscheinlich, der Wortfolge „*Säge ehrte Frau schnitt*“ zu begegnen; viel wahrscheinlicher ist dagegen die Wortfolge „*Sehr geehrte Frau Schmidt*“. Auf diese Weise können zum einen akustisch nicht unterscheidbare Wörter korrekt erkannt werden, zum anderen können Fehler ausgeglichen

werden, die durch ungenaue Aussprache oder ungenaue Modellierung der Wörter entstehen würden.

Sprachmodelle lassen sich z. B. für Diktiersysteme an Hand großer Mengen geschriebenen Textes automatisch erlernen. Das Prinzip besteht darin, die Häufigkeit einzelner Wörter und Worttupel (z. B. Wortpaare) zu zählen. Je häufiger ein Worttupel im *Trainingstext* auftritt, desto höher wird die Wahrscheinlichkeit, die ihm durch das Sprachmodell zugewiesen wird. Mit Hilfe von geeigneten Interpolations- und Glättungsverfahren erhält man auch für Wortkombinationen, die im Trainingstext nie aufgetreten sind, plausible Wahrscheinlichkeiten. In Diktiersystemen werden i.d.R. sogenannte *Trigramme* verwendet, das sind Sprachmodelle, die sich auf die Auszählung der Worttupel von bis zu drei Wörtern Länge stützen. Je besser die Trainingstexte und die gesprochenen Sätze übereinstimmen, desto höher wird die der gesprochenen Wortfolge zugewiesene Sprachmodellwahrscheinlichkeit, und desto genauer wird die Erkennung.

Alternativ kann ein Sprachmodell auch durch eine formale Grammatik vorgegeben werden. Dies bietet sich z. B. bei einfachen Kommandosteuerungen an, in denen der Benutzer die möglichen Kommandos und Wortkombinationen kennt. Ein Beispiel hierfür wäre eine Grammatik, die u.a. Kommandos der Form „*NAME anrufen*“ oder „*NAME löschen*“ gestattet, wobei *NAME* einer von mehreren gespeicherten Eigennamen sein kann. An die Stelle der automatisch erlernten Sprachmodellwahrscheinlichkeiten treten hierbei in aller Regel die beiden „Wahrscheinlichkeitswerte“ Eins (zulässige Wortfolge) und Null (nicht zulässige Wortfolge).

3.4 Suche

Bei der Bestimmung der am wahrscheinlichsten gesprochenen Wortfolge zu einer gegebenen Folge von Merkmalvektoren handelt es sich um ein komplexes Suchproblem. Mit der Länge des Satzes nimmt die Zahl der möglichen Wortfolgen exponentiell zu. Bereits bei einer Wortschatzgröße von 1000 Wörtern und einer Satzlänge von 8 Wörtern ergeben sich 10^{24} kombinatorisch mögliche Wortfolgen. Erschwerend kommt hinzu, dass die Wortgrenzen in aller Regel nicht feststellbar sind, ohne die Wörter selbst bereits bestimmt zu haben.

Effiziente Suchverfahren wie die *Viterbi-Suche*, die daraus abgeleitete *Strahlsuche* (*beam search*) sowie der $A^\star$-Algorithmus werden in unterschiedlichen Kombinationen eingesetzt, um dieses Problem in Echtzeit lösen zu können. Bei großen Wortschätzen kann zudem durch Anordnung des HMM–Lexikons in Form eines Baumes, an dessen Unterbäumen Wörter mit übereinstimmenden Wortanfängen angeordnet sind, der Rechenaufwand in Grenzen gehalten werden.

4 Natürlichsprachlicher Dialog und Übersetzung

Die im vorangegangenen Abschnitt erläuterte Technologie ermöglicht eine mehr oder weniger genaue Umsetzung der gesprochenen Sprache in eine textuelle Repräsentation,

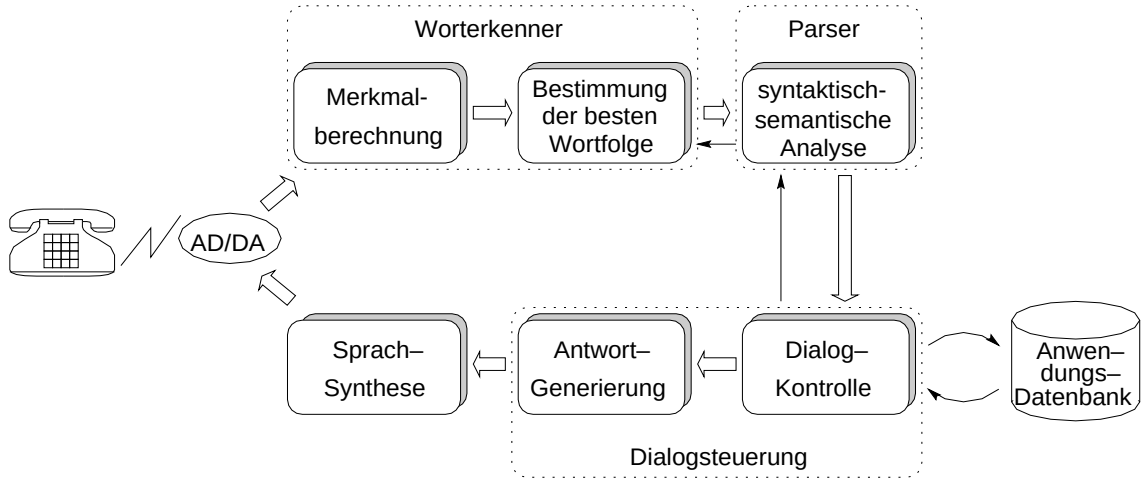

Abbildung 3: Aufbau eines automatischen Dialogsystems zur Informationsabfrage (ohne prosodische Analyse).

beispielsweise in eine sogenannte „Wortkette“ oder einen „Wortgraphen“. Während die Wortkette im Falle eines Diktiersystems bereits das erwünschte Ergebnis darstellt, erwartet der Benutzer eines „sprachverstehenden” Systems eine geeignete Systemreaktion. Im Falle eines Kommandoerkenners oder eines einfachen Menüsystems ist die Umsetzung des erkannten Schlüsselwortes in die entsprechende Systemreaktion relativ trivial.

Erheblich komplizierter wird es, wenn ein „intelligentes“ Dialogverhalten erwartet wird, mit dem das Verhalten eines menschlichen Gesprächspartners imitiert werden soll. Eine typische Architektur für ein natürlichsprachliches Dialogsystem zur Informationsabfrage ist in Abbildung 3 dargestellt.

Bereits die Interpretation einer Datums- und/oder Uhrzeitangabe (z. B. „*diesen Donnerstag am späten Nachmittag so ab fünf Uhr*“) erfordert eine relativ komplexe *syntaktisch-semantische Analyse* des Spracherkennungsergebnisses. Sollen darüberhinaus z. B. die beiden verschieden intonierten Äußerungen „*Natürlich nicht am Montag*“ und „*Natürlich nicht. Am Montag*“ unterschieden werden, so benötigt man neben der Spracherkennung noch eine sogenannte *prosodische Analyse* des Sprachsignals. Weiterhin ist in jedem Falle eine *Dialogsteuerung* notwendig, die dafür verantwortlich ist, dass das System in sinnvoller Weise auf die Benutzeräußerung reagiert bzw. den Benutzer in geeigneter Weise durch den Dialog führt. Schließlich und endlich erwartet der Benutzer in aller Regel auch, dass das System sich in natürlicher, gesprochener Sprache ausdrücken kann.

Ähnliche Teilprobleme ergeben sich, neben einigen zusätzlichen Anforderungen, auch im Falle einer automatischen Übersetzung gesprochener Sprache. Im Rahmen des BMBF–Projekts VERBMOBIL [5, 2] wurde unter Beteiligung unseres Lehrstuhls bereits der Prototyp eines automatischen Übersetzungssystems realisiert, der einen spontansprachlichen Dialog zwischen einem deutschsprachigen und einem englisch- bzw. japanischsprachigen Gesprächspartner gestattet. Die Gespräche müssen sich bisher allerdings auf die Bereiche „Terminabsprache“ und „Reiseplanung“ beschränken.

Die obengenannten Teilprobleme und existierende Lösungsansätze werden in den folgenden Abschnitten kurz erläutet.

4.1 Syntaktisch-semantische Analyse

Aufgabe dieses Verabeitungsschrittes ist die Extraktion der Bedeutung der im Spracherkennermodul generierten Wortkette bzw. des Wortgraphen. In den meisten sprachverstehenden Systemen werden hierfür Verfahren verwendet, die für die Analyse von Texten entwickelt und auf die Erfordernisse der Spracherkennung angepasst wurden. Diese Anpassung ist notwendig, um fehlerhafte Erkennung, Worthypothesen-Alternativen und (im Sinne der „Schriftsprache“) ungültige Äußerungen verarbeiten zu können. Syntaktisches Wissen dient hierbei dazu, die Einheiten im Strom der Worthypothesen zu bestimmen, denen eine Bedeutung (Semantik) zuzuordnen ist. Die Syntax dient also zur Suchraumbeschränkung, denn ohne dieses Wissen müsste man für alle kombinatorisch möglichen Unterketten einer Wortkette überprüfen, ob man dieser Teilkette eine Bedeutung zuordnen kann (im obigen Beispiel z. B. für ..., *Donnerstag*, *Donnerstag am*, ..., *Nachmittag so*, ...).

Typische Analyseansätze sind beispielsweise HPSG (Head driven Phrase Structure Grammar) [16] und TUG (Trace Unification Grammar) [20]. Die Ansätze basieren darauf, dass linguistisches (syntaktisches und semantisches) Wissen über die Wörter in einem Lexikon kodiert ist.

Abbildung 4 zeigt den Lexikon-Eintrag für das Wort „*der*“ sowie eine Regel, die es erlaubt, zwei Worthypothesen, die mit den Platzhaltern `det` (für *determiner*, Artikel) und `n` (für Nomen) bezeichnet sind, zu einer mit `np` bezeichneten Nominalphrase zusammenzubauen. Die Regel kann feuern, wenn der Bedingungsteil erfüllt ist (die Nummern beziehen sich auf die Nummern im Bild):

1) Der Syntaxeintrag im Lexikon für die Worthypothese `det` besagt, dass es sich um einen Artikel handelt.

2) Der Syntaxeintrag im Lexikon für die Worthypothese `n` besagt, dass es sich um ein Nomen handelt.

3a) - 3c) Die Syntaxeinträge im Lexikon für die beiden Hypothesen stimmen in Bezug auf Kasus, Numerus und Genus überein.

Falls die Regel feuert, wird ein neues Konstrukt `np` erzeugt, das folgende Eigenschaften hat:

4) Der Syntaxeintrag besagt, dass es sich um eine Nominalphrase handelt.

5a) - 5c) `np` übernimmt den Eintrag für Kasus, Numerus und Genus von `det`.

6) `np` übernimmt die semantische Bedeutung von `n`.

Lexikoneintrag für „der“:

$$\left[\begin{array}{l} Syntax: \left[\begin{array}{l} syntaktische\_Klasse : Artikel \\ Kasus : Nominativ \\ Numerus : Einzahl \\ Genus : männlich \end{array}\right] \\ Semantik : \left[\ type : dummy\ \right] \end{array}\right]$$

Regel zur Kombination von Artikel und Nomen:

```
np ===> det, n if

/* Bedingungen */

 1)  det:Syntax:syntaktische_Klasse:Artikel
 2)  n:Syntax:syntaktische_Klasse:Nomen
3a)  [det,Syntax,Kasus]     == [n,Syntax,Kasus],
3b)  [det,Syntax,Numerus]   == [n,Syntax,Numerus],
3c)  [det,Syntax,Genus]     == [n,Syntax,Genus],

/* Aktionen beim Feuern der Regel */

 4)  np:Syntax:Nominalphrase,
5a)  [np,Syntax,Kasus]      == [det,Syntax,Kasus],
5b)  [np,Syntax,Numerus]    == [det,Syntax,Numerus],
5c)  [np,Syntax,Genus]      == [det,Syntax,Genus],
 6)  [np,Semantik]          == [n,Semantik].
```

Abbildung 4: Lexikon-Eintrag für das Wort „der“ und Beispielregel einer Phrasenstrukturgrammatik

Mit Hilfe solcher Regeln kann man aus den Worthypothesen immer größere Konstrukte erzeugen, bis dadurch das Eingabesignal komplett überspannt wird. Das Kontrollmodul, welches entscheidet, welche Regel für welche Teilkonstrukte als nächstes angewendet wird, wird Parser genannt. Ziel von vielen Projekten im Bereich der geschriebenen Sprache ist es, Grammatiken und Parser zu entwickeln, welche einen möglichst hohen Anteil von möglichst unbeschränktem Text verarbeiten können. Diese Ansätze setzen allerdings meistens eine fehlerfreie und syntaktisch korrekte Eingabe voraus. Bereits die Annahme der Fehlerfreiheit ist für gesprochene Sprache nicht gegeben. Selbst die weltweit besten Spracherkenner sind in Bezug auf Fehlerfreiheit über sehr viele Anwendungen hinweg eine Größenordnung oder mehr schlechter als der Mensch [13]. Allerdings sind in einem sprachverstehenden System nicht immer vollständige Analysen notwendig; so erfordert

z. B. die Anwendung „Zugauskunft“ für die Äußerung „*ich möchte äh ich meine meine Frau und ich möchten nach Hamburg fahren*“ eigentlich nur die Information, dass es sich bei dem Zielort um Hamburg handeln soll. Wenn die vollständige Analyse scheitert, kann u.U. auf einer Teilanalyse aufgesetzt werden. Ähnlich ist es auch bei einer Übersetzungsaufgabe, wo es in der Regel genügt, die Intention des Sprechers und den wesentlichen semantischen Gehalt des Gesprochenen in die Fremdsprache zu übertragen.

4.2 Prosodische Analyse

Die Prosodie beschäftigt sich mit suprasegmentalen (lautübergreifenden) sprachlichen Ereignissen. Diese Ereignisse überlagern sprachliche Einheiten, die mehr als einen Laut umfassen, also *Silben, Wörter, Phrasen, Sätze*, usw. Zur spektralen Dimension zählen *Klangfarbe, Tonhöhe, Stimmlage* und *Stimmqualität*, zur Intensität *Lautheit*, und die zeitliche Dimension umfasst *Pausensetzung, Dauerverhältnisse, Rhythmus, Sprechgeschwindigkeit* und *Tempo*. Diesen perzeptiven Einheiten entsprechen akustische Parameter. So ist z. B. die *Grundfrequenz* des Sprachsignals das akustische Korrelat der *Tonhöhe*.

Der Zuhörer extrahiert Information aus den wahrgenommenen prosodischen Ereignissen, d. h. wir können den Ereignissen funktionale Rollen zuordnen. Als wichtigste Funktionen werden allgemein die prosodische Markierung von *Satz- und Phrasen-Grenzen, Betonung, Satzmodus* und *Gemütszustand (Emotion)* angesehen. Betrachten wir die folgenden Äußerungen sowie eine Übersetzung ins Englische, so sehen wir die Wichtigkeit prosodischer Information. Diese Äußerungen können in einem Übersetzungsszenario durchaus vorkommen.

„*Vielleicht. Am Montag bei mir. Paßt das?*“
„*Maybe. On Monday, at my place. Is that OK?*“

„*Vielleicht am Montag. Bei mir paßt das.*“
„*Maybe on Monday. That's possible for me.*“

„*Dann müssen wir noch einen Termin ausmachen.*“
„*Then we still have to fix a date.*“

„*Dann müssen wir noch einen Termin ausmachen.*“
„*Then we have to fix another date.*“

Obwohl die Bedeutung der prosodischen Information in der Mensch-Mensch-Kommunikation allgemein anerkannt wird, wird diese Informationsquelle in der automatischen Sprachverarbeitung bisher nur spärlich benutzt. Die wichtigste Rolle der Prosodie, die Segmentierung und Disambiguierung von Äußerungen, kam in den bisherigen Anwendungen oft nicht zum Tragen, da entweder diese Analyseaufgaben nicht notwendig waren (z. B. in Diktiersystemen) oder da die Äußerungen der Benutzer zu kurz waren. So ist zum Beispiel die durchschnittliche Länge einer Äußerung in einem Feldexperiment mit einem Zugauskunftssystem 3,5 Wörter, vgl. [6]. Für die Verarbeitung von Spont-

ansprache, von relativ langen Redebeiträgen sowie bei der Anforderung, diese Beiträge nicht nur zu erkennen, sondern auch inhaltlich zu erschließen, hat sich die Verwendung prosodischer Information jedoch als unabdingbar erwiesen [14, 12].

4.3 Dialogsteuerung

Aufgabe der Dialogsteuerung ist es zum einen, die semantische Repräsentation der Benutzeräußerung in den Kontext des bis dahin geführten Dialogs einzubetten, und zum anderen, die nächste Aktion des Systems zu planen. Im Falle eines starren Menübaumes sind diese Schritte eher einfach, in einem freien Mensch-Maschine-Dialog ist jedoch ein „Gedächnis“ notwendig, um unvollständige (elliptische) Äußerungen - so wie sie in Spontansprache typisch sind - korrekt interpretieren zu können. So kann die Benutzeräußerung

den Josef

nur korrekt interpretiert werden, wenn die letzte Systemäußerung des „eisernen Fräuleins vom Amt“

ich habe zwei Müller in meinem Verzeichnis, Josef Müller und Hans Müller. Wen möchten Sie sprechen?

bekannt ist. Noch mehr „Intelligenz“ von Seiten der Dialogsteuerung ist erforderlich, wenn z. B. der Benutzer eines Zugauskunftssystems die Frage

Sie wollen von Köln nach Bamberg fahren?

mit

Nein, nach Amberg.

beantwortet, um zu erkennen, dass hierdurch — trotz des „Nein“ zu Beginn der Äußerung — der Abfahrtsbahnhof Köln implizit bestätigt wird.

Nachdem die Benutzeräußerung im Dialogkontext interpretiert ist, muss die Dialogsteuerung u.U. überprüfen, ob alle notwendigen Informationen vorhanden sind, um eine Anwendungs-Datenbankanfrage zu starten. Im Falle einer Zugauskunft kann z. B. keine Datenbankanfrage gestartet werden, wenn nur der Zielort bekannt ist. In diesem Fall initiiert das System eine Nachfrage beim Benutzer, ansonsten kann die Dialogsteuerung direkt eine Datenbankanfrage starten, die abgefragte Information aufbereiten und an den Benutzer weitergeben.

4.4 Sprachsynthese

Es gibt eine Reihe von Möglichkeiten, Computer zum Sprechen zu bringen. Es hängt von der jeweiligen Applikation ab, welche Methode vorzuziehen ist.

Die einfachste Möglichkeit besteht darin, dem Benutzer Äußerungen vorzuspielen, die zuvor aufgenommen und digital gespeichert wurden. Eine Variante hiervon ist die Verkettung von einzeln gespeicherten Wörtern oder Satzfragmenten zu einer Gesamtäußerung (*canned speech*). Für unser Fahrplanauskunftssystem wurde beispielsweise das Satzfragment „*Sie möchten nach ...*“, jeder einzelne Ort mit IC-Bahnhof, sowie das Wort „*fahren*“ von einem Mitarbeiter einzeln gesprochen und aufgenommen. Während des Dialoges werden hieraus Systemäußerungen wie „*Sie möchten nach Bamberg fahren?*“ zusammengebaut. Gegenüber einer „echten“ Sprachsynthese zeichnet sich dieses Verfahren in aller Regel durch eine bessere Verständlichkeit aus.

Ist der aktive Wortschatz des Systems jedoch zu groß, ist eine solche Vorgehensweise nicht mehr praktikabel. In diesem Falle greift man auf Sprachsyntheseverfahren zurück, die unter den Bezeichnungen *text-to-speech* (TTS) oder *concept-to-speech* (CTS) verbreitet sind, oder auch auf Kombinationen dieser Verfahren [10]. Im TTS-Verfahren wird zunächst eine linguistische Analyse des zu sprechenden Textes durchgeführt, um z. B. die zu betonenden Wörter und Silben sowie eine geeignete Intonation zu ermitteln. Die Wörter selbst werden an Hand von Aussprachelexika aus einem Inventar von Laut- oder Silbenbausteinen zusammengesetzt. Im Rahmen von Sprachdialogsystemen können CTS-Systeme, denen anstelle einer Folge von Wörtern und Satzzeichen *semantische Konzepte* als Eingabe dienen, günstiger sein. Hier kann eine sinnvolle Intonation ohne den Umweg über die Generierung des zu sprechenden Textes und dessen anschließender linguistischer Analyse festgelegt werden.

5 Anwendungen und Perspektiven

Bereits heute ist absehbar, dass Spracherkennungstechnologie den Umgang von Menschen mit Computern und Maschinen enorm verändern wird. Im Folgenden werden einige der aussichtsreichsten Anwendungsmöglichkeiten skizziert. Ein tabellarischer Überblick über eine Auswahl dieser Anwendungen findet sich in Tabelle 1, in der die Anwendungen an Hand der in Abschnitt 2 eingeführten Leistungskriterien eingeordnet werden. Ein umfangreicher Überblick über Anwendungen und Produkte in diesem Bereich findet sich in [21].

5.1 Diktiersysteme

Diktiersysteme sind mittlerweile auf dem besten Wege, Bestandteil der Grundausstattung von Personalcomputern zu werden. Zu beobachten ist ein dramatischer Preisverfall bei gleichzeitiger Steigerung der Leistungsfähigkeit der Systeme. Es befinden sich Systeme mehrerer Anbieter auf dem Markt, die zur Zeit Vokabulargrößen von etwa 60.000 Wörtern bewältigen. Die Sprachmodelle und Vokabulare lassen sich durch den Erwerb spezieller Branchenlösungen an die Erfordernisse bestimmter Berufsgruppen (z. B. Ärzte, Juristen) anpassen.

	Diktier-systeme	Geräte-bedienung im Auto	Auskunfts-Dialog (z.B. Fahrplan)	Behinderten-unter-stützung
Sprecherabh.	ja	nein	nein	ja
Sprechart	kontin.	diskret/kontin.	spontan	diskret
Wortschatz	ca. 60.000	ca. 100	ca. 2000	ca. 100
Gramm. Kompl.	groß	gering	mittel	gering
Eingabemedium	Headset	Mikrophon, z.B. am Innenspiegel	Telefon	Headset

Tabelle 1: Übersicht über Leistungsmerkmale heutiger Spracherkennungssysteme in unterschiedlichen Anwendungen. Einzelheiten zu den Leistungsmerkmalen finden sich in Abschnitt 2. Die aufgeführten Zahlenwerte sind lediglich als grobe Richtwerte zu verstehen.

Die Erkennungsraten dieser Systeme liegen nach Herstellerangaben jeweils bei etwa 92 bis 95 Prozent, wobei dies voraussetzt, dass der Benutzer das System vorher an seine Sprechweise adaptiert hat. Gleichzeitig ist es aber auch erforderlich, dass der Benutzer deutlich und in gleichmäßigem Tempo spricht, und z. B. die Endungen von Wörtern nicht verschluckt. Auch ungewöhnliche Formulierungen bereiten den Systemen Schwierigkeiten, sodass beispielsweise das Diktieren literarischer Texte von geringem Erfolg gekrönt sein wird.

Hinzu kommt, dass es Sprecher gibt, denen es beim besten Willen nicht gelingt, mit solchen Systemen befriedigende Erkennungsergebnisse zu erzielen; andere wiederum werden überdurchnittlich gut „verstanden“. Dies hängt damit zusammen, dass die Adaptionsfähigkeit der heutigen Systeme an die spezifische Sprechweise einzelner Sprecher noch ungenügend ist, wodurch Sprecher benachteiligt sind, die — etwa in den Eigenschaften ihres Vokaltrakts oder in ihrer dialektalen Färbung — zu stark vom „Durchschnittssprecher“ abweichen. Dieses Thema ist Gegenstand intensiver Forschungsarbeit, die weitere Verbesserungen der einschlägigen Produkte erwarten lässt.

Der Einsatz solcher Systeme im Büroalltag setzt in jedem Falle die individuelle Bereitschaft voraus, sich auf diese stellenweise noch etwas gewöhnungsbedürftige Technologie einzustellen. Hinzu kommen natürlich Anforderungen an den Arbeitsplatz, da die Kollegen nicht durch das Diktieren gestört werden wollen, und da darüberhinaus möglichst wenig Hintergrundgeräusche vorhanden sein sollten.

5.2 Sprachsteuerung

Sprachsteuerungen für PC-Bedienoberflächen sind bereits heute erhältlich. Diese orientieren sich bislang streng an der auf Maus und Tastatur zugeschnittenen Anwenderschnittstelle und sind daher im Normalfall keine echte Erleichterung im Umgang mit dem PC. Wirklich interessant sind diese momentan nur für Anwender, die aus dem einen

oder anderen Grund keine Hand frei haben, mit der sie den PC bedienen könnten.

Die Zukunft für Sprachsteuerungen wird vermutlich eher in der Bedienung von Geräten liegen, die nicht über die bewährten, aber in vielen Fällen ungeeigneten Schnittstellen Bildschirm, Maus und Tastatur verfügen, besonders im Bereich der sogenannten *Embedded Systems*. Im Zuge der fortschreitenden Miniaturisierung wird die für Spracherkennung notwendige Rechenleistung in wenigen Jahren auf winzigen Chips Platz finden. Gleichzeitig nimmt die Funktionalität elektronischer Geräte wie Mikrowellengeräte, Videorecorder und Mobiltelefone immer mehr zu. (Bereits heute ist Umfragen zufolge die überwiegende Zahl der Besitzer eines Videorecorders nicht in der Lage, diesen zu programmieren.) Es ist daher absehbar, dass Sprache als Mensch-Maschine Schnittstelle hier eine entscheidende Rolle spielen wird, und Kommandos wie „*Nimm mir bitte morgen den Film gleich nach der Tagesschau auf*“ schon bald technisch möglich werden. Hierbei ist es von zentraler Bedeutung, dass an die Stelle einer reinen Kommandosteuerung ein natürlichsprachlicher Dialog tritt, der dazu dient, Unklarheiten auszuräumen und Fehler zu korrigieren.

Ein besonders interessantes Anwendungsgebiet für Sprachsteuerungen liegt im Automobilbereich. Auf der einen Seite sollte der Fahrer weder den Blick von der Fahrbahn nehmen noch die Hände vom Lenkrad, auf der anderen Seite werden ihm immer komplexere Geräte wie Telefon, CD-Player, Navigationssystem, und bald auch Internetanschluss und Auto-PC zur Verfügung gestellt. Für die Bedienung der genannten Geräte gibt es zur Entwicklung leistungsfähiger Sprachkommunikationsschnittstellen praktisch keine Alternative, da die Spracheingabe Hände und Augen nicht in Anspruch nimmt.

5.3 Telefonieapplikationen

Einer der interessantesten Märkte für Spracherkennungstechnologie liegt im Bereich der sogenannten *Telefonie*, der alle Anwendungen umfasst, die die Übertragung von gesprochener Sprache über Telefon einschließen. Gemeint sind hierbei insbesondere sogenannte *Call Center*. Hierzu gehören beispielsweise Service- und Bestell-Hotlines, Informations- und Auskunftsdienste sowie Telefonvermittlungen. Das in der Einleitung beispielhaft wiedergegebene Gespräch eines Anrufers mit dem Prototypen eines Fahrplanauskunftssystems veranschaulicht das Automatisierungpotential, das Sprachtechnologie in diesem Bereich bietet.

Im Rahmen der fortschreitenden *Computer Telephony Integration (CTI)*, also der Einführung von Computertechnologie in die Telefonkommunikation, hat auch die automatische Spracherkennung bereits Eingang in zahlreiche kommerzielle Systeme gefunden. In der Regel handelt es sich um relativ einfache Systeme, die die Navigation in einem menüartig strukturierten Dialogablauf gestatten, und die auf Einzelworterkennern oder sogenannten *Word Spottern* basieren. Word Spotter sind Spracherkenner, die in einer Äußerung nach einem bestimmten Schlüsselwort suchen, und hierfür neben den Schlüsselwörtern sogenannte „Füllwörter“ bereitstellen. Anders als bei „echten“ Spracherkennern spielt die Reihenfolge der „Füllwörter“ keine Rolle, und bei der Aus-

wertung des Erkennungsergebnisses werden sie ignoriert. Dem Anrufer werden i.d.R. mehrere Schlüsselwörter vorgegeben (z. B. „*Informationen*“, „*Kontoabfrage*“), und der Word Spotter akzeptiert dann auch den Satz „*Ich möchte Informationen bitte.*“. Solche einfachen Systeme haben gegenüber komplexen, natürlichsprachlichen Dialogsystemen (siehe Abschnitt 4) den Vorteil, dass sie einfach und ohne spezielle Fachkenntnisse für bestimmte Anwendungen konfiguriert werden können und damit preisgünstig einzusetzen sind. Mittelfristig werden sich wesentlich leistungsfähigere Systeme etablieren, die einen quasi-natürlichen Dialog mit dem Anrufer ermöglichen.

5.4 Behindertenunterstützung

Spracherkennungstechnologie bietet für die Unterstützung von Körperbehinderten ein enormes Potential, sowohl im privaten Bereich als auch am Arbeitsplatz. Das sprachgesteuerte Öffnen und Schließen von Fenstern und Türen, das Bedienen von Radio und Telefon, und schließlich auch der Umgang mit PCs sind nur einige Beispiele für Tätigkeiten, die bereits heute mit Hilfe von kommerziell erhältlicher Sprachtechnologie selbst solchen Menschen ermöglicht werden können, die vom Hals abwärts gelähmt sind.

Ein tragbares, sprachgesteuertes Navigationssystem für Blinde und Sehbehinderte befindet sich bereits in der Entwicklung. Noch Zukunftsmusik sind dagegen sprachbegabte Roboter, die als flexible und nützliche Helfer in Haushalt und Beruf fungieren („*Bring mir doch bitte mal eine Tasse Kaffee!*“), oder Fernsehgeräte, die automatisch Untertitel für hörbehinderte Menschen generieren.

5.5 Weitere Anwendungen

Es gibt zahlreiche weitere Anwendungsfelder für Spracherkennungstechnologie: Die Zugangskontrolle, die Unterstützung beim Erlernen von Fremdsprachen und die Sprachtherapie sind hier nur einige Beispiele. Mit zunehmender Leistungsfähigkeit der Technologie ergeben sich weitere kommerziell interessante Applikationen, wie etwa die Verschriftung von Radio- und Fernseh-Archiven, die Kriminalistik oder die automatische Übersetzung.

5.6 Aktuelle Forschungsprojekte

Die natürlichsprachliche Kommunikation zwischen Mensch und Maschine ist Gegenstand zahlreicher aktueller Forschungsvorhaben. Die DFG finanziert grundlagenorientierte Projekte, wie z.B. [1, 8]. Eine Übersicht über eher anwendungsnahe Projekte der EU in diesem Bereich findet sich unter [7]; Schwerpunkte liegen hier u.a. in der Automatisierung von Call Centern (ACCESS, IDAS), im Einsatz von Spracherkennung im Auto (VODIS), sowie in der automatischen Zugauskunft (ARISE).

Die automatische Übersetzung spontan gesprochener Sprache steht im Mittelpunkt BMBF–Projekts VERBMOBIL [5, 2]. Zunehmend in den Blickpunkt, z.B. im Rahmen der BMBF-Leitprojekte zur Mensch–Technik–Interaktion [3], wie SMARTKOM [4] und EMBASSI, des japanischen MITI–Projekts RWC [18] oder des amerikanischen NSF–Projekts

STIMULATE [15], rückt die Verschmelzung von Spracherkennungstechnologie mit weiteren Informationsquellen, wie etwa der optischen Wahrnehmung von Mimik und Gestik. Auf dieser Grundlage soll letztlich ein natürlicher und intuitiver Umgang des Menschen mit Computern ermöglicht werden.

6 Zusammenfassung

Automatische Spracherkennung ist heute von der Robustheit und Flexibilität der menschlichen Sprachwahrnehmung noch weit entfernt. Probleme liegen unter anderem noch in dem starken Einfluss des akustischen Kanals und der Sprechereigenheiten, in der Empfindlichkeit gegenüber Hintergrundgeräuschen, in der Behandlung ungrammatischer, spontan gesprochener Sprache, im Umgang mit Akzenten und Dialekten, und in der mangelnden Robustheit gegenüber Themenwechseln oder „unvorhergesehenen" Äußerungen. Unter gewissen Einschränkungen sind jedoch bereits heute beeindruckende Erkennungsergebnisse möglich, die zu einer Reihe kommerzieller Geräte führten.

Die Funktionsweise nahezu aller leistungsfähigen Spracherkennungssysteme basiert auf Hidden-Markov-Modellen. Dieser statistische Ansatz erlaubt es, Wortmodelle mit großen Mengen gesprochener Sprache automatisch zu trainieren. Als zusätzliche Wissensquelle werden Sprachmodelle verwendet, die die Wahrscheinlichkeit für die Aufeinanderfolge bestimmter Wörter festlegen. Statistische Sprachmodelle lassen sich mit in geschriebener Form vorliegenden Äußerungen, aber z. B. auch mit Texten in Briefen oder Zeitungen, ebenfalls automatisch trainieren. Die besten Erkennungsergebnisse werden erzielt, wenn die zum Trainieren der Wort- und Sprachmodelle herangezogenen Daten den in der Erkennungsphase vorliegenden Äußerungen möglichst ähnlich sind.

Mit steigender Komplexität der Anwendungen gewinnt das „Verstehen" des Gesprochenen gegenüber dem reinen „Erkennen" zunehmend an Bedeutung. Wird eine quasinatürliche Kommunikation zwischen Mensch und Maschine angestrebt, so sind neben der eigentlichen Spracherkennung noch weitere Probleme zu lösen. Hierzu gehören die syntaktische und prosodische Analyse von Äußerungen, die Dialogsteuerung und die Sprachsynthese. Besonders die intelligente Dialogsteuerung ist von zentraler Bedeutung für die Akzeptanz sprachverstehender Systeme.

Das Spektrum der Anwendungsmöglichkeiten geht weit über reine Diktiersysteme hinaus. Von besonderer Bedeutung sind die Automatisierung von Telefondialogen, die Steuerung von Geräten und die Unterstützung von Behinderten; darüberhinaus gibt es jedoch vielfältige weitere Einsatzmöglichkeiten.

Literatur

[1] Universität Bielefeld. SFB 360: Situierte Künstliche Kommunikatoren. http://www.techfak.uni-bielefeld.de/sfb/.

[2] T. Bub and J. Schwinn. Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. In *Int. Conf. on Spoken Language Processing*, volume 4, pages 1026–1029, Philadelphia, October 1996.

[3] E. Bulmahn. Presseerklärung zu den Sieger–Leitprojekten des Wettbewerbs Mensch–Technik–Interaktion in der Wissensgesellschaft am 02.03.1999 in Bonn. http://www.bmbf.de/deutsch/veroeff/presse/pm99/pm030299.htm.

[4] DFKI. SmartKom — Dialogische Mensch–Technik–Interaktion durch koordinierte Analyse und Generierung multipler Modalitäten. http://www.dfki.de/SmartKom/.

[5] DFKI. Verbmobil. http://www.dfki.de/Verbmobil/.

[6] W. Eckert, E. Nöth, H. Niemann, and E. Schukat-Talamazzini. Real Users Behave Weird — Experiences made collecting large Human–Machine–Dialog Corpora. In *Proceedings of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, pages 193–196, Vigsø, Denmark, 1995.

[7] EU. HLT Project's Page. http://www.linglink.lu/hlt/projects/.

[8] F. Gallwitz. *Integrated Stochastic Models for Spontaneous Speech Recognition.* Dissertation. Technische Fakultät der Universität Erlangen–Nürnberg, erscheint 1999.

[9] F. Gallwitz, M. Aretoulaki, M. Boros, J. Haas, S. Harbeck, R. Huber, H. Niemann, and E. Nöth. The Erlangen Spoken Dialogue System EVAR: A State–of–the–Art Information Retrieval System. In *Proceedings of 1998 International Symposium on Spoken Dialogue (ISSD 98)*, pages 19–26, Sydney, Australia, 1998.

[10] D. Gibbon, R. Moore, and R. Winski, editors. *Handbook of Standards and Resources for Spoken Language Systems.* Mouton de Gruyter, Berlin; New York, 1997.

[11] Frederick Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, Massachusetts, 1998.

[12] R. Kompe. *Prosody in Speech Understanding Systems.* Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.

[13] R.P. Lippmann. Speech Recognition by Machines and Humans. *Speech Communication*, 22(1):1–15, 1997.

[14] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann. Prosodische Information: Begriffsbestimmung und Nutzen für das Sprachverstehen. In *Mustererkennung 1997*, Informatik FB, pages 37–52, Heidelberg, 1997. Springer–Verlag.

[15] NSF. Presseerklärung zum Projekt STIMULATE (Speech, Text, Image and Multimedia Advanced Technology Effort) am 18.2.1997 in Arlington, Virginia. http://www.nsf.gov/od/lpa/news/press/pr9714.htm.

[16] C. Pollard and I. Sag. *Information–based Syntax and Semantics, Vol. 1*, volume 13 of *CSLI Lecture Notes*. CSLI, Stanford, CA, 1987.

[17] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey, 1993.

[18] RWC. Real World Computing — Real World Intelligence Technology. http://www.rwcp.or.jp/outline/div/development-e.html#kenkyu1.

[19] E. G. Schukat-Talamazzini. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen*. Vieweg, Braunschweig, 1995.

[20] N. Sikkel. *Parsing Schemata*. CIP-Gegevens Koninklijke Bibliotheek, 1993.

[21] Axel Susen. *Spracherkennung — Kosten, Nutzen, Einsatzmöglichkeiten*. VDE Verlag, Berlin; Offenbach, 1999.