



Spracherkennung und Sprachdialog: Stand der Technik, Einsatzbeispiele und zukünftige Trends

Dr. Jürgen Haas,, Dr. Florian Gallwitz, Jens Kornwachs, Dr. Martin Schröder
Sympalog Voice Solutions GmbH
Karl-Zucker-Straße 10
91052 Erlangen
09131/616 61-0
haas@sympalog.de

Abstract: Das automatische Erkennen und Verstehen von gesprochener Sprache ist schon seit Jahrzehnten Gegenstand intensiver Forschungsarbeit. Mittlerweile sind nun verschiedene Anwendungen auf Grundlage der hieraus hervorgegangenen Technologie kommerziell verfügbar, und es zeichnet sich ab, dass spracherkennende und -verstehende Systeme schon bald aus dem betrieblichen Alltag nicht mehr wegzudenken sind. In diesem Beitrag werden neben den technologischen Grundlagen von Sprachsystemen die Möglichkeiten und auch die Grenzen dieser Technologie dargestellt. Heutige Anwendungen werden anhand konkreter Beispiele vorgestellt und es wird ein Ausblick auf kommende Entwicklungen gegeben

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. 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.

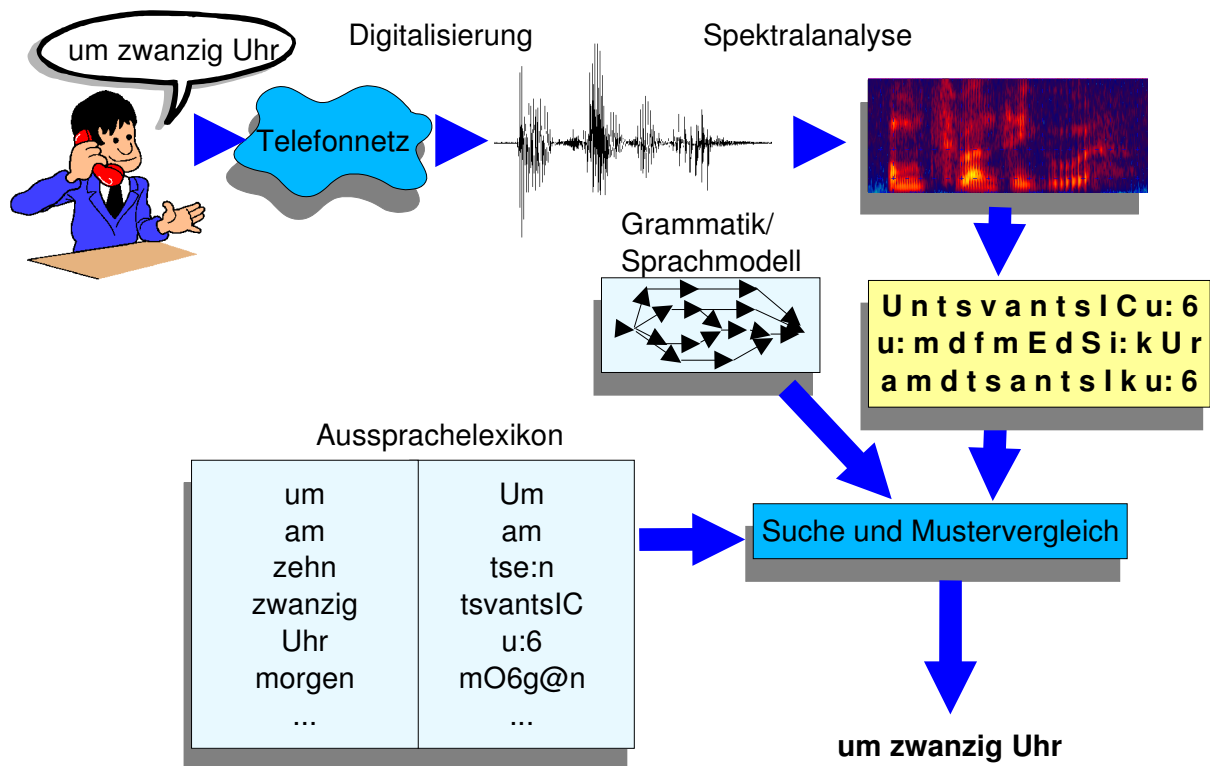
Dennoch hat es gerade im Bereich des automatischen Verstehens von gesprochener Sprache in den vergangenen Jahren erhebliche Fortschritte gegeben, die in Form von Produkten auch dem Endverbraucher zugänglich gemacht werden. Hierzu gehören beispielsweise Mobiltelefone, die bei der Nennung eines Namens 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 produktive Systeme in den unterschiedlichsten Bereichen, die das Verhalten einer menschlichen Auskunftsperson erfolgreich simulieren.

2. Wie funktioniert automatische Spracherkennung

Nachfolgend 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 Spracherkennungler zunächst an Hand von Beispielen trainiert wird, d.h. die Aussprache bestimmter Laute oder Wörter wird automatisch erlernt.

- ◆ **Vorverarbeitung und Merkmalsberechnung:** Am Beginn der Verarbeitungskette werden die in Form von Luftdruckschwankungen vorliegenden Schallwellen mittels eines Mikrophons in ein elektrisches Signal umgewandelt. Aus dem digital vorliegenden Signal gilt es nun, Informationen über die jeweils gesprochenen Laute (bzw. *Phoneme*, die kleinsten bedeutungsunterscheidenden Lauteinheiten) zu gewinnen und daraus einen Merkmalvektor zu bestimmen.
- ◆ **Wortmodellierung und -klassifikation:** Die Aufgabe besteht 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.
- ◆ **Sprachmodellierung:** Unter dem Begriff *Sprachmodellierung* (engl. *language modeling*) fasst man Verfahren zusammen, die dem Spracherkennungler Wissen über die Wahrscheinlichkeit von bestimmten Wortfolgen vermitteln, ohne die eine befriedigende Erkennungsleistung in den meisten Fällen nicht möglich wäre. 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, zum Anderen können Fehler ausgeglichen werden, die durch ungenaue Aussprache oder ungenaue Modellierung der Wörter entstehen würden.
- ◆ **Suche:** Bei der Bestimmung der am wahrscheinlichsten gesprochenen Wortfolge zu einer gegebenen Folge von Merkmalvektoren handelt es sich um ein komplexes Suchproblem, dessen Aufwand mit der Länge des Satzes exponentiell zunimmt. Effiziente Suchverfahren wie die *Viterbi-Suche*, die daraus abgeleitete *Strahlsuche* (beam search) sowie der A*-Algorithmus werden in unterschiedlichen Kombinationen eingesetzt, um dieses Problem in Echtzeit zu lösen. Bei großen Wortschätzen kann zudem durch Anordnung des HMM-Lexikons in Form eines Baumes, an dessen Unterbäumen Wörter mit gleichen Wortanfängen angeordnet sind, der Rechenaufwand in Grenzen gehalten werden.

In der nachfolgenden Abbildung sind sämtlich Schritte, die beim Spracherkennungsprozess durchlaufen werden, nochmals am Beispiel eines Telefonsystems dargestellt. Ein anderer Eingabekanal, z.B. ein Mikrophon hat dabei lediglich Auswirkungen auf die Übertragung des Signals bis zum Arbeitsschritt "Digitalisierung", jedoch nicht für den Spracherkennungsprozess an sich.



3. Leistungskriterien für die automatische Spracherkennung

Entwickler und Forscher im Bereich der automatischen Spracherkennung sehen sich häufig mit Aussagen konfrontiert, wie: „Spracherkennung, wieso? Das gibt' sdoch 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 jener geübter Computerbenutzer vergleichbar ist. 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 100.000 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. Es lassen sich fünf Leistungsachsen definieren:

1. 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 das Vorlesen einiger Übungssätze zugemutet werden kann, ist dies zum Beispiel bei einem Fahrplanauskunftssystem oder gar bei einem sprachgesteuerten Getränkeautomaten nicht praktikabel.

2. Sprechart

Die Unterscheidung zwischen *diskreter* und *kontinuierlicher Sprache* verliert, zumindest im Zusammenhang mit Diktiersystemen, zunehmend an Bedeutung: Die kurzen Sprechpausen zwischen den Wörtern, die zu Beginn 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 verarbeitet werden, die jedoch die automatische Verarbeitung drastisch erschweren. So muss zum Beispiel in einem Fahrplanauskunftssystem mit der folgenden Anfrage gerechnet werden: „*äh ja hallo also ähm nach Hamburg wollt' ich fahr' rab München Pasing so gegen acht gegen zwanzig Uhr heut' a' rd.*“ Zusätzliche Schwierigkeiten ergeben sich noch durch Sprecher mit regionalem Dialekt oder ausländischem Akzent.

3. 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 100.000 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 in einem Diktiersystem nur auf Wunsch in den Erkennungswortschatz aufgenommen.

4. 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 Eigename. 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 Vorgängervörter bereits bekannt sind.

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

5. Eingabemedium

Von großer Bedeutung für automatische Spracherkennungssysteme ist der sogenannte „Eingabekanal“, hierzu gehören das Mikrofon oder auch ein Mikrofonarray und, z.B. im Falle einer Telefonanwendung, auch die Art der Übertragung des Signals (Festnetz- vs. Mobiltelefon). 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 Mikrofon in der Nähe des Mundwinkels positioniert wird. Dennoch wird der Einfluss der Qualität des Mikrofons oft überschätzt; viel wichtiger ist es, dass der Spracherkennungssystem mit Daten *trainiert* bzw. *adaptiert* wurde, die nach Möglichkeit mit dem gleichen Mikrofon unter möglichst ähnlichen akustischen Bedingungen aufgenommen wurden.

Besonders schwierig wird es, wenn das Mikrofon 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. Fahrgeräusche im Auto oder Geräusche in einer Bahnhofshalle), oder gar mehrere Sprecher, die gleichzeitig reden ("Cocktailparty-Effekt") erschweren die Spracherkennung zusätzlich oder machen sie in extremen Fällen 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.

4. Dialogsteuerung

Im Gegensatz zu einem Diktiersystem, bei welchem die vom Spracherkennungssystem gelieferte Wortkette schon das Ergebnis darstellt, wird bei einem *sprachverstehenden* System eine geeignete Systemreaktion erwartet. Die Systemreaktion wird durch eine Dialogsteuerung erreicht, die eine Interpretation des Gesagten vornimmt, eine entsprechende Aktion auslöst und dies dem Benutzer, auf welchem Weg auch immer, mitteilt.

Im Falle eines Kommandoerkennters oder eines einfachen Menüsystems ist die Umsetzung des erkannten Schlüsselwortes in die entsprechende Systemreaktion relativ trivial. Der Anrufer bewegt sich durch eine vorgegebene Menüstruktur, je nach Äußerung verzweigt das System in den vorgesehenen Pfad, z.B. „Wollen Sie zum Bereich Verkauf, Buchhaltung oder Technik“ und stellt dementsprechend weitere Fragen oder löst die passende Aktion, z.B. die Vermittlung zu dem Ansprechpartner, aus.

Erheblich komplizierter wird es, wenn ein intelligentes Dialogverhalten erwartet wird, mit dem das Verhalten eines menschlichen Gesprächspartners imitiert werden soll. Bereits die Interpretation einer Datums- und/oder Uhrzeitangabe (z.B. „*diesen Donnerstag am späten Nach-*

mittag so ab fünf Uhr“) erfordert eine relativ komplexe *syntaktisch-semantische* Analyse des Spracherkennungsergebnisses. Sollen darüber hinaus z.B. die beiden verschiedenen 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 erwartet der Benutzer in aller Regel auch, dass das System sich in natürlicher, gesprochener Sprache ausdrücken und über den aktuellen Zustand informieren kann.

- ◆ **Syntaktisch-Semantische Analyse:** Aufgabe dieses Verarbeitungsschrittes ist die Extraktion der Bedeutung der im Spracherkennungsmodule generierten Wortkette. Syntaktisches Wissen dient hierbei dazu, die Einheiten im Strom der Worthypothesen zu bestimmen, denen eine Bedeutung (Semantik) zuzuordnen ist. Viele Ansätze setzen allerdings meistens eine fehlerfreie und syntaktisch korrekte Eingabe voraus. Bereits die Annahme der Fehlerfreiheit ist jedoch für gesprochene Sprache nicht gegeben. Selbst die weltweit besten Spracherkennungssysteme sind in Bezug auf Fehlerfreiheit über sehr viele Anwendungen hinweg eine Größenordnung oder mehr schlechter als der Mensch. Allerdings sind in einem sprachverstehenden System nicht immer vollständige Analysen notwendig; so erfordert z.B. eine Anwendung „Fahrplanauskunft“ 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.
- ◆ **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. Als wichtigste Funktionen werden allgemein die prosodische Markierung von *Satz- und Phrasen-Grenzen, Betonung, Satzmodus* und *Gemütszustand (Emotion)* angesehen. Betrachten wir die folgende Äußerung, so erkennen wir die Wichtigkeit prosodischer Information: „*Vielleicht. Am Montag bei mir. Passt das?*“ versus „*Vielleicht am Montag bei mir passt das?*“ Obwohl die Bedeutung der prosodischen Information in der Mensch-Mensch-Kommunikation allgemein anerkannt wird, wird diese Informationsquelle in der automatischen Sprachverarbeitung bisher jedoch nur spärlich benutzt.
- ◆ **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. 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.

5. Sprachsynthese

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

Prerecorded oder Canned-Speech

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*

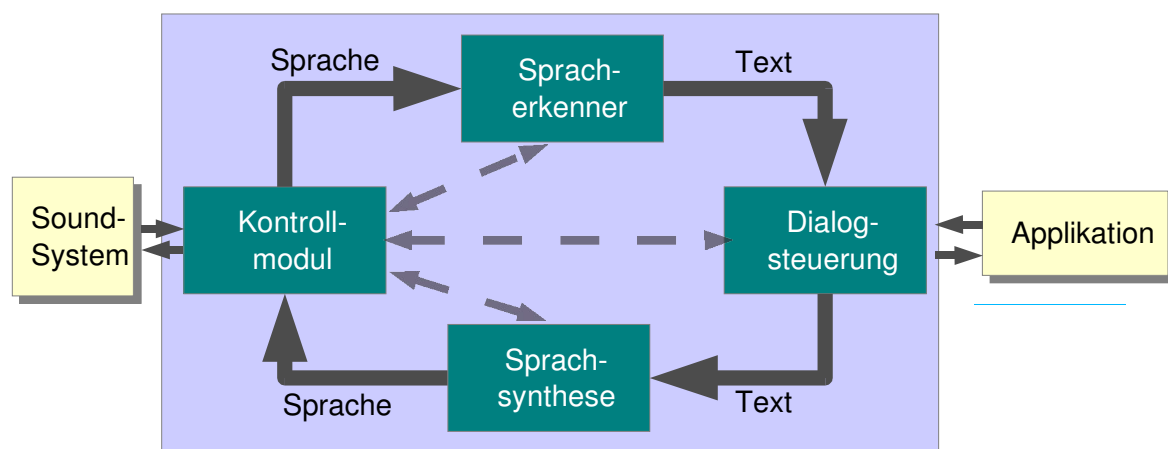
speech). Für Sympalogs Bundesligaauskunftssystem BErTI wurden Satzfragmente wie z.B. "...hat gegen ...", jeder einzelne Bundesligaverein, sowie einzelne Wörter wie z.B. „gewonnen“ von einem Radiosprecher einzeln gesprochen und aufgenommen. Während des Dialoges werden hieraus Systemäußerungen wie „Der VfB Stuttgart hat gegen den 1. FC Kaiserslautern auswärts drei zu zwei gewonnen“ zusammengebaut. Hierzu ist es allerdings notwendig, dass alle möglichen Äußerungen des Systems vorab bekannt sind, so dass eine passende Segmentierung in Satzchnipsel erfolgen kann. Gegenüber einer „echten“ Sprachsynthese zeichnet sich dieses Verfahren in aller Regel dadurch aus, dass die Verständlichkeit deutlich besser ist.

Vollsynthese bzw. text-to-speech (TTS)

Ist der aktive Wortschatz des Systems zu groß oder zu dynamisch, 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. 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.

6. Systemarchitektur Sprachdialogsystem

Wie in den vorherigen Abschnitten erläutert, benötigt man für den Aufbau eines Sprachdialogsystems die drei Komponenten Spracherkennung, Dialogsteuerung und Sprachsynthese. Für ein produktives System werden diese üblicherweise durch ein Kontrollmodul komplettiert, das für die Anbindung an die Signalquelle (z.B. Soundkarte oder Telefonanlage) zuständig ist und bei Bedarf auch die Verteilung der Rechnerlast in einem Rechnernetzwerk leisten kann. In der folgenden Abbildung ist die übliche Systemarchitektur dargestellt.



Die durchgezogenen Pfeile deuten dabei die konzeptionelle Vorgehensweise beim Sprachdialog an, die gestrichelten Pfeile die technische Kommunikation der Module untereinander. Konzeptionell wird wie folgt vorgegangen: das eingehende Sprachsignal wird zum Spracherkennung geschickt, dort wird die wahrscheinlichste Wortkette ermittelt. Diese

wird der Dialogsteuerung übergeben, die die syntaktisch-semantische Analyse durchführt, den aktuellen Zustand im Sinne der zu bedienenden Applikation interpretiert und daraus den nächsten Zustand generiert. Dabei kann die Dialogsteuerung mit der Applikation kommunizieren, entweder um zusätzliche Informationen für den weiteren Dialog abzuleiten oder entsprechende Aktionen anzustoßen. Die Dialogsteuerung generiert zum neuen Dialogzustand die entsprechende Wortkette. Diese wird der Sprachsynthese zur Verfügung gestellt, die diese Äußerung in ein Sprachsignal verwandelt, das dem Benutzer vorgespielt wird. Dann wird auf die nächste Benutzeräußerung gehört und der Zyklus beginnt von Neuem.

Technisch läuft die Kommunikation der Module üblicherweise über Programmierschnittstellen, um die notwendigen Funktionalitäten in den einzelnen Modulen zum richtigen Zeitpunkt anzustoßen und diesen die jeweils notwendige Information zur Verfügung zu stellen. So findet das Spracherkennungsergebnis i.d.R. nicht direkt den Eingang zur Dialogsteuerung, sondern das Erkennungsergebnis wird dem Kontrollmodul gesendet, von wo aus es an die Dialogsteuerung weiter gegeben wird. Die Kommunikation basiert häufig auf standardisierten Protokollen, wie z.B. TCP/IP oder HTTP, und ist, gerade wenn die Produkte für den Telefonbereich konzipiert sind, netzwerkfähig, so dass unterschiedliche Module auf mehrere Rechner verteilt werden können bzw. sogar mehrere Instanzen eines Moduls in einem Netzwerk zur Verfügung stehen.

7. Anwendungen – Praxisbeispiele

Heutzutage sind bereits zahlreiche Sprachapplikationen in den verschiedensten Einsatzbereichen erfolgreich im produktiven Betrieb. Nachstehend werden einige Anwendungen aus verschiedenen Bereichen skizziert:

Diktiersysteme

Diktiersysteme sind mittlerweile erfolgreich im Einsatz, besonders im medizinischen und juristischen Bereich.

Beispiele:

- In Baden-Württemberg wurden kürzlich 1.000 Richterarbeitsplätze mit Spracherkennungssoftware zum Diktieren von Urteilen ausgestattet.
- In medizinischen Einrichtungen, in denen ein Großteil der Arbeit aus dem Diktieren von Befunden besteht, z.B. in der Radiologie, sind Diktiersysteme im täglichen Einsatz.

Sprachsteuerung

Eine Sprachsteuerung (speech control) von Geräten und Maschinen über Mikrofon bringt in vielen Fällen Vorteile. Besonders wenn der Benutzer die Hände nicht frei hat, die Gefahr von Verschmutzung oder Kontamination gegeben ist, die Konsole und Arbeitsplatz voneinander entfernt sind oder die Bedienungshierarchien sehr komplex sind. Ein Haupttreiber der Sprachtechnologieentwicklung in diesem Bereich ist im Automobilumfeld zu finden.

Beispiele:

- Steuerung von Monitoren oder Einrichtungen im chirurgischen Bereich über Kommandos, zum Teil angereichert durch Dialogfunktionalität um Mehrdeutigkeiten aufzulösen
- Sprachsteuerung im Automobil zur Steuerung und Eingabe eines Telefons

Spracherkennung zur Datenerfassung

In vielen Anwendungsfällen kann die Spracherkennung sinnvollerweise als Kanal für die Eingabe von zu protokollierenden Daten eingesetzt werden, ähnlich wie bei der Sprachsteuerung von Geräten in solchen Gebieten, in denen der Benutzer üblicherweise die Hände nicht frei hat oder das Mitführen anderer Protokollierungshilfsmittel umständlich ist. Solche Situationen ergeben sich u.a. im medizinischen Umfeld, z.B. bei der Protokollierung von Operationen oder aus dem industriellen Umfeld in der Qualitätssicherung.

Beispiele:

- Kommissionierung per Sprache („pick-by-voice“)
- Spracheingabe bei Kfz-Hauptuntersuchungen zur Erstellung des Prüfberichts

Sprachdialogsysteme

Einer der interessantesten Märkte für Spracherkennungstechnologie liegt im Bereich der Telefonie (Anwendungen, die die Übertragung von gesprochener Sprache über Telefon einschließen). Vor allem im Callcenter-Bereich ergibt sich zur Vorqualifizierung und automatischen Bearbeitung von Gesprächen ein breites Spektrum unterschiedlicher Anwendungen wie Service- und Bestellhotlines, automatische Vermittlungen und Auskunftsdienste.

Beispiele:

- Intelligentes Vermittlungsportal bei der Sixt AG: Vermittelt die Anrufer an den richtigen Ansprechpartner oder Bereich, abhängig vom Anliegen, gewünschten Gesprächspartner oder der Abteilung. („Ich möchte bitte ein Auto mieten“)
- Bürgerinformationssystem bei der Stadt Würzburg: Anrufer bekommen Auskünfte zum für ihr Anliegen zuständigen Amt („Wo muss ich mich denn hinwenden, wenn ich mein Auto ummelden will?“)

7. Aktuelle Forschungsprojekte

Die multimodale Mensch-Maschine-Interaktion ist ein Forschungsthema von großer Relevanz und ein zukünftiger Forschungstreiber. Aus diesem Grund bestehen in diesem Umfeld wichtige Forschungsprojekte. Exemplarisch wird nachfolgend das aktuelle SmartWeb-Projekt als eines der Bedeutendsten vorgestellt. Als Ergebnis des Projekts soll die Recherche im Internet zukünftig einfacher und effektiver möglich sein - und das auch über UMTS-Telefone und andere mobile Endgeräte. An dem Projekt unter Leitung des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI) arbeiten insgesamt 14 Partner aus Wirtschaft und Wissenschaft zusammen, darunter DaimlerChrysler, die Deutsche Telekom und Siemens. SmartWeb wird vom Bundesministerium für Bildung und Forschung (BMBF) mit insgesamt 13,7 Millionen Euro gefördert. Sympalog ist in SmartWeb für die automatische Spracherkennung im mobilen Anwendungsszenario verantwortlich.

Das World Wide Web (WWW) hat den weltweiten Zugang zu digital gespeicherter Information drastisch vereinfacht und beschleunigt. Allerdings gibt es bisher zwei Zugangshürden.

- Der Zugang zu den Inhalten ist größtenteils auf PCs mit großen Bildschirmen optimiert. Statt eines einfachen, intuitiven Zugangs mittels natürlicher Sprache über das Mobiltelefon suchen derzeit Suchmaschinen textuell nach Inhalten, die nicht in jeder Modalität, z.B. nur mittels Sprache, dem Benutzer zugänglich gemacht werden können.

- Bislang waren die Inhalte im WWW nur maschinenlesbar, ohne maschinell verstehbar zu sein. Da Information im WWW meist in natürlicher Sprache präsentiert wird, sind die bei einer Suche gefundenen Dokumente bislang nur für den Menschen voll verständlich. Zudem entsprechen die Resultate trotz verbesserter Such- und Rankingtechniken oftmals nicht den Intentionen der Benutzer.

In *SmartWeb* werden die führenden Forscher aus dem Bereich der Intelligenten Benutzerschnittstellen, des SemanticWeb und der Informationsextraktion Methoden und Technologien erforschen und umsetzen, um diese Hürden zu beseitigen. Das Semantische Web basiert auf der inhaltlichen Beschreibung digitaler Dokumente mit standardisierten Vokabularen, die eine maschinell verstehbare Semantik haben. Damit wird der Übergang von einem "Netz aus Verweisstrukturen" zu einem "Netz aus Inhaltsstrukturen" vollzogen. Dies eröffnet völlig neue Dimensionen in den Bereichen Internetdienste, Information Retrieval, Mobile Computing, E-Commerce und E-Work.

SmartWeb bildet eine wichtige Stufe bei der Realisierung des Internets der nächsten Generation, das breitbandige Multimediadienste mobil und individualisiert bereitstellen wird. *SmartWeb* ist auf der Ebene der Softwaresysteme angesiedelt, welche die Infrastruktur für spezielle Anwendungsprojekte zur Implementierung neuartiger Mehrwertdienste im Internet der nächsten Generation bildet. *SmartWeb* ist abgestimmt auf die Entwicklungen auf dem Gebiet des Mobiltenet und der darunter liegenden Schicht der Hochleistungsfestnetze, welche die Mobilfunkstationen breitbandig mit Datenströmen versorgen. Damit nimmt *SmartWeb* eine zentrale Stellung bei der Verbindung innovativer Kommunikationstechnologien mit völlig neuartigen Anwendungsfeldern. *SmartWeb* baut auf die Konvergenz der verschiedenen Mobilfunktechnologien auf und sichert eine bedarfsgerechte Informationsversorgung sowie nutzerzentrierte Informationslogistik.

8. Ausblick/Schluss

Bereits heute ist erkennbar, dass Spracherkennungstechnologie den Umgang von Menschen mit Computern und Maschinen weitreichend verändert hat und weiter verändern wird. Es wurde auf den vorigen Seiten schon auf die vielfältigen Einsatzmöglichkeiten und existierende Anwendungen in diesen Bereichen eingegangen. Die Steuerung von Geräten und Maschinen sowie automatisierte telefonische Kommunikation mit Voice-Portalen werden in Zukunft ebenso alltäglich sein, wie heute schon das gewohnte Surfen im Web.

Aufbauend auf den Möglichkeiten moderner Spracherkennung und Sprachsteuerung sind viele Unternehmen heute schon dabei, Projekte zur Integration solcher Technologien in ihre Produkte und Abläufe voranzutreiben. Eine wichtige Rolle spielt in diesem Bereich die Automobilindustrie und die Telekommunikationsindustrie, die durch ihre hohe Breitenwirkung beim Endkunden die Akzeptanz der Spracherkennung noch steigern werden. Auch der Callcenterbereich treibt den Markt voran, da sich in diesem Sektor enorme Automatisierungs- und Rationalisierungspotentiale in der Kundenkommunikation ergeben.

Ein weiterer Trend der sich in Zukunft abzeichnen wird, ist die Verschmelzung der verschiedenen Kommunikationsmöglichkeiten im Bereich Mensch-Maschine Kommunikation, hin zu übergreifenden sogenannten „*Multichannel-Ansätzen*“. Die Kommunikation mit Datenbanken und Applikationen oder die Steuerung von Geräten kann über verschiedensten Kanäle erfolgen, die der jeweiligen Situation des Benutzers angepasst sind.