Zeitschrift: Bulletin des Schweizerischen Elektrotechnischen Vereins, des Verbandes Schweizerischer Elektrizitätsunternehmen = Bulletin de l'Association suisse des électriciens, de l'Association des entreprises électriques suisses

Herausgeber: Schweizerischer Elektrotechnischer Verein ; Verband Schweizerischer Elektrizitätsunternehmen

Band: 87 (1996)

Heft: 19

Artikel: Warum Computer sich mit der menschlichen Sprache schwer tun : Stand und neue Ansätze in der Spracherkennung

Autor: Pfister, Beat

DOI: https://doi.org/10.5169/seals-902359

Mitte der siebziger Jahre herrschte in Fachkreisen die Meinung, dass dank leistungsfähigen Digitalrechnern das Problem der Spracherkennung innert weniger Jahre gelöst sein werde. Heute sind namhafte Forscher der Ansicht, dass auch in 20 Jahren Computer den Menschen noch nicht zu konkurrenzieren vermögen, wenn es um das Verstehen gesprochener Sprache geht. Dieser Beitrag zeigt den Stand, den die Forschung im Bereich Spracherkennung erreicht hat, mit welchen Methoden gearbeitet wird und in welche Richtung neue Ansätze an der ETHZ gehen.

Warum Computer sich mit der menschlichen Sprache schwer tun

Stand und neue Ansätze in der Spracherkennung

■ Beat Pfister

Unter dem Begriff Spracherkennung können verschiedene Aufgaben verstanden werden, die sich nicht zuletzt im Schwierigkeitsgrad extrem unterscheiden. So ist es relativ einfach, einen Spracherkenner zu verwirklichen, der unter günstigen Umständen einige Dutzend Wörter unterscheiden kann, die stets von derselben Person gesprochen werden. Trotzdem ist auch ein solcher Spracherkenner, der genauer als *sprecherabhängiger Worterkenner für kleines Vokabular* bezeichnet wird, nicht trivial, weil kein Mensch dasselbe Wort mehrmals exakt gleich aussprechen kann, wie dies aus Bild 1 ersichtlich ist.

Erheblich schwieriger wird es für die Spracherkennung, wenn die Wörter von beliebigen Personen gesprochen werden. Ein derartiger *sprecherunabhängiger Worterkenner* muss sehr stark unterschiedliche Sprachsignale, wie sie beispielsweise in Bild 3 dargestellt sind, demselben Wort zuordnen können.

Eine noch viel grössere Knacknuss stellt sich der Spracherkennung im Falle der *kontinuierlich gesprochenen Sprache:* Eine Äusserung ist nun nicht mehr ein Wort, sondern typischerweise ein Satz mit einer unbekannten Anzahl von Wörtern, wobei zwischen den Wörtern im allgemeinen keine Pausen sind und vorhandene Pausen meistens keine Wortgrenzen markieren, sondern Verschlusslaute.

Selbstverständlich wirken sich bei jeder Spracherkennungsaufgabe die phonetische Ähnlichkeit der zu erkennenden Wörter und Ausdrücke, das akustische Umfeld der sprechenden Person (Umgebungslärm und Raumakustik) und die Art der Signalübertragung (z.B. über das Telefon) sehr stark auf die erzielbare Erkennungsrate aus. Generell gilt: Je stärker die Veränderung oder Beeinträchtigung des Sprachsignals, desto schwieriger ist es, eine zuverlässige Spracherkennung zu erreichen.

Im folgenden werden für die drei obenerwähnten Spracherkennungsaufgaben je ein heute gebräuchlicher Ansatz und seine Vor- und Nachteile aufgezeigt.

Worterkennung durch Zuordnung zu Mustern

Der einfachste Fall der Spracherkennung, die Erkennung von Wörtern, die stets vom selben Sprecher gesprochen werden, kann mit einem Mustererkennungsansatz gelöst werden. Das Vorgehen ist wie folgt: Die zu erkennenden Wörter werden gesprochen und als Referenzen abgespeichert. Diese Vorbereitung wird

Adresse des Autors
Dr. *Beat Pfister*, Institut für Technische Informatik ETH Zentrum, 8092 Zürich

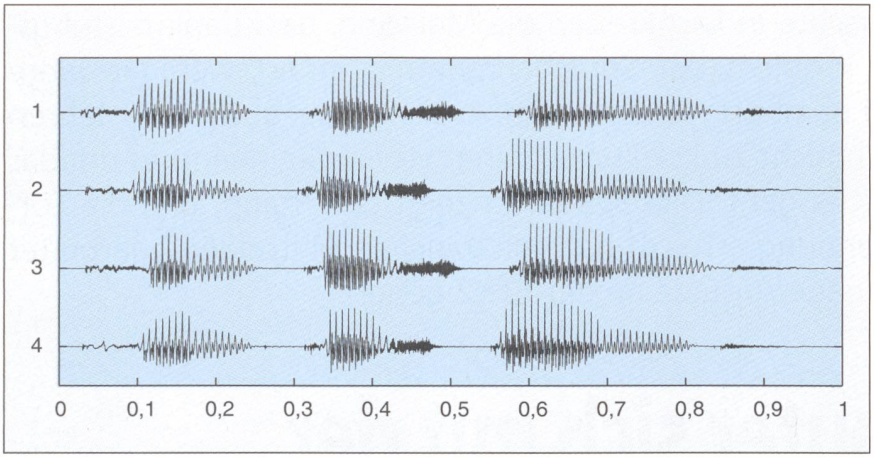

Bild 1 Das Oszillogramm zeigt, dass das von einem Sprecher viermal gleich gesprochene Wort «Kontostand» jedesmal ein bisschen anders ausfällt.

gewöhnlich als Training bezeichnet. Nach dem Training kann der Erkenner ein neu gesprochenes Wort mit allen Referenzen vergleichen und dasjenige Wort als erkannt melden, dessen Referenz am besten mit dem neu eingegebenen Wort übereinstimmt.

Man benützt dabei also die Tatsache, dass eine Person dasselbe Wort zwar nicht genau gleich, aber doch recht ähnlich artikuliert. Wie Bild 1 zeigt, eignet sich jedoch der zeitliche Verlauf des Sprachsignals (Oszillogramm) nicht direkt für den Mustervergleich, weil beispielsweise das Differenzsignal zwischen zwei Äusserungen kein gutes Mass für die Ähnlichkeit zweier Sprachmuster ist.

Die einzelnen Äusserungen unterscheiden sich nicht nur in der Signalform, sondern auch hinsichtlich der zeitlichen Struktur und der Grundfrequenz (bzw. der Periodendauer) in den stimmhaften Segmenten. Um Sprachmuster miteinander vergleichen zu können, muss also die Ähnlichkeit so gemessen werden, dass geringere Veränderungen der zeitlichen Struktur und der Tonhöhe das Ähnlichkeitsmass nicht wesentlich beeinflussen.

Um dies zu erreichen, werden die zwei zu vergleichenden Sprachsignale (Referenz- und Testsignal) je in ein geglättetes Kurzzeitspektrum umgewandelt, wie es in Bild 2 gezeigt wird. Zur Bestimmung des Kurzzeitspektrums wird das Sprachsignal in Abschnitte von 30 ms unterteilt, die sich um 20 ms überlappen, und aus jedem dieser Abschnitte wird das Fourierspektrum berechnet. Durch das Glätten wird der Einfluss der Sprachgrundfrequenz auf das Ähnlichkeitsmass eliminiert.

Um nun den Vergleich noch von Variationen der zeitlichen Struktur unabhängig zu machen, wird die Zeitachse des Test-Kurzzeitspektrums so verzerrt, dass es zeitlich optimal zur Referenz passt. Man nennt diese Operation Zeitnormalisation. Dabei werden an den Stellen, wo das Test-Kurzzeitspektrum zu strecken beziehungsweise zu stauchen ist, einzelne Teilspektren verdoppelt beziehungsweise eliminiert. Diese Optimierung wird mittels dynamischer Programmierung gelöst. Schliesslich wird als Ähnlichkeitsmass die mittlere Distanz zwischen den geglätteten und zeitlich angeglichenen Kurzzeitspektren bestimmt.

Die Vorteile dieses Spracherkennungsverfahrens sind, dass der Benutzer eines solchen Erkenners das zu erkennende Vokabular frei zusammenstellen kann und dass das Verfahren bei der Erkennung relativ wenig Rechenleistung benötigt. Auf der andern Seite sind Mustererkenner sprecherabhängig, das heisst, sie haben den Nachteil, dass ein Erkenner, der für einen Sprecher trainiert worden ist, in der Regel für andere Sprecher unbefriedigend funktioniert.

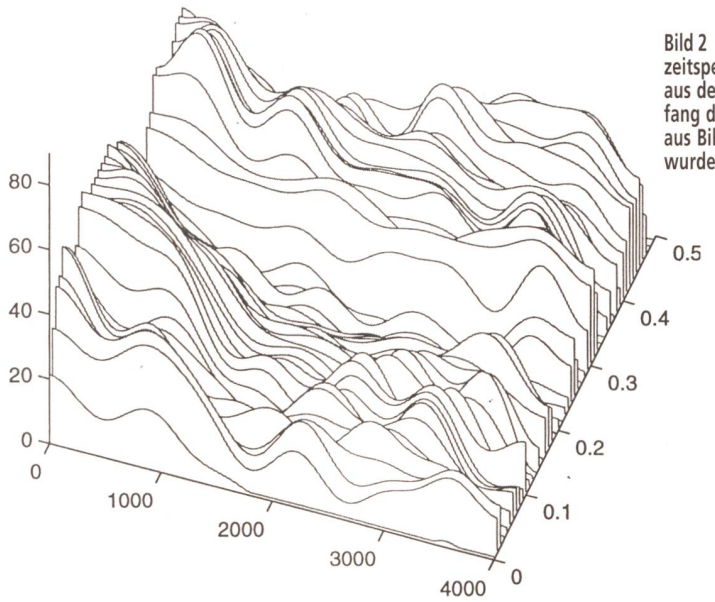

Bild 2 Geglättetes Kurzzeitspektrum, welches aus dem 0,5 s langen Anfang des ersten Signals aus Bild 1 ermittelt wurde.

Worterkennung mittels statistischer Modelle

Werden die zu erkennenden Wörter von verschiedenen Personen gesprochen oder über stark verschiedene Übertragungskanäle zum Erkenner geschickt, dann ist die Variation der Sprachsignale so gross, dass mit dem Mustererkennungsansatz keine zuverlässige Erkennung erzielt werden kann. Die grosse Variabilität, wie sie Bild 3 veranschaulicht, ist zwar komplex, aber nicht rein zufällig. Sie kann mit statistischen Mitteln beschrieben werden.

Ein mathematisches Modell, das sich zur statistischen Beschreibung von Sprachsignalen als sehr gut geeignet herausgestellt hat und heute praktisch als Standard gilt, ist der HMM-Ansatz (Hidden Markov Model). Eine kurze Einführung dazu findet sich im Anhang. Um beispielsweise für das Wort «sieben» ein Hidden Markov Model zu bestimmen, das für Sprachbeispiele, wie sie in Bild 3 gezeigt sind, eingesetzt werden kann, wird folgendermassen vorgegangen:

Datenbeschaffung: Zuerst muss für das Wort «sieben» repräsentatives Sprachmaterial beschafft werden, das heisst, das Wort muss von mehreren hundert Personen gesprochen und aufgenommen werden.

Merkmalsextraktion: Wie bei der Mustererkennung (siehe oben) ist auch bei der statistischen Methode das Sprachsignal als Beobachtungsgrösse (des HMM) schlecht geeignet. Das Sprachsignal wird deshalb wiederum abschnittweise vorverarbeitet. Konkret werden aus jedem Signalabschnitt gewisse Merkmale bestimmt. Üblich sind beispielsweise das geglättete Kurzzeit-

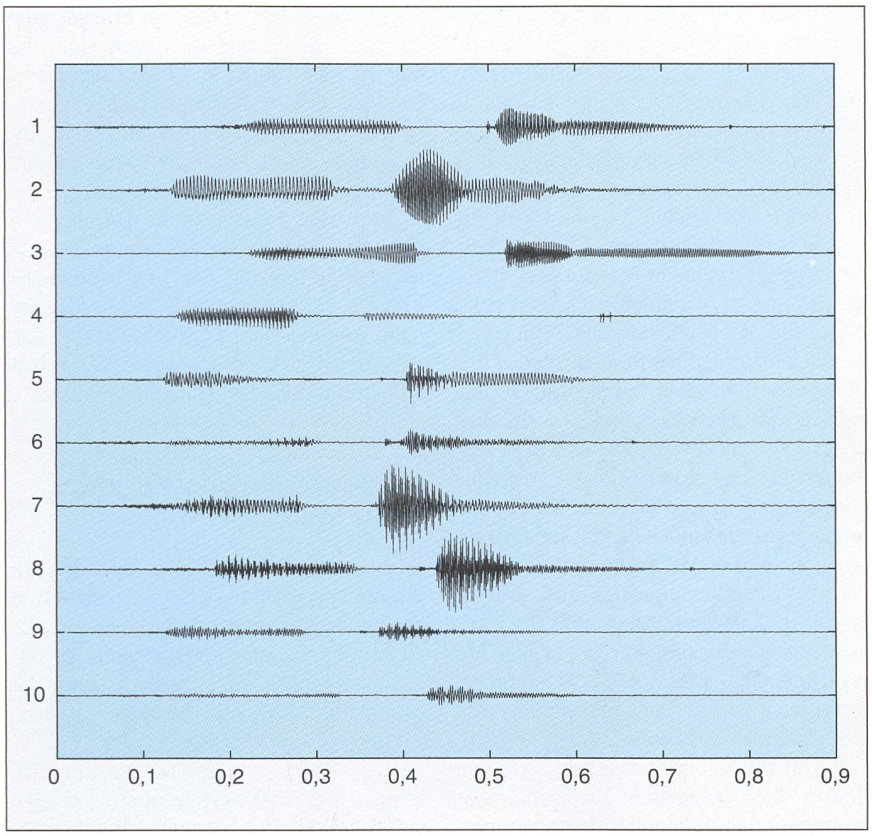

Bild 3 Wenn zehn Personen das Wort «sieben» über verschiedene Telefonverbindungen sprechen, dann resultieren sehr unterschiedliche Sprachsignale.

spektrum und die lokale Signalleistung sowie deren zeitliche Ableitungen.

HMM-Training: Bestimmung der Modellgrössen (Werte A und B_j) so, dass das HMM die vorliegenden Merkmalssequenzen mit maximaler Wahrscheinlichkeit produziert. Dazu wird das Baum-Welch-Verfahren (siehe beispielsweise [1] oder [2]) eingesetzt.

Für jedes Wort des zu erkennenden Vokabulars muss dieses Prozedere durchgeführt werden, bis für jedes Wort ein sogenanntes Ganzwort-HMM vorliegt.

Beim Einsatz dieser Ganzwort-HMM zur Erkennung eines gesprochenen Wortes wird zuerst das Sprachsignal in derselben Art vorverarbeitet wie für das Training. Aus dem Sprachsignal wird also eine Merkmalssequenz ermittelt. Anschliessend wird für jedes HMM des Vokabulars die Wahrscheinlichkeit berechnet, mit der dieses HMM diese Beobachtungs- oder Merkmalssequenz erzeugt. Als erkannt wird dasjenige Vokabularwort deklariert, dessen HMM die Beobachtungssequenz mit der grössten Wahrscheinlichkeit generiert.

Der Vorteil eines Worterkenners, der auf Ganzwort-HMM beruht, ist, dass er nach dem Training personenunabhängig eingesetzt werden kann. Bei vielen Anwendungen ist dies unabdingbar. Der wesentlichste Nachteil der Ganzwort-HMM-Erkenner ist, dass eine Erweiterung des Vokabulars sehr aufwendig sein kann, nämlich dann, wenn das für das Training benötigte Sprachmaterial, das von Hunderten von Personen gesprochen werden muss, nicht vorhanden ist.

Dieser Nachteil lässt sich umgehen, indem statt ganzer Wörter als sprachliche Einheiten Wortteile genommen werden, beispielsweise Laute. Nach dem Training der Laut-HMM können zur Erkennung von Wörtern einfach die entsprechenden Laut-HMM zusammengehängt werden. Ein HMM für das Wort «sieben» setzt sich also aus den HMM der Laute /z/,/i:/,/b/, /ə/ und /n/ zusammen (Bezeichnung der Laute in phonetischer Schrift). Wenn einmal alle Laut-HMM vorhanden sind, dann können so beliebige Wörter erkannt werden. Allerdings ist es sehr aufwendig, die Laut-HMM zu trainieren, weil benachbarte Laute sich gegenseitig beeinflussen. Zudem macht ein auf Laut-HMM beruhender Worterkenner mehr Erkennungsfehler als einer mit Ganzwort-HMM.

Erkennung kontinuierlich gesprochener Sprache

In ähnlicher Art und Weise, wie mit Laut-HMM ein Erkenner für Wörter zusammengesetzt werden kann, ist auch aus Wort-HMM ein Erkenner für Sätze machbar. Diese theoretische Möglichkeit entpuppt sich in der Praxis jedoch schnell als nicht gangbarer Weg, weil bei einem gegebenen Vokabular von beispielsweise $K = 1000$ Wörtern die Anzahl der möglichen Sätze mit einer maximalen Länge von L-Wörtern sehr gross ist (ungefähr K^L, wenn jedes Wort an jeder Position im Satz stehen kann) und damit auch die Anzahl der Satz-HMM.

Es wird deshalb nicht wie im Fall der Worterkennung für jeden möglichen Satz ein einzelnes HMM angesetzt, sondern ein Netz mit den Vokabularwörtern als Kanten. In diesem Netz stellt jeder Pfad vom Anfangsknoten zum Endknoten einen möglichen Satz dar. Wenn wiederum jedes Vokabularwort an jeder Position des Satzes stehen kann (Bild 4), ergibt sich dieselbe Anzahl Sätze wie oben, aber die Suche des wahrscheinlichsten Satzes lässt sich in diesem Fall mit dynamischer Programmierung viel effizienter gestalten.

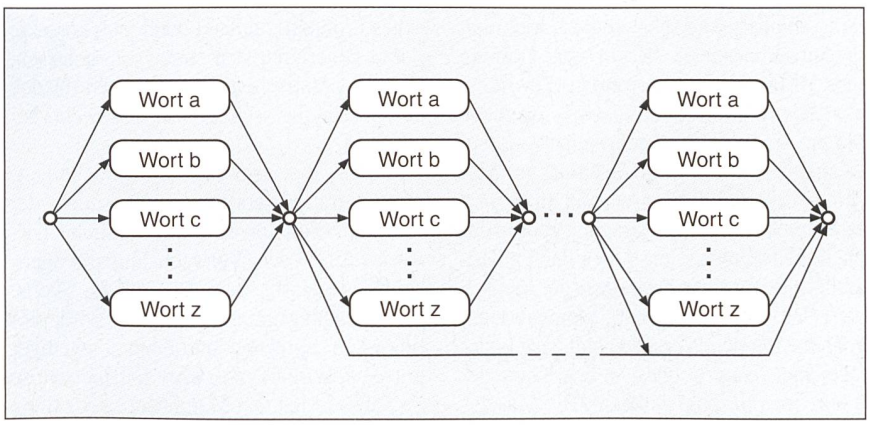

Bild 4 Einfaches Erkennungsnetz für beliebige Folgen von Wörtern aus dem Vokabular $Wort_a$ bis $Wort_z$. Als Kanten figurieren Wort-HMM, wie sie in Bild 5 dargestellt sind.

Da natürlichsprachliche Sätze selbstverständlich nicht Folgen beliebiger Wörter sind, sondern nach bestimmten Regeln aufgebaut sind, kann auf ein gegebenes Wort nicht ein beliebiges Wort des Vokabulars folgen, sondern nur eine relativ kleine Auswahl. So lässt sich die Anzahl der Verzweigungen nach jedem Wort von $K = 1000$ im Mittel auf wenige Dutzend reduzieren. Die Information, welche Wörter nacheinander folgen können und allenfalls mit welcher Wahrscheinlichkeit diese Wortpaare auftreten, muss aus Textmaterial bestimmt werden, das für die betreffende Anwendung repräsentativ ist. Die Reduktion des Verzweigungsfaktors kann also nur über ein anwendungsspezifisches Training erreicht werden. Weil dazu nur Textmaterial nötig ist, also nicht Sprachsignale von Hunderten von Personen wie oben, ist dies nicht so problematisch.

Warum der Mensch Sprache besser erkennt als die Maschine

Es ist eine Tatsache, dass die maschinelle Spracherkennung noch bei weitem nicht das zu leisten vermag, was der Mensch scheinbar ganz mühelos schafft. Die Diskrepanz ist besonders bei der Erkennung kontinuierlich gesprochener Sprache gross. Aber auch bei relativ einfachen Aufgaben, wie dem Erkennen der zehn Ziffern über das Telefon, vermag die Maschine den Menschen (noch) nicht zu konkurrenzieren. Für diese Unterlegenheit gibt es unter anderem die folgenden Erklärungen:

1. Die heute erfolgreichsten Spracherkennungsansätze basieren auf HMM, sind also eine statistische Beschreibung von Sprachsignalstücken, seien diese nun Wörter, Laute oder andere sprachliche Einheiten. Insbesondere bei der Sprachübertragung über das Telefon werden die Sprachsignale, die schon ursprünglich recht verschieden sind (sprecherspezifische Unterschiede), durch die Übertragung zusätzlich stark verändert. Die sehr grosse Variabilität der Sprachbeispiele, die für das Training eines HMM verwendet werden, bewirken, dass das HMM unscharf wird. Ähnliche Wörter sind so nur sehr unzuverlässig auseinanderzuhalten. Im Gegensatz zur Maschine versucht der Mensch sich an die Art der Stimme und die Übertragung zu gewöhnen. Einen Hinweis, dass dies der Fall ist, stellt die Beobachtung dar, dass der Mensch am Telefon oft im ersten Moment auch Probleme mit dem Verstehen hat, sich in der Regel aber relativ schnell an den Klang der zu hörenden Stimme gewöhnt. Die Maschine jedoch arbeitet bei der Erkennung des zehnten Wortes noch immer gleich wie beim ersten. Auf die Spracherkennung bezogen heisst dies, dass eine geeignete adaptive Vorverarbeitung des Sprachsignals dann erfolgversprechend ist, wenn dadurch die Varianz sprachlicher Einheiten reduziert oder die Differenz zwischen sprachlichen Einheiten vergrössert werden kann.

2. Die Merkmale, die in heutigen Spracherkennungssystemen aus dem Sprachsignal extrahiert werden, beschreiben den groben Verlauf des Kurzzeitspektrums und der Signalleistung. Hingegen werden der Verlauf der Sprechmelodie, die Information über die Stimmhaftigkeit und über die zeitlichen Verhältnisse nicht verwertet, obwohl absolut klar ist, dass der Mensch auf diese Information für eine gute Verständigung angewiesen ist. Geflüsterte Sprache oder hinsichtlich Rhythmus oder Sprechmelodie verfremdete Sprache bieten dem Menschen eindeutig mehr Mühe. Es ist deshalb durchaus möglich, dass durch ein geschicktes Ausnutzen von Information, die heute noch als störend eliminiert wird, die Spracherkennung wesentlich verbessert werden kann.

3. Um Sprache zu verstehen, setzt der Mensch alle verfügbaren, zu einem sinnvollen Resultat führenden Informationen ein. Dazu gehören etwa: welche sinntragenden Elemente es in der Sprache gibt (lexikalische Information), wie daraus Wortformen erzeugt werden (Morphologie), welche Wortformen korrekte Ausdrücke oder Sätze bilden (Syntax), welche Bedeutung Wörter und Ausdrücke haben (Semantik) usw.

Im Vergleich zum Menschen machen heutige Spracherkennungssysteme nur sehr begrenzt Gebrauch von all dieser verfügbaren Information. Das Problem liegt einerseits darin, dass die Informationsmenge enorm gross ist, anderseits gibt es im Zusammenhang mit natürlicher Sprache sehr viel Mehrdeutigkeiten. Da es bisher beispielsweise noch nicht gelungen ist, eine einigermassen umfassende, maschinengerechte Spezifikation der Syntax für Deutsch zu verwirklichen (linguistisches Problem), anhand derer sich korrekte Sätze generieren oder analysieren lassen, sind die praktischen Möglichkeiten für den Einsatz linguistischen Wissens recht beschränkt.

4. Für die maschinelle Spracherkennung wirkt sich erschwerend aus, dass der Mensch beim Sprechen die lautliche Differenzierung den Verwechslungsmöglichkeiten anpasst, das heisst, wo keine Verwechslungsgefahr besteht, wird bei flüssigem Sprechen automatisch weniger präzis artikuliert. So wird beispielsweise im Satz «Er hat den Mann heute getroffen» statt /. . . dən man . . . / bei schneller Sprechweise eher /. . . dəman . . ./ gesprochen. Das bedeutet, dass im Sprachsignal gar nicht alle Laute zu finden sind, die aufgrund der Wörter vorhanden sein müssten. Derartige Auslassungen oder Veränderungen von Lauten können nur unter Anwendung sprachlichen Wissens erkannt und kompensiert werden.

Die heute verfügbaren Spracherkennungsverfahren gehen nur sehr beschränkt auf diese Probleme ein. Für wesentliche Fortschritte, insbesondere in der Erkennung kontinuierlich gesprochener Sprache, ist jedoch eine eingehendere Beschäftigung mit diesen Problemen unabdingbar.

Die Spracherkennungsforschung an der ETH Zürich

Die allgemeine Lösung des Spracherkennungsproblems, die alles verfügbare akustische und linguistische Wissen unter Anwendung aller Möglichkeiten der Signalverarbeitung, der Statistik und der Informatik einbezieht, ist noch nirgends auf der Welt verwirklicht worden.

Es soll hier nicht der falsche Eindruck erweckt werden, dass wir nun als einzige auf der Welt den richtigen Weg eingeschlagen haben, aber wir beabsichtigen tatsächlich, mit unserem Ansatz einen wichtigen Schritt in Richtung der allgemeinen Lösung zu tun. Dieser Ansatz geht davon aus, für jedes Teilproblem der Spracherkennung methodisch angepasst vorzugehen und die verschiedenen Komponenten eng zu koppeln.

Bezüglich der Methoden heisst dies, dass beispielsweise für die Erkennung lautlicher Ereignisse in einem Sprachsignal sinnvollerweise statistische Methoden (Markov-Modelle, neuronale Netze usw.) eingesetzt werden, weil die charakteristischen Merkmale der Laute eine grosse Varianz aufweisen. Hingegen ist linguistisches Wissen in den Bereichen Morphologie, Syntax, Semantik usw. zweckmässigerweise wissensbasiert anzuwenden (in der Form von Lexika und Produktionsregeln), weil sich beispielsweise die Konjugationsformen eines Verbs kaum statistisch sinnvoll beschreiben lassen.

Die enge Kopplung aller Komponenten ist deshalb erforderlich, weil der Spracherkenner nur aufgrund aller wesentlichen Anhaltspunkte (akustische, lexikalische, syntaktische, anwendungsspezifische usw.) aus dem Sprachsignal entscheiden kann, was die Aussage ist. Eine Verarbeitung, in der alle Komponenten sequentiell durchlaufen werden, ist in diesem Fall nicht zweckmässig.

Der Ansatz für unser Spracherkennungssystem Arcos, das aus den beiden

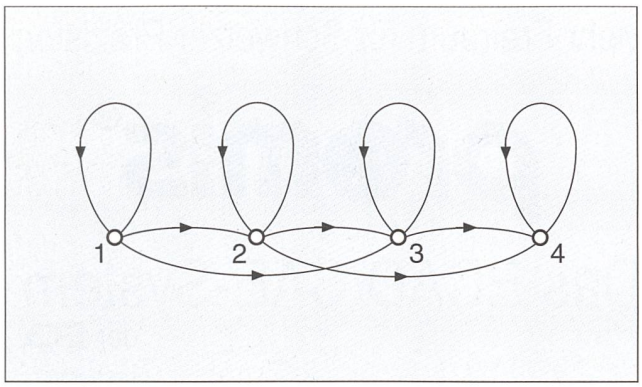

Bild 5 Markov-Modell mit vier Zuständen
Nur die von Null verschiedenen Zustandsübergangswahrscheinlichkeiten sind als Bögen eingetragen.

Teilen *Grundelement-Erkenner* und *linguistische Stufe* besteht, trägt diesen Gesichtspunkten insofern Rechnung, als der erste Teil statistisch gelöst wird (neuronales Netz kombiniert mit HMM) und die linguistische Stufe wissensbasiert ist (Finite State Transducer, Chart-Parser, Lexika usw.). Um die enge Kopplung zu erreichen, werden alle Komponenten von einer zentralen Einheit gesteuert, die als Strategiekomponente bezeichnet wird.

Anhang

Ein HMM ist ein doppelt stochastischer Prozess, wovon der eine beobachtbar und der andere verdeckt (hidden) ist. Der verdeckte Prozess ist ein Markov-Prozess erster Ordnung mit N–Zuständen, das heisst, falls sich der Prozess zum Zeitpunkt t–1 im Zustand i befunden hat, dann geht er zur Zeit t mit der Wahrscheinlichkeit a_{ij} in den Zustand j über. In der Spracherkennung werden gewöhnlich nur für einen Teil der a_{ij} Werte zugelassen, die grösser als Null sind. Dies sind sogenannt nichtergodische Modelle, wie es als Beispiel mit vier Zuständen in Bild 5 dargestellt ist.

Der zweite Zufallsprozess besteht aus n unabhängigen Zufallsvariabeln, je eine pro Zustand, welche die beobachtbaren Ausgaben $O(t)$ des HMM produzieren, wobei nur der jeweils aktive Zustand eine Ausgabe tätigt. Diese Ausgaben werden pro Zustand j durch eine Beobachtungs-Wahrscheinlichkeitsverteilung B_j bestimmt. Ein HMM mit der in Bild 5 gezeigten Konfiguration wird also durch die folgenden Parameter beschrieben:

N	Zahl der Zustände
$A = a_{ij}$	Zustandsübergangs-Wahrscheinlichkeitsmatrix ($N \times N$-Matrix)
B_j	Beobachtungs-Wahrscheinlichkeitsverteilung pro Zustand (diese kann kontinuierlich oder diskret sein)

Im allgemeinen kommen noch die Anfangszustands-Wahrscheinlichkeiten dazu. Weil jedoch im vorliegenden Fall zu Beginn nur der Zustand 1 eingenommen werden kann, sind die andern Anfangszustände ausgeschlossen. Während N ein wählbarer Parameter ist, werden die Grössen A und B_j aus Sprachsignalen berechnet.

Literatur

[1] *L. R. Rabiner* und *B. H. Juang:* An Introduction to Hidden Markov Models. IEEE ASSP Magazine, 3(1), January 1986.

[2] *L. R. Rabiner:* Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. of the IEEE, pages 257–286, February 1989.

Pourquoi les ordinateurs ont tant de peine avec le langage humain

Etat actuel et nouvelles solutions dans le domaine de l'identification de la parole

Au milieu des années soixante-dix, les spécialistes étaient en majorité d'avis que les ordinateurs numériques performants permettraient de résoudre en quelques années le problème de l'identification de la parole. Actuellement, des chercheurs de renom pensent que même dans 20 ans, l'ordinateur ne pourra pas encore faire concurrence à l'homme en matière de compréhension de textes parlés. Le présent article montre où en est la recherche dans le domaine de l'identification de la parole, les méthodes appliquées et la direction des nouvelles solutions envisagées à l'EPF de Zurich.