# Mit dem Computer reden

***IBM-Produkte wie Voicetype und Dragon Dictate sollen gesprochene Wörter und Sätze für den PC verständlich machen. CHIP erläutert das Prinzip dieser Spracherkennungsprogramme.***

Fiorito

Die Erkennung der menschlichen Sprache per Computer hat viele Anwendungsmöglichkeiten: die Bedienung moderner Maschinen, die „Fernbedienung“ vom PC aus einigen Metern Abstand, das Diktieren von Briefen und so weiter. Produkte für den Einsatz auf dem PC reichen von Lösungen mit kleinem Wortschatz wie Microsofts Sound System und das sprecherunabhängige Speechpower bis hin zu Diktiersystemen wie Dragon Dictate.

OS/2 Warp 4 (Codename Merlin), IBMs neue PC-Betriebssystem-Version, hat Spracherkennung sogar eingebaut. Untypisch für Big Blue, wurde das Sprachmodul portiert und ist als *Voicetype 3.0 für Windows 95* erhältlich. Ein Pentium-PC mit 16 Megabyte Arbeitsspeicher und Soundkarte werden von IBM als Mindestvoraussetzung genannt. Das verwendete Mikrofon sollte – wie bei anderen Produkten auch – nicht vom Wühltisch sein, denn das Ergebnis des Diktiervorgangs ist von der Qualität des Spracheingabematerials stark abhängig.

Ein prominentes Beispiel für die isolierte Einzelworterkennung (meist Befehle) sind Telefon-Sprachcomputer in Firmen, die Kunden ohne klassische Telefonzentrale mit der gewünschten Abteilung verbinden. Der informationstechnische Aufwand für solche Sprachcomputer ist vergleichsweise gering, da sich der Wortschatz dieser Programme auf einige Dutzend Wörter beschränkt.

Diktiersysteme dagegen sind eine hochentwickelte Form der Spracherkennung. Aber auch sie arbeiten nur mit isoliert gesprochenen Wörtern zufriedenstellend, da sich das System zur Restrukturierung des Redeflusses in Einzelwörter umfassende Kenntnisse über den Kontext jeder Wortgruppe verschaffen müßte. Voicetype beispielsweise fordert, etwas gewöhnungsbedürftig, mindestens 100 Millisekunden Pause zwischen zwei Wörtern.

Problematisch bei der Spracherkennung sind zusammengesetzte Wörter (Komposita). Beispiel „Untersuchungsleiter“: Das Wissen um „Untersuchung“ und „Leiter“ bringt den Algorithmus des Programms nicht weiter. Erst durch eine umfassende Kontextanalyse würde klar, daß ein Polizeibeamter und nicht ein neuartiger Operationstisch gemeint ist.

Die dudenkonforme Groß- und Kleinschreibung in der Schriftsprache ist gleichfalls oft nur durch Kenntnis des Kontextes realisierbar. Das dies auch inhaltlich relevant sein kann, verdeutlichen „Der geliebte Floh“ und „Der Geliebte floh“.

## Algorithmen zur Spracherkennung

Die kleinste Einheit der gesprochenen Sprache ist das 10 bis 40 Millisekunden andauernde Phonem. Ein Wort besteht in der Regel aus mehreren Phonemen. Im ersten Schritt der Spracherkennung wird etwa alle 10 Millisekunden das akustische Kurzzeit-Frequenzspektrum mittels Fourier-Transformation hergestellt.

Als Ergebnis dieser digitalen Signalumwandlung liegt ein Frequenzintensitäts-Diagramm vor, dessen Frequenzlinien zu numerischen Vektoren vereinheitlicht und mit Muster- oder Referenzvektoren des Systems verglichen werden. Der Vergleich ist sehr rechenintensiv und muß daher möglichst stark optimiert sein. Drei Verfahren haben sich dabei durchgesetzt: die dynamische Programmierung, die Darstellung mittels Hidden-Markov-Modellen und die Künstliche Intelligenz.

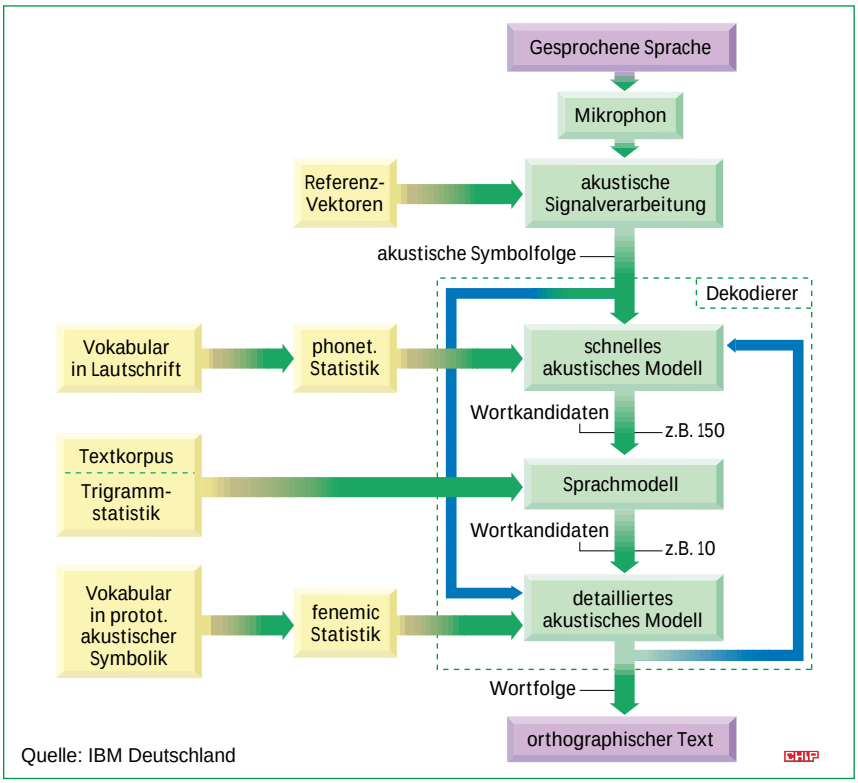
Bei der dynamischen Programmierung wird der Vergleich zwischen Kurzzeit- und Mustervektoren durch einen rekursiven Algorithmus realisiert. Das bedeutet, daß sich die Vergleichsprozedur wieder selbst mit neuen, verbesserten Parametern aufruft. Ziel ist die Bestimmung des Referenzmusters, das die höchste Übereinstimmung mit den Ausgangsdaten aufweist.

Die Erkennungsmethode mit Hidden-Markov-Modellen berechnet die Übergangswahrscheinlichkeiten von einem Phonem zum nächsten. Zum Erzeugen der benötigten Modellklassen ist eine Trainingsphase des Systems nötig. In der Erkennungsphase wird aus dem zu erkennenden Musterverlauf unter Zuhilfenahme der erlernten Modellklassen die größte Erzeugungswahrscheinlichkeit bestimmt. Der Rechenaufwand für Hidden-Markov ist enorm; er potenziert sich mit der Wortlänge. In der Praxis finden darum verkürzte, optimierte Abwandlungen des Verfahrens Verwendung.

Die Erkennung mittels neuronaler Netze (Künstliche Intelligenz) ist an die Funktion natürlicher Nervenzellen (Neuronen) angelehnt. Die Modellierung dieser „Intelligenz“ funktioniert in groben Zügen so: Ein Eingangs-(Bit-)Muster – zum Zwecke der Spracherkennung sind es die Vektoren eines Wortes – wird auf die Eingangsschicht des neuronalen Netzes gelegt. An der Ausgangsschicht liegt dann das Muster für das Wort selbst an. Zwischen Eingangs- und Ausgangsschicht liegen ein oder mehrere (verborgene) Zwischenschichten.

Beginnend bei der Eingangsschicht befinden sich von jedem „Neuron“ Gewichte zu allen „Neuronen“ der nächsthöheren Schicht. Im Verlaufe der Trainingsphase des Netzes verschieben sich diese Gewichte zugunsten der trainierten Eingangsmuster. Diese „unscharfe Logik“ ist in der Arbeitsphase ein geeignetes Mittel, um Ähnlichkeiten herauszufinden. Auch zur Schrifterkennung (OCR) wird sie häufig eingesetzt.

**Ablauf der Spracherkennung bei IBMs Voicetype**




Quelle: IBM Deutschland

## IBM Voicetype dekodiert dreistufig

Voicetype von IBM arbeitet folgendermaßen (siehe Bild oben): Der Redefluß wird in eine Symbolfolge der Phoneme zerlegt. Der eigentliche Sprachdekodierer ist dreistufig: In der ersten Stufe (schnelles akustisches Modell) werden die Wörter mit dem Hidden-Markov-Modell selektiert, welche die größte Wahrscheinlichkeit für das richtige Wort besitzen. Die Zahl der Wortkandidaten beträgt etwa 150. Zu jedem Phonem gibt es ein Markov-Modell.

In der zweiten Stufe (Sprachmodell) wird die Anzahl der Kandidaten auf zehn bis 20 Wörter eingeschränkt. Entscheidend ist, mit welcher Schreibweise das Wort mit zwei anderen Wörtern am häufigsten aufgetreten ist. Diese Kontextprüfung heißt „Trigrammtechnik“ und erlaubt die sichere Unterscheidung auch sehr ähnlich klingender Wörter. Hier wird auch im sogenannten „Cache Language Model“ im Laufe der Benutzung eine Datenbank mit neu erlernten Wörter und ihren Trigrammstatistiken gespeichert.

Die dritte Stufe (detailliertes akustisches Modell) ordnet die verbliebenen Wortkandidaten zu einer Art Hitliste. Das Auswahlverfahren dafür ist dem des schnellen Modells vergleichbar mit dem Unterschied, daß als Vergleichsmaterial nicht die Lautschrift, sondern Symbolfolgen für Phoneme verwendet werden. Das Wort an Platz 1 der Hitliste wird wieder der ersten Stufe zugeführt und durchläuft so ein zweites Mal die Dekodierkette.

Das Grundvokabular von Voicetype beträgt rund 30 000 Wortformen und kann um 34 000 benutzerdefinierte Wörter ergänzt werden. Nach etwa zweistündigem Training auf die Sprechweise des Benutzers ist die Erkennungsrate etwa 95 Prozent. *Jan Kleinert*

Diessner, Herwig: Spracherkennung auf dem Computer; Diplomarbeit; Fraunhofer-Institut für Arbeitswissenschaft und Organisation

Voicetype Diktiersystem 3.0 für Windows 95; ca. 1500 Mark; IBM Deutschland, Tel. (018 03) 31 32 33, *http://www.software.ibm.com/workgroup/voicetyp*

Als Demoversion gab es Voicetype auch auf der CD-ROM zur CHIP-Ausgabe 11/96 (zu bestellen beim CHIP-Leserservice, Vertrieb 731, 97064 Würzburg; siehe auch Impressum)

Dragon Dictate 2.2, ca. 1690 Mark, Dragon Systems, Im Buhles 4, 61479 Glashütten, Tel. (061 74) 9 66 10

Speechpower; ca. 980 Mark; S. Punkt – Gesellschaft für Software, Pauwelstraße 19, 52075 Aachen, Tel. (02 41) 446 84 00