



Mann und Maschine, ins Gespräch vertieft: Was in Stanley Kubricks Science-Fiction-Film „2001: Odyssee im Weltraum“ noch ein Gedanken-spiel war, wird nun Wirklichkeit. Computer, die menschliche Kommunikation immer besser verstehen, sollen in den nächsten Jahren auf den Markt kommen. Linguisten haben entdeckt, dass maschinelle Spracherkennung nach den gleichen Mustern wie beim Menschen funktioniert.

Wenn der **Computer** aufs Wort gehorcht

*Sprache funktioniert nur, weil es bestimmte Wahrscheinlichkeiten und Regeln gibt, die Sprecher und Hörer gleichermaßen verstehen. Erst mithilfe dieser grundlegenden Muster kann der Mensch einzelne Laute zu Wörtern zusammensetzen und den Sinn erkennen. So lautet die zentrale These der rund 40 Wissenschaftler, die sich auf Einladung des **MAX-PLANCK-INSTITUTS FÜR PSYCHOLINGUISTIK** im holländischen Nijmegen zum Workshop „Spracherkennung als Klassifizierung von Mustern“ trafen.*

„Computer, sag’ mir alles über den Planeten Erde!“ In der Fernsehserie „Raumschiff Enterprise“ erteilt Captain Kirk seinen Bordgeräten Befehle per Zuruf. Maus und Tastatur sind überflüssig. Intelligente Computer, die unsere Sprache verstehen und mit Menschen problemlos sprechen können, haben nicht nur Fernsehproduzenten und Filmemacher wie Stanley Kubrick in Bann geschlagen, dessen legendäre Denkmaschine „HAL“ in „2001: Odyssee im Weltraum“ den Menschen von den Lippen lesen konnte und zum Leidwesen der Astronauten bald auch ein Eigenleben entwickelte.

FOTO: DEFF-MOVIES

Was lange Zeit als kühne Vision galt, beginnt allmählich Wirklichkeit zu werden. Für die automatische Spracherkennung und für die Sprachsynthese, also die künstliche Spracherzeugung, interessieren sich neben den Forschern auch Wirtschaftsunternehmen. Die Erwartungen an neue Technologien sind groß: Spracherkennung sei „die Zukunft des Computers“ hat Microsoft-Gründer Bill Gates unlängst verkündet. Doch bei aller Euphorie über Wachstumschancen und Wissensgewinn müssen sich die Wissenschaftler eingestehen: Von menschlicher Sprachkompetenz sind die Maschinen noch ein gutes Stück entfernt.

Solange es keinen intelligenten, verständigen Computer wie „HAL“ gibt, müssen die Maschinen bei den Menschen in die Lehre gehen. Das Erstaunliche dabei ist: Sie müssen genau die Sprachsysteme lernen, die sich kleine Kinder im Alter von wenigen Monaten scheinbar mühelos aneignen. Das erste, was Kleinkinder verstehen lernen, sind Sprachlaute, die einfache Wörter wie „Mama“ oder „Papa“ bilden. Schon bald erkennen Kinder unbewusst, dass die Aussprache dieser ganz elementaren Äußerungen auf bestimmten Sprachmustern beruht. Auch wenn es nicht die Mutter ist, die ein Wort ausspricht, sondern ein Fremder, können Kinder das Wort „Mama“ erkennen – ganz egal, wer das Wort ausspricht. Ein solches Muster („M-a-m-a“) muss zu allen möglichen Aussprachevarianten passen, egal, ob der Sprecher nun einen sächsischen, schwäbischen oder bayerischen Dialekt hat – solange die Dialekte nicht zu extrem sind. Andernfalls kann die Übersetzung der akustischen Signale in sinnvolle Wort- und Satzeinheiten nicht funktionieren.

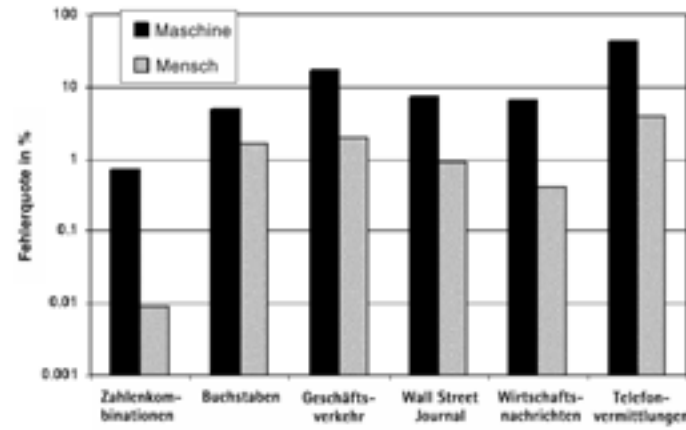
Schon ein so scheinbar einfacher Befund stellt die Wissenschaftler, die sich mit automatischer Spracherkennung befassen, vor erhebliche Probleme. Im Zeitalter der globalisierten

Welt verändert sich unsere Sprache täglich; gleichzeitig wird es immer wichtiger, rasch miteinander zu kommunizieren und Sprachbarrieren zu überwinden. Menschliche und maschinelle Spracherkennung spielt dabei eine erhebliche Rolle, wenn wir wissen wollen, wie Spracherkennung eigentlich funktioniert. Erstmals haben deshalb rund 40 Wissenschaftler ganz unterschiedlicher Fachrichtung auf Einladung des Max-Planck-Instituts für Psycholinguistik im niederländischen Nijmegen drei Tage lang über das Thema „Spracherkennung als Klassifizierung von Mustern“ diskutiert. Die spannende Frage vor Beginn der Konferenz war: Gibt es überhaupt Gemeinsamkeiten zwischen Linguisten, Psychologen, Ingenieuren, Technikern und Computerspezialisten, die sich allesamt mit Teilaspekten der Spracherkennung beschäftigen?

ALLE LERNEN VONEINANDER

Das Ergebnis der Konferenz in Nijmegen ist ermutigend. Wissenschaftler, die sich ausschließlich mit automatischer Spracherkennung beschäftigen, können von ihren Kollegen, die auf menschliche Hörer spezialisiert sind, eine Menge lernen. Umgekehrt lässt sich das Gleiche feststellen. Der Psycholinguist Roel Smits, der am Max-Planck-Institut in Nijmegen forscht und das Treffen mitorganisiert hat, fasst ein Ergebnis der Tagung zusammen: „Seit den achtziger Jahren gab es einen hohen Grad der Spezialisierung in der Spracherkennungsforschung. Jetzt ist die Diskussion zwischen den Fachbereichen wieder in Gang gekommen.“

In Nijmegen wurde deutlich, dass die Wissenschaftler eine gemeinsame Basis haben. Die Forscher stimmen darin überein, dass Sprache nur funktioniert, wenn es bestimmte Wahrscheinlichkeiten und Regeln gibt, die Sprecher und Hörer gleichermaßen verstehen. Erst mithilfe



Linguisten vom Max-Planck-Institut für Psycholinguistik haben die Fehlerquoten bei verschiedenen Textsorten verglichen und herausgefunden, dass menschliche Hörer den Maschinen noch immer weit überlegen sind, selbst bei relativ einfachen Aufgaben wie dem Erkennen von Zahlenreihen.

dieser Regeln kann der Empfänger einer sprachlichen Äußerung einzelne Laute zu Wörtern zusammensetzen und den Sinn erkennen. Der Schritt vom Muster zu einer Kategorie ist für jeden, der sich eine Sprache aneignet, das Ergebnis vieler Erfahrungswerte. Das gilt für Menschen wie für Maschinen. Kleine Kinder etwa müssen akustische Signale immer wieder hören, bis sie Worte und später ganze Sätze verstehen und irgendwann eigenständig bilden können: Sie lernen, einzelne Muster („b“, „a“, „l“) einer bestimmten Kategorie („Ball“) zuzuordnen. Grundlage sind Wahrscheinlichkeiten, nach denen schon Kinder lernen, Kategorien zu bilden. „Menschen sind geborene Statistiker“, sagt Roel Smits.

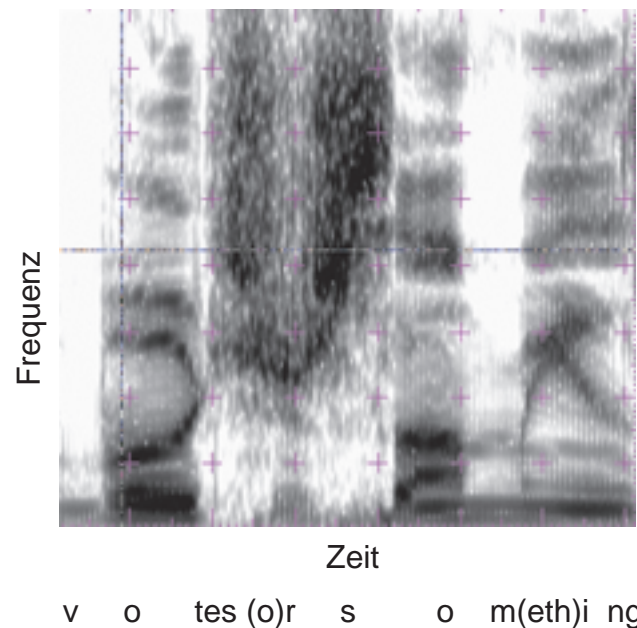
COMPUTER WERDEN GEFÜTTERT

Etwas ganz Ähnliches macht ein Spracherkennungscomputer. Damit eine Maschine sprachliche Muster und Kategorien erkennen kann, muss sie mit einer gewissen Datenmenge „gefüttert“ worden sein. Viele hundert Stunden an Aufnahmen sind nötig; dabei müssen sprachliche Äußerungen unterschiedlicher Sprecher abgespeichert werden. Die Maschine soll schließlich in der Lage sein, genau das zu leisten, was ein Mensch täglich leistet: Etwa im Stimmengewirr eines Kaufhauses

einen bestimmten Sprecher herauszufiltern. Oder einen nuschelnden Gesprächspartner zu verstehen. Und was ist mit einer im Flüsterton gehaltenen Liebeserklärung? Auch diese muss der Sprachcomputer erkennen, wenn er es mit den Menschen aufnehmen will. Im Wettlauf zwischen Mensch und Maschine sind Erstere noch immer klar im Vorteil:

„Hörer passen sich schnell an eine veränderte Umwelt an. Maschinen reagieren dagegen sensibel auf einen Sprecherwechsel oder ein verändertes Sprechtempo. Sie lassen sich zu sehr durch Hintergrundgeräusche ablenken“, sagt der Linguist Smits.

Maschinen, die den natürlichen Redefluss verstehen und sprachliche Eigenheiten wie Akzente oder Dialekte erkennen, sind schon verfügbar, aber noch stark verbesserungsfähig. Der Markt für solche Programme ist zweifellos vorhanden: Amerikanische Unternehmen versuchen bereits, mittels automatischer Spracherkennung Kosten zu kürzen und Personal zu sparen. Wenn solche Voice-Systeme einmal reibungslos funktionieren, werden teure Call-Center der Vergangenheit angehören, und die Telefonistin, zu Beginn des 20. Jahrhunderts ein Symbol des kommunikativen Fortschritts, wäre endgültig Geschichte. Stattdessen



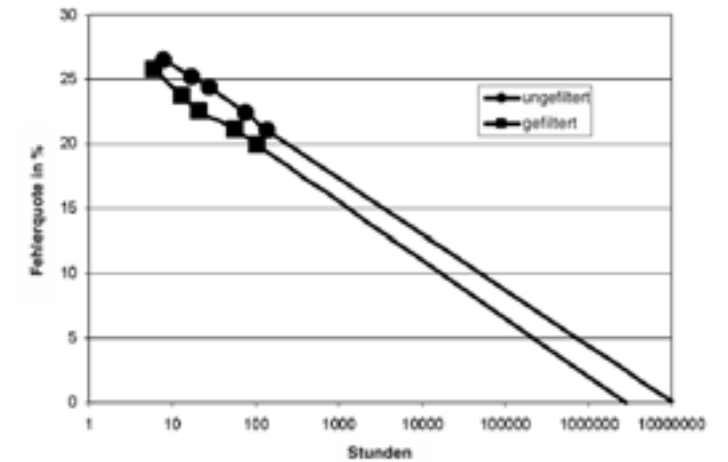
Sprachcomputer zeichnen menschliche Äußerungen spektrografisch auf. Die drei Wörter „votes or something“ sind an ihrer Frequenz erkennbar, die dunklen Stellen weisen auf die Vokale hin. Der Sprecher – das zeigt das Spektrogramm – hat allerdings die Worte nur kryptisch ausgesprochen: „votes 'r som'ing“. Solche Verschleifungen stellen Spracherkennungsprogramme vor riesige Herausforderungen. Empfindlich reagiert der Computer auch auf unregelmäßige Sprechpausen.

würden Sprachcomputer detaillierte Kundenwünsche bearbeiten, Bestellungen aufnehmen, Fragen beantworten, Beschwerden abwickeln – und dabei nie die Geduld verlieren.

Schon jetzt gibt es erste Systeme, die auf automatischer Spracherkennung basieren: Wer etwa in den Niederlanden eine Zugverbindung buchen will, spricht mit einer Maschine, die mit freundlicher Stimme die Bestellung entgegennimmt. Nur wenn die automatische Bestellung auch beim zweiten Versuch nicht geklappt hat, greift ein Mitarbeiter zum Hörer. Offenbar funktioniert der Dialog zwischen Mensch und Maschine nur in sehr engen sprachlichen Bahnen.

DIE KONKURRENZ DER WÖRTER

Noch ist es notwendig, sich mit Grundlagenforschung zu beschäftigen, wie bei der Tagung in Nijmegen deutlich wurde. Denn von der Flexibilität menschlicher Hörer können die Maschinen viel lernen. Ein Beispiel dafür lieferte der Sprachwissenschaftler Gareth Gaskell von der University of York. Gaskell beschrieb, wie Erwachsene neue Wörter in ihr Vokabular einfügen. Folgt man seinen Annahmen, dann funktioniert Spracherkennung beim Menschen fast so wie bei einem digitalen Lexikon, das mit einer Suchmaschine arbeitet. Wenn ein Hörer nur den Anfang eines Wortes wahrnimmt, werden gleichzeitig alle Wörter, die mit diesem Laut beginnen, aktiviert – sie konkurrieren in diesem frühen Stadium miteinander. Die Silbe „spek“ aktiviert etwa die Wörter „Spektakel“, „Spektrum“, „Spekulant“ oder „Spekulatius“. Diese Wörter werden erst deaktiviert, wenn der Input nicht mehr zur gespeicherten Form passt. Folgt nach der Silbe „spek“ ein „u“, so bleiben nur noch „Spekulant“ und „Spekulatius“ übrig. Dies alles läuft im Gehirn des Hörers in Sekundenbruchteilen ab. Der Kreis der mögli-



Es gibt ihn nicht, den perfekten Spracherkennungscomputer – aber mit viel Übung sinkt die Fehlerquote kontinuierlich. Nach dieser von Anne Cutler und Roger Moore erstellten Berechnung müsste das Programm zehn Millionen Stunden lang mit Daten gefüttert werden, um bei der Spracherkennung eine Fehlerquote unter einem Prozent zu erreichen. Die obere Linie zeigt den Versuchsverlauf mit ungefilterten Daten aus der Alltagskommunikation, bei einem weiteren Test war das Trainingsmaterial selektiert, die Fehlerquote daher geringer.

chen Wörter wird immer enger, bis das richtige Wort identifiziert ist. Ähnlich arbeiten im Prinzip auch Sprachcomputer: Sie berechnen bei der Worterkennung die Wahrscheinlichkeit, dass eine bestimmte Lautfolge ein Wort ergibt. Eine Vielzahl lexikalischer Varianten wird dabei ständig als unwahrscheinlich verworfen und ausgeschlossen.

Wie aber werden nun neue Wörter in diesen lexikalischen „Wettbewerb“ eingeführt? Was geschieht, wenn ein neues Wort mit denselben Lauten beginnt wie ein bereits existierendes? Gareth Gaskell fand heraus, dass es zunächst einmal keine Konkurrenz gibt, sofern der Hörer das neue Wort nicht mehrere Tage hintereinander hört. Der vorhandene Wortschatz passt sich nur allmählich an neue Eingaben an – wahrscheinlich, um zuvor gelernte Wörter gegen eine übermäßige Beeinträchtigung zu schützen. James McQueen und Anne Cutler vom Max-Planck-Institut für Psycholinguistik fanden andererseits Beweise dafür, dass Hörer ihre Kriterien für die Identifizierung von Sprachlauten immer wieder flexibel anpassen, wenn die Umstände das erfordern.

Eine Testreihe im Sprachlabor brachte Gewissheit: Die Hälfte der Probanden hörte einen Laut, der

zweideutig zwischen „s“ und „f“ am Ende einer Reihe von Lauten wie „kara-“ vorkam, was dem niederländischen Wort „karaf“ (Karaffe) ähnelt. Sie hörten auch eindeutige „s“-Laute am Ende von Wörtern wie „karkas“ (Kadaver). Die andere Hälfte der Versuchspersonen hörte das Gegenteil: ein eindeutiges „f“ in Wörtern, die mit „f“ enden wie „karaf“ – und den zweideutigen Laut am Ende der Reihe wie „karka-“. Der Kontrast zwischen den zweideutigen Lauten „s“ oder „f“ und die Kenntnis niederländischer Wörter bei den Hörern veranlasste die erste „f“ zu hören, während die zweite Gruppe ein „s“ vernahm.

HÖRER SIND TOLERANT UND FLEXIBEL

Das Ergebnis zeigt: Eine radikale Veränderung in der Aussprache eines Sprachlauts, der ein Wort von einem anderen unterscheidet, stellt den Hörer vor keine Schwierigkeiten, sofern er genug Informationen bekommt, um den Laut in ein bestehendes Wort einzureihen. Hörer – das wurde in Nijmegen in mehreren Vorträgen klar – tolerieren unterschiedliche Aussprachen von existierenden Wörtern und sie passen sich an neue Sprachäußerungen an. Sie haben schon so viele verschiedene Aus-

Abb.: MPI für Psycholinguistik

sprachen gehört, dass sie leicht und schnell eine neue Verbindung zwischen einem Signalmuster und einer gespeicherten Kategorie herstellen können.

Um mehr über das Wesen der Spracherkennung zu erfahren, müssen die Wissenschaftler das System in ihrer Entstehungsphase beleuchten. Schon sehr kleine Kinder lernen, Sprachsignale nach Muster und Kategorien zu ordnen. Linguisten können die Reaktion von Babys testen, indem sie das Saugverhalten der Probanden beobachten. Dazu wird das Kind im Sprachlabor verschiedenen Lauten ausgesetzt. Wenn das Baby etwa den Wechsel in der Abfolge „da da da“ und „pa pa pa“ verstanden hat, saugt es automatisch schneller. Über komplizierte Versuchsanordnungen lassen sich die kindlichen Reaktionen elektronisch messen. Bei einem anderen Versuch

sitzen die Kinder vor Bildschirmen, auf denen Begriffe wie „Blume“ oder „Ball“ auftauchen. Wird das passende Wort von einem Sprecher artikuliert, können die Forscher an der Blickrichtung der Kinder sehen, ob sie die Worte verstehen oder nicht.

VERTRAUHEIT STEUERT LERNEN

Im Alter von sechs bis zwölf Monaten werden Kinder für unterschiedliche Laute empfänglich. Sie können den Klang der Wörter verstehen – vorausgesetzt, der Sprecher ist ihnen vertraut. Schwierigkeiten traten im Versuch immer dann auf, wenn die Wörter von unterschiedlichen Personen gesprochen wurden. Wird ein bereits bekanntes Wort sofort von einem zweiten Sprecher wiederholt, dann können es bereits sieben Monate alte Kinder verstehen. Der Wiederholungseffekt hat funk-

tioniert. Experimente zeigen aber auch: Liegt zwischen den Äußerungen desselben Worts ein Tag, kann das Kind dieses Wort nicht mehr erkennen, weil es den zweiten Sprecher nicht kennt. Es ist auf eine ganz bestimmte Stimmfrequenz gepolt.

Menschen lernen also in der frühen Kindheit, Sprachsignale zu entschlüsseln und diese Signale zu einem sinnvollen Ganzen zusammenzusetzen. Bis das Puzzle gelöst ist, vergehen einige Jahre. Auch Sprachcomputer benötigen eine solche Lernphase. Um eine Maschine dazubringen, erst Laute, dann ganze Wörter und Sätze und zuletzt Texte zu erkennen, müssen die Wissenschaftler Algorithmen entwerfen – das sind komplizierte Zeichenreihen, die nach einer mathematischen Logik funktionieren. Das Sprachsignal wird dabei digitalisiert und in eine Form gebracht, die ein Computer

verarbeiten kann. Aus jedem Laut muss die Maschine ein Referenzmuster erzeugen, sie muss akustische Einheiten bilden. Diese Lautsignale werden in einem weiteren Schritt in Worte umgewandelt. Technisch funktioniert das, vereinfacht gesagt, indem der Computer bereits bestehende Frequenzdiagramme ständig mit den gerade erkannten akustischen Signalen abgleicht.

Übung macht den Meister – das gilt auch für die Sprachmaschine. Je besser ein Erkennungssystem trainiert ist, desto weniger Fehler passieren. Viele der bisher auf dem Markt erschienenen Programme basieren allerdings meist auf einem relativ bescheidenen Wortschatz. Sie können nur ganz bestimmte Arten von Informationen verstehen – etwa einfache Bestellungen. „Wenn Sie reservieren wollen, dann sagen Sie bitte „eins“, wenn Sie einen Mitarbeiter sprechen wollen, sagen Sie „neun“: Nach diesem Muster funktioniert die derzeit gängige Form der Spracherkennungsprogramme, die man als Kommando-Empfänger bezeichnen könnte. Wesentlich größer ist der Wortschatz von automatischen Diktiergeräten, die allerdings einen entscheidenden Nachteil haben: Sie sind fast immer auf einen individuellen Sprecher eingestellt und reagieren äußerst empfindlich, wenn die Stimmfrequenz plötzlich wechselt. Um unbekannte Sprecher zu verstehen, muss die Maschine ein Vielfaches an Trainingseinheiten absolvieren. „Automatische Spracherkennung basiert auf Statistik“, sagt Roel Smits, „je mehr Daten, desto genauer das Verständnis.“

Menschen sind offenbar besser in der Lage, die sprachlichen Nuancen eines Alltagsgesprächs zu erkennen und phonetischen Fallstricken aus dem Weg zu gehen. Gerade die deutsche Sprache ist voller Mehrdeutigkeiten. Homophone – also gleich klingende Wörter – sind häufig, was Fremdsprachler vor große Probleme

stellt. Um die Unterscheidung zwischen dem Adjektiv „viel“ und dem Verb „fiel“ zu verstehen, muss der Computer den Kontext begreifen. Ähnliche Schwierigkeiten machen Doppelpaare wie „Meer/mehr“, „wieder/wider“ oder „Lerche/Lärche“. Selbst Muttersprachler sind da oft irritiert, Spracherkennungsprogramme erst recht.

Es gibt aber auch Bereiche, in denen Maschinen den Menschen schon heute überlegen sind. Bei der bloßen Identifizierung von unbekanntem Sprechern können Computer bestimmte Signale sehr präzise entschlüsseln. Die Gefahr, sich von Stimmenimitationen überrumpeln zu lassen, ist bei ihnen fast ausgeschlossen: Die Frequenzdiagramme sprechen eine klare Sprache. Kein Wunder, dass automatische Spracherkennung und vor allem Sprecherkennung in der Kriminologie eine immer wichtigere Rolle spielen.

THEORIE IST DER PRAXIS VORAUSS

Doch was das Erkennen komplizierter Sprechereinheiten angeht, müssen die Sprachcomputer noch stark verbessert werden. Wie lange es dauern würde, bis eine Maschine eine Irrtumsrate von nahezu null Prozent erreichen würde, haben Anne Cutler vom Max-Planck-Institut und Roger Moore von der englischen Firma 20/20 Speech in England modellhaft ausgerechnet: Zwischen zwei bis neun Millionen Übungsstunden wären theoretisch nötig, um einen Computer so zu schulen, dass er die Kapazität eines nahezu perfekten menschlichen Hörers erreicht. Von Perfektion kann bei der jetzigen Generation der Computerprogramme noch keine Rede sein. Doch die Konzerne – etwa IBM mit seinem System ViaVoice – arbeiten mit Hochdruck daran, Spracherkennung verlässlicher zu machen. Noch nutzen weniger als ein Prozent der Bevölkerung automatische Spracherkennung –

doch das dürfte sich bald ändern. In einigen Jahrzehnten, glaubt der Psycholinguist Roel Smits, werden Mikrochips in Haushaltsgeräten menschliche Befehle erkennen und an intelligente Maschinen weiterleiten. Fernbedienungen werden ebenso auf Zuruf funktionieren wie PCs. Mobiltelefone verfügen bereits über Chips mit Spracherkennungstechnologie. Erhebliche Fortschritte sind außerdem bei automatischen Hörhilfen zu erwarten.

Wie so oft ist die Theorie der Praxis weit voraus, und bei der Spracherkennung gilt das besonders. Hynek Hermansky, der an der Oregon Health and Sciences University in Portland lehrt, stellte in Nijmegen ein Erkennungssystem vor, das auf der Struktur des menschlichen Gehörs basiert. Dieses System verwendet Intervalle von einer Sekunde Sprechdauer, die bis zu 15 Sprachlaute enthalten. Eine solch differenzierte automatische Spracherkennung wäre weit weniger anfällig gegenüber Hintergrundgeräuschen oder Verzerrungen als bisherige Systeme, die Energiemengen in Intervallen von der Länge eines einzigen Sprachlauts aufweisen.

Wann also werden wir künftig mit Sprachcomputern so kommunizieren, wie es Captain Kirk im Raumschiff Enterprise vorgemacht hat? Auf dem Workshop in Nijmegen ging es weniger um konkrete Anwendungen und auch nicht um visionäre Projekte, die den Bereich der künstlichen Intelligenz streifen. Die Wissenschaftler wollten vielmehr den Stand der Grundlagenforschung darstellen. Denn ohne einen Wissenstransfer zwischen menschlicher und automatischer Spracherkennung wird es keine wesentlichen Fortschritte bei den neuen Technologien geben, glaubt Roel Smits. „Wir stoßen an eine Grenze – auch wenn es gelingt, die Datenkapazität der Computer weiter zu steigern.“

CHRISTIAN MAYER



Foto: Corbis - StockMarket

Im Alter von wenigen Monaten lernen Kleinkinder, einfache Sprachmuster zu unterscheiden und nach Wahrscheinlichkeiten zu kategorisieren. Besonders leicht lernen Kinder neue Worte, wenn ihnen die Stimme vertraut ist.