Spracherkennung

Mensch, ich versteh' Dich

Die Spracherkennung per Computer macht sich auf zu neuen Ufern: Die Technik stellt sich auf den jeweiligen Benutzer ein. Die Systeme verstehen Inhalte und werden dialogfähig. *Von Manfred Flohr*

Es grenzt an Zauberei: Der Versuchsperson im Fahrsimulator wurde lediglich gesagt, welche Funktionen des Fahrzeugs sie per Spracheingabe oder mit Gesten steuern soll. Weitere Hinweise gab es beim Einsteigen in die mit Messelektronik gespickte Karosse nicht. Lenken muss der Proband den 7er BMW selbst. Der Trip durch die virtuellen Landschaften auf der Großbildleinwand erfordert sein ganzes fahrerisches Können. Alles, was der Testperson sonst noch abverlangt wird, setzt sie durch Sprache und Zeichen um. „Lauter. Anderer Sender. Bayern Drei. Weiter". Egal, welche Worte er zum Steuern des Radios benutzt – es klappt. Die richtigen Sender werden eingestellt, die Lautstärke nach Wunsch ausgesteuert. Um die Klimaanlage zu regeln, reicht eine lässige Handbewegung. Hat die Kommunikation mit dem Bordcomputer eines Autos über Nacht einen Quantensprung gemacht? Ist die automatische Spracherkennung durch Zauberei perfekt geworden?

In der Tat heißt das Verfahren, das hier an der Technischen Universität München angewandt wird „Wizzard of Oz". Den Zauber macht dabei allerdings ein Versuchsleiter, der in einem Kontrollraum hinter einer Glasscheibe sitzt. Auf mehreren Monitoren verfolgt der Wissenschaftler das Geschehen im Fahrzeug und drum herum. Während der Proband meint, der Computer würde all seine Kommandos ausführen, erfüllt der Mann hinter der Scheibe die Wünsche des Testfahrers – und zwar manuell.

Mit Versuchen wie diesem wollen die Forscher herausfinden, wie ein Mensch den Computer bedienen würde, wenn er gleich von Beginn an von dessen Fähigkeiten überzeugt wäre. Diese Testreihen sollen dazu dienen, intelligente Maschinen zu verbessern und dem Menschen näher zu bringen. Usability-Labs wie die Einrichtung in München, wo verschiedene Formen der Kommunikation zwischen Mensch und Maschine erforscht werden, sind typisch für

SERIE

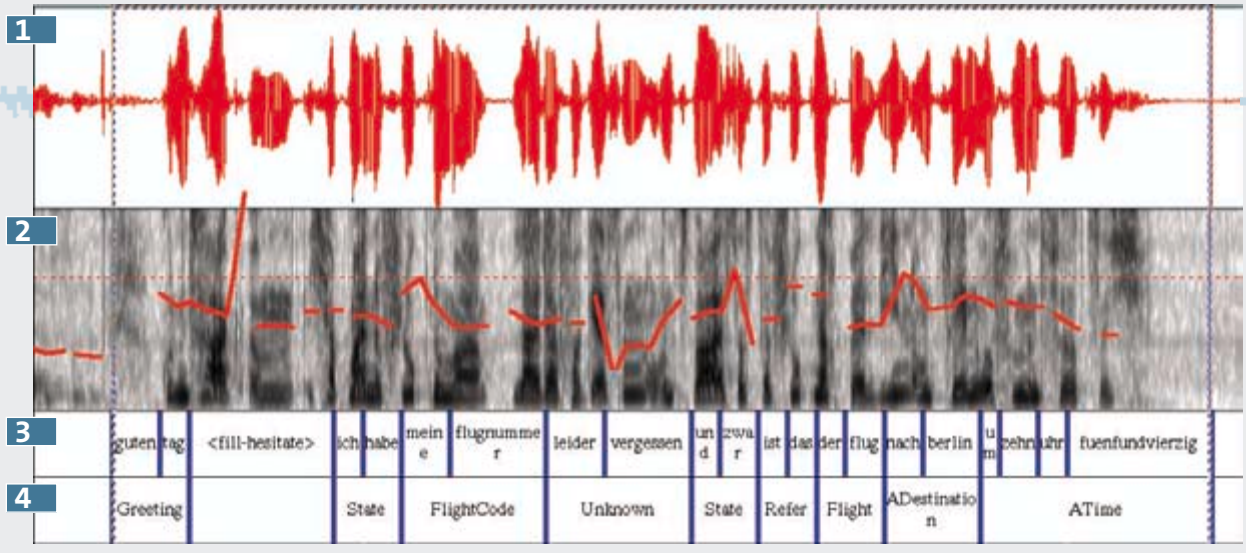
NEUES AUS DEN FORSCHUNGSLABORS

Sie stellen die Weichen für die Zukunft: Wissenschaft und Industrie arbeiten an Schlüsseltechnologien, die unser Leben dramatisch verändern werden. CHIP zeigt in dieser Serie, wie die Welt von morgen aussieht.

Fotos: K. Weichbrodt; Daimler Chrysler; dpa; Ullstein Bild; K. Bogon, Wildlife; EBV; M. Hüttinger

FLUGAUSKUNFT: SO ANALYSIERT DER COMPUTER SPRACHINHALTE

Die Zuordnung von Klängen und Sprache ist eine aufwendige Arbeit. „Verschriften" nennen Wissenschaftler den Vorgang, aus gesprochenen Lauten, Phonemen, Wörtern, Satzteilen oder ganzen Sätzen einen geschriebenen Text zu machen. Mit einem Annotationstool behalten sie den Überblick: Die Software zeigt, ob das Gesprochene richtig erkannt wurde. Im letzten Schritt wird den ermittelten Begriffen eine Bedeutung zugeordnet – Grundlage für einen Dialog zwischen Mensch und Computer.



1 Frequenzdiagramm: Es zeigt die digitalisierten Ausgangsdaten, die es zu erkennen gilt. Hier wird eine Soundkarte eingesetzt, welche die Sprache aus der Mikrofonaufnahme digitalisiert.

2 Spektrum: Hier werden alle Elemente sichtbar gemacht, aus denen sich Sprache zusammensetzt. Alle zehn Millisekunden extrahiert das Programm aus einem kurzen Abschnitt des Signals mehrere Merkmale und fasst sie zu einem Vektor zusammen, mit dem sich Laute identifizieren lassen. Als Beispiel ist rot die Grundfrequenz eingezeichnet, welche die Sprachmelodie charakterisiert.

3 Spracherkennung: Erst werden den aufbereiteten Daten passende Phoneme , also kurze Klangelemente, zugeordnet, dann wird nach den passenden Wörtern gesucht, die sich daraus bilden.

4 Semantische Erkennung: Benutzte Vokabeln, Satzbau und Zusammenhang erlauben es dem System, den Inhalt des gesprochenen Satzes zu analysieren.

neue Entwicklungen in der Spracherkennung. „Wir schränken den Wortschatz drastisch ein und bringen dafür mehr Semantik ins Spiel", erläutert Professor Gerhard Rigoll einen der aktuellen Ansätze, die automatische Spracherkennung per Computer effizienter und alltagstauglich zu machen. Rigoll ist Ordinarius am Lehrstuhl Mensch-Maschine-Kommunikation an der TU München.

Kleinerer Wortschatz – besseres Verständnis

Das Motto heißt Spezialisierung. Während sich andere Spracherkennungsprogramme mit Vokabularen jenseits der 100.000 Wörter herumschlagen und die 100-Prozent-Marke als Trefferquote nie erreichen werden, kommen Spezialprogramme mit einem Wortschatz von weniger als 1.000 Vokabeln aus. Dafür versuchen sie, das Gesagte inhaltlich zu verstehen und die Absicht des Sprechenden zu erkennen. Ziel ist ein richtiger Dialog zwischen Mensch und Maschine – allerdings über ein sehr begrenztes Thema.

Als Musteranwendung nennt Rigoll eine automatische Flugauskunft, bei der ein Computer in Dialog mit dem Benutzer tritt (siehe Grafik oben). Das System erkennt nicht nur, was der Benutzer sagt, sondern sortiert die einzelnen Begriffe auch nach deren Bedeutung ein. Schon bei der Eingabe wird der Inhalt semantisch untersucht. Das funktioniert allerdings nur, wenn die genannten Schlüsselwörter auch tatsächlich etwas mit der Flugauskunft zu tun haben. Hat der Computer etwas nicht verstanden, fragt das System gezielt nach, ehe etwas falsch interpretiert wird.

Gerhard Rigoll macht es sich auf dem schwarzen Ledersofa in seinem Büro bequem und beginnt über die Geschichte der Spracherkennung zu plaudern, die zu einem guten Teil auch seine eigene Geschichte ist. Seit über zwanzig Jahren beschäftigt sich der 45-Jährige mit dem Thema Speech Recognition. „Am Anfang habe ich selber daran gezweifelt, dass wir je mit einem Computer einen Dialog führen können", erzählt Rigoll. →

WIE DER COMPUTER SPRACHE ERKENNT

»Datenbasen, Algorithmen und Statistik

Beim kontinuierlichen Sprechen sind Wörter oft lückenlos aneinander gereiht. Für einen Computer ist es schwer, diese Klangmuster in Wörter zu zerlegen.

Stufe 1: Phoneme erkennen. Ein ganzes Wort besteht aus mehreren Phonemen (lt. Duden: „kleinste bedeutungsentscheidende, aber nicht selbst bedeutungstragende lautsprachliche Einheit"). Bei normaler Sprechgeschwindigkeit hat ein Phonem die Dauer von 10 bis 40 Millisekunden. Bei der Spracherkennung werden in Abständen von etwa 10 Millisekunden Kurzzeitspektren der Akustik erstellt. Daraus errechnet das System einzelne Kennwerte und fasst sie zu einem Merkmalsvektor zusammen. Die zeitliche Folge von Merkmalsvektoren bildet die Grundlage für die Entscheidung, welche Wortfolge gesprochen wurde. Dazu werden die Merkmalsvektoren mit gespeicherten Referenzmustern verglichen.

Usability-Lab: Training verhilft zum besseren Dialog zwischen Mensch und Maschine.

Stufe 2: Hidden-Markov-Modelle. Um diese Vergleiche möglichst schnell bei optimaler Erkennung vorzunehmen, wird ein statistisches Verfahren benutzt, das auf so genannten Markov-Ketten basiert. Das sind Ketten von Übergangswahrscheinlichkeiten von einem Phonem zum nächsten. Nach einer Trainingsphase wird bei der Erkennung für einen unbekannten Musterverlauf die Wahrscheinlichkeit dafür berechnet, dass das Modell diesen Verlauf erzeugen kann. Diese Berechnung wird wiederholt ausgeführt. Das erfordert hohen Rechenaufwand.

Stufe 3: Bi- und Trigramme. Damit ein Spracherkennungsprogramm eine noch höhere Erkennungsgenauigkeit erreichen kann, gibt es neben dem Hidden-Markov-Modell ein weiteres statistisches Verfahren. Durch die Bi- beziehungsweise Trigrammstatistik, die während des Diktierens permanent ihre Berechnungen ausführt, wird eine Kontextprüfung vollzogen. Das System passt sich dadurch immer mehr an den Sprecher und seinen individuellen Sprachstil an.

1986 ging er zur IBM-Forschung in die USA. Damals begannen die Boom-Jahre der Spracherkennung und erste kommerzielle Produkte kamen auf den Markt. Das Konzept der so genannten Hidden-Markov-Modelle (siehe oben) hatte den Entwicklern ein Werkzeug an die Hand gegeben, mit dem Sprache anhand statistischer Merkmale erkennbar wird. Ermittelt wird dabei die Wahrscheinlichkeit, mit der bestimmte Merkmale, etwa Frequenzen, produziert werden. Anhand der Merkmale wird auf Wörter geschlossen.

Bevor sich die Spracherkennung per Statistik anbahnte, hatte man versucht, die digitalisierten Signale in Laute zu unterteilen. „Sprache ist aber ein Prozess, der tief in unseren Sinnesorganen sitzt. Es ist daher nicht einfach, sie in Wenn-dann-Regeln zu packen", doziert Rigoll. Die Hidden-Markov-Modelle brachten für die Spracherkennung den Durchbruch. IBM wurde mit dem darauf basierenden Diktiersystem ViaVoice Marktführer, auch andere Anbieter bedienten sich bald dieser Grundlage.

Kommunikations-Probleme in der Praxis

Doch die Euphorie ist mittlerweile vorbei. In den meisten Büros wird auf Spracherkennung verzichtet, weil für die Praxis die Fehlerquote immer noch zu hoch ist. Das weiß auch Rigoll, in dessen Vorzimmer sogar noch eine rote mechanische Schreibmaschine steht. Versuche im eigenen Büro hat er rasch wieder bleiben lassen. Für den Wissenschaftler Rigoll geht es jetzt um ein „gnadenloses Verfeinern und Optimieren" der Algorithmen.

So viel Wert darauf gelegt wird, den Menschen bei der Kommunikation mit der Maschine zum Maß der Dinge zu machen, so wenig ist er nötig, wenn es darum geht, noch mehr Erkennungsleistung aus den Systemen herauszuholen.

> »Inzwischen geht es um ein gnadenloses Verfeinern und Optimieren der Algorithmen.
>
> Professor Gerhard Rigoll, Technische Universität München

„Bei uns redet heute kein Mensch mehr mit Computern", verrät Rigoll. Vorbei sind die Zeiten, da Forscher dem Rechner einen Text vorlasen und anschließend prüften, wie viel die Software fehlerfrei erkennen konnte. Inzwischen sorgen normierte Datenbanken für reproduzierbare Testbedingungen, die auch kleinste Fortschritte sichtbar machen. Zwar ist Englisch internationaler Standard, doch bei der Verfeinerung der Algorithmen spielt die Sprache keine Rolle.

Das Ganze ist eine recht mühselige Prozedur, die vorwiegend aus Programmier-Arbeit besteht. Erst nach stundenlangen Computerläufen, in extremen Fällen sogar erst nach Wochen, wissen die Forscher, ob Änderungen an den Berechnungen etwas gebracht haben oder nicht.

Fahrsimulator: **Unter der Haube des Autos in einem Usability-Lab der TU München arbeitet kein Motor, sondern ein älterer Pentium-PC, der die Simulation steuert (links oben). Im Kontrollraum (links) werden die Sprachbefehle des Testfahrers noch manuell umgesetzt – Vorarbeit für einen automatischen Dialog zwischen Fahrer und Fahrzeug.**

„Kern aller statistischer Methoden ist die Frage: Welches Wort wurde mit welcher Wahrscheinlichkeit gesprochen, wenn bestimmte Merkmale beobachtet werden?“, bringt Professor Günther Ruske das Prinzip auf den Punkt. Ruske, der ebenfalls an der TU München forscht, kommt mit den notwendigen Abstraktionen gut zurecht.

Kurven-Diskussion führt ans Ziel

Die Mustererkennung funktioniere bei der Sprache wie bei der Erkennung von Handschriften, Buchstaben oder Gesichtern. Aus vielen Ansichten wird herausgefiltert, welche Merkmale wichtig sind. „Der Mensch muss schließlich auch nicht alle Häuser der Welt kennen, um ein Haus als solches zu erkennen“, veranschaulicht er die Bedeutung effizienter Algorithmen. Es habe wenig gebracht, nur auf den Menschen zu schauen, zur automatischen Sprachverarbeitung seien die Spitzen in den aufgezeichneten Kurven viel aussagekräftiger. Ruske beugt sich den Fakten: „Ich persönlich halte das für einen sehr primitiven Ansatz, aber wir verwenden ihn, weil die Ergebnisse ganz gut sind.“

Anhand der spektralen Maxima können die Wissenschaftler Vokale recht gut erkennen. Das A liegt beispielsweise immer weit vorne im Spektrum und kann damit für eine gewisse Eichung dienen. Wo genau diese Resonanzstelle bei einem bestimmten Menschen sitzt, entscheidet das passive „Rohr“, wie Ruske den Vokaltrakt nennt. Die Stelle hängt von der Länge des Vokaltrakts ab und wird durch Mund und Zunge verschoben. Per logarithmischer Verschiebung lassen sich diese Vokallängen im System normieren und so auf einen Sprecher anpassen. „Ein System, das sich binnen Sekunden an einen Sprecher adaptiert, wäre ein Traum“, zeigt Ruske die aktuellen Grenzen auf.

Wie schwierig die Erkennung ist, wenn es keine Einschränkungen des Vokabulars gibt, und das System nicht auf einen Sprecher angepasst wird, zeigen Highend-Anwendungen der Spracherkennung, die bereits eingesetzt werden. Für die Maschine verschwinden da zum Beispiel die Unterschiede zwischen einem amerikanischen Flugzeugträger und einem deutschen Konferenzsaal. In beiden Fällen zeichnet ein System über Stunden die gesamte Kommunikation auf.

» Wir sind noch weit entfernt davon, das wirklich Wesentliche zu erkennen.

Professor Günther Ruske, Technische Universität München

Im Prinzip ein großes Diktiergerät – an der Entschlüsselung der Inhalte hapert es. Immerhin gibt die automatische Erfassung auch Leuten Informationen, die nicht selbst am Ort des Geschehens waren. „Meeting Transcriptions“ sind der neueste Schrei für Manager, die informiert sein wollen, ohne an einer Besprechung teilgenommen zu haben. →

VIELREDNER: Egal, ob Konferenzsaal oder Flugzeugträger – die Software zeichnet alles auf. Mangels Spezialisierung hat sie aber Probleme mit den Inhalten.

Ein ganz heißes Eisen ist die Spracherkennung für mobile Geräte. UMTS-Handys, mit denen künftig auch im Internet gesurft werden soll, werden allein mit der kleinen Tastatur auf Dauer kaum auskommen. „Das verlangt nach einem neuen User-Interface, und ich bin überzeugt, dass im mobilen Bereich die Spracherkennung einziehen wird", gibt sich Gerhard Rigoll zuversichtlich.

Die Ansprüche sind hier allerdings sehr hoch, denn es muss nicht nur mit unterschiedlichen Benutzern gerechnet werden, sondern auch mit schlechter Tonqualität und störenden Nebengeräuschen. Auch wenn die Spracherkennung weiter verbessert und robuster gemacht werden kann, wird ein Handy schwerlich in der Lage sein, die dafür erforderliche Rechenleistung zu erbringen.

Ein Ausweg könnte die verteilte Erkennung sein. Das Handy berechnet aus der aufgenommenen Sprache nur die wichtigsten Merkmale und schickt die entsprechenden Daten zu einem Server beim Provider. Dort würde ein Großrechner die eigentlichen Berechnungen zur Spracherkennung ausführen. Siemens arbeitet in München bereits an einem Chip, der die entsprechenden Leistungen auf der Handy-Seite erbringen soll. Ob und wann er eingesetzt wird, steht in den Sternen – denn noch braucht kein Handy die Hilfe eines Großrechners.

manfred.flohr@chip.de

LINKS

www.mmk.ei.tum.de/forschung/
http://verbmobil.dfki.de/Vm.Info.Phase2.html

WORT FÜR WORT

»Geschichte der Spracherkennung

1952 Die Bell Laboratories führen ein System ein, das die übers Telefon gesprochenen Ziffern 0 bis 9 erkennt.

1959 Ein System des MIT erkennt 93 Prozent der Vokale. Sieben Jahre später werden 50 Wörter erkannt.

1962 Das erste Gerät zur Sprachausgabe kommt auf den Markt. IBM 7772 redet ziemlich blechern.

1968 Science Fiction ist um Jahrzehnte voraus. Im Film „2001 – Odyssee im Weltraum" spricht der Computer HAL mit den Astronauten.

1976 Bruce Lowerre entwickelt das System Harpy, das komplette Sätze und einfache grammatikalische Strukturen erkennt. Dazu war Parallelverarbeitung durch 50 Computer nötig.

1977 Die Bausparkasse Wüstenrot nimmt als erster kommerzieller Anwender in Deutschland ein System zur „Sprachausgabe" in Betrieb.

1978 Texas Instruments packt einen Sprachprozessor auf einen Chip.

1986 IBM Tangora 4 erkennt in Echtzeit statistische Strukturen.

1988 Dragon bringt die erste Spracherkennungssoftware für den PC.

1996 OS/2 Warp ist das erste Betriebssystem mit Sprachsteuerung.

1997 Immer mehr Programme sind marktreif. Im Juli kommt Dragon NaturalySpeeking, das 23.000 Wörter erkennt. IBM bringt im August Via Voice auf den Markt. Philips sowie Lernout & Hauspie ziehen nach.

2000 Sprache-zu-Sprache-Umwandler „Verbmobil" von Wolfgang Wahlster.