Ulrike Kuhlmann

Achtung Aufnahme

Der sprachgesteuerte Videorecorder

Ivica Rogina hat sich des 'Haushaltsgerätes' angenommen, das schon viele Anwender zur Verzweiflung getrieben hat. Der Diplom-Informatiker entwickelte einen Videorecorder, der auf Zuruf Videos vorspielt, Sendungen aufnimmt und über das Fernsehprogramm informiert.

Auf die Anweisung 'Nimm heute abend den Actionfilm mit Bruce Willis auf', reagiert das Gerät prompt: 'Ich zeichne den Film 'Stirb langsam' um 22.15 Uhr in der ARD auf', schnarrt es aus den Lautsprechern. Zugleich erscheint auf dem Fernsehschirm ein Vorschaubild des gesuchten Films inklusive Inhaltstext im Programmzeitschriftenstil.

Das muss auch Opa können – diesem Credo hat sich Ivica Rogina verschrieben. Der Videorecorder des Wissenschaftlers am Institut für Logik, Komplexität und Deduktionssysteme der Universität Karlsruhe wird diesem Anspruch gerecht: Er lässt sich mit einer eigens entwickelten Hard- und Software ohne Fernbedienung per Spracheingabe programmieren. Sollte das Gerät die Anweisung nicht verstanden haben, bittet die synthetische Stimme um eine erneute Ansage.

Ohne Training

'Die audiovisuelle Bestätigung der Eingabe ist enorm wichtig', erläutert Entwickler Rogina. Nur mit dem entsprechenden Feedback lasse sich sicherstellen, dass die Programmierung im Sinne des Anwenders ausgeführt wird. Liefen beispielsweise zwei Filme mit Actionheld Willis, muss man auswählen können, welcher Film aufgezeichnet werden soll. 'Die Erkennungsleistung liegt bei circa 90 Prozent; da wird auch einmal etwas falsch oder gar nicht erfasst. Ohne Rückmeldung würde die Mensch-Maschine-Schnittstelle in einem solchen Fall versagen.'

Eine Genauigkeit von rund 90 Prozent erreichen auch die herkömmlichen Diktiersysteme von IBM, Dragon und anderen. Doch anders als diese Programme gehorcht der sprechende Videorecorder dem Anwender ohne jegliches Training – der Erkenner ist sprecherunabhängig. Solche sprecherunabhängigen Systeme stellen sich auf eine beliebige Aussprache, Betonung, Wortwahl und Satzstellung ein und verstehen auch Akzente. 'Bei einem starken Dialekt tut sich unser System allerdings schwer', schränkt der Diplom-Informatiker ein. Dabei erfasst der Recorder nicht nur beliebige Wörter, sondern deutet auch ihren Sinn. Denn die Frage nach einem Actionfilm mit Willis beinhaltet auch, dass nur Filme der entsprechenden Kategorie herausgefiltert werden.

Suche ist Wissenschaft

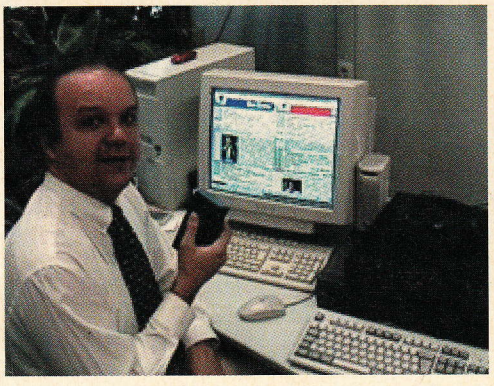
Um die Erkennungsgenauigkeit der Software zu erhöhen, will der Wissenschaftler weitere Anwender um eine Spracheingabe bitten. Aus dem Sprachprofil von einigen hundert Studenten konnte Ivica Rogina bereits ein ausgefeiltes akustisches Modell erstellen. Auf Basis des Hidden-Markov-Modells zerlegte er Wörter des allgemeinen Sprachgebrauchs in Lautfolgen. Das so generierte Aussprachelexikon beschreibt, aus welchen Phonemen jedes Wort zusammengesetzt ist. Da die Phoneme je nach räumlichem Kontext anders betont werden (das 'a' in 'Laus' wird anders gesprochen als das 'a' in 'Land'), zergliedert man sie in so genannte Subpolyphone. Diese Untergruppen fasst man wiederum mit einem Clustering-Verfahren zusammen; die Klassen bilden die atomaren Einheiten des Erkenners. Daneben hält ein Sprachmodell Aussagen über die Wahrscheinlichkeit von Wortfolgen bereit (siehe auch [1]).

Nach der Spracheingabe zerlegt das System alle Wörter in Lautfolgen und vergleicht sie mit den atomaren Einheiten: Je genauer ein akustisches Atom mit einem Laut übereinstimmt, umso höher ist die ihm zugewiesene Wahrscheinlichkeit. Ein ausgefeilter Algorithmus ermittelt die in Frage kommenden atomaren Einheiten. Die Summe aller Wahrscheinlichkeiten entscheidet am Ende über das identifizierte Wort. 'Die Suche nach den passenden Atomen muss man durch Entscheidungskriterien einschränken; hierin liegt die Wissenschaft der Spracherkennung.'

Während die Modelle (akustisches und sprachliches) beim sprecherabhängigen System durch Vorlesen beziehungsweise Einlesen von Texten an den Anwender angepasst werden, sind sie beim sprecherunabhängigen System fest und dementsprechend aufwändig.

Vorschriftsmäßig

'Ein großes Problem bei der Spracherkennung ist immer der Benutzer. Nur ein kooperativer Benutzer erzielt Genauigkeiten von über 90 Prozent. Es gibt etliche Anwender, die durch falsches Eingabeverhalten Misserfolge erzielen; die wenden sich dann ab und sagen, 'das funktioniert doch gar nicht'.' Die Hoffnung, das Gerät über größere Kanäle vertreiben zu können, habe sich genau an diesem Punkt zerschlagen, erklärt Rogina. Ohne eine größere Robustheit gegenüber Anwendungsfehlern und eine Genauigkeit von 99 Prozent würden große Unternehmen wie Sony oder Philips den sprechenden Videorecorder nicht ins Programm aufnehmen. Außerdem sei die notwendige Hardware für den breiten Einsatz derzeit zu teuer: Die Rechnerplattform, ein Pentium II mit 400 MHz und 128 MByte Arbeitsspeicher, kostet im Einkauf etwa fünfmal so viel wie ein herkömmlicher Videorecorder. Einige mittelständische Unternehmen aus der Unterhaltungsbranche hätten dennoch Interesse signalisiert. Weitere Anfragen hat Ivica Rogina von Blindenverbänden bekommen. Da die Spezialausstattungen für Menschen mit Behinderungen ohnehin deutlich teurer als vergleichbare Massenprodukte sind, sieht man hier die Kosten für den sprechenden Videorecorder weniger kritisch. (uk)



Der sprachgesteuerte Videorecorder beruht auf einem sprecherunabhängigen Erkenner und läuft auf einem Standardrechner unter Linux.

Daten aus dem Internet

Um auf die Anfrage 'Kommt heute abend ein Sciencefiction?' angemessen reagieren zu können, muss der sprachgesteuerte Videorecorder genaue Daten über das Fernsehprogramm auswerten. Während diese Datenbasis beim vorgestellten Prototyp noch per Hand eingegeben wurde, könnte man das komplette TV-Programm nebst Fotomaterial, Kurzzusammenfassung, beteiligten Schauspielern und einer Klassifizierung aller Sendungen über das Internet beziehen. Einige dieser Daten werden bereits jetzt mit den Fernsehsignalen übermittelt, mit einem entsprechenden Decoder ließen sie sich also verwerten. Und spätestens mit steigender Verbreitung der Settop-Boxen dürfte der Zugriff auf das notwendige Datenmaterial kein Problem mehr darstellen.

Literatur

[1] N. Haberland et all, Sprachunterricht, Wie funktioniert computerbasierende Spracherkennung? c't 5/98, S. 120 c't