Spracherkennung

Kommunikation mit Maschinen

Die Trefferquote bei der Spracherkennung wird immer besser. Dafür sorgen neue Software-Algorithmen und Sprachdatenbanken. Dies fördert den Absatz und die Akzeptanz solcher Produkte.

Von Dipl.-Ing. Micha Baum, Dr. Gregor Erbach, Dr. Markus Kommenda

Bei der Spracherkennung gibt es zwei große Klassen von Systemen, die heute im praktischen Einsatz sind: einerseits Diktiersysteme, zum Beispiel IBM ViaVoice, Philips FreeSpeech oder Naturally Speaking von Lernout & Hauspie sowie andererseits Sprachdialogsysteme, die vor allem in Telefonapplikationen Verwendung finden.

Das Vokabular bestimmt die Leistungsfähigkeit

Die Leistungsfähigkeit eines Spracherkennungssystems wird nach dem Umfang des Vokabulars beurteilt, das bei einer vorgegebenen Fehlerrate erkannt werden kann. Der Umfang des Vokabulars, das zuverlässig erfasst wird, ist stark abhängig von der Umgebung, in der das System eingesetzt ist. Die wichtigsten Einflussfaktoren sind:

- *Variation der Sprecher*: Die besten Erkennungsraten erreichen Systeme, die auf die Stimme eines einzelnen Sprechers trainiert werden. Aber auch dann treten Variationen auf, da ein Wort nie zweimal genau gleich ausgesprochen wird. Einflüsse wie Krankheit, Stimmungsschwankungen oder Müdigkeit führen bereits zu Abweichungen. Bei sprecherunabhängigen Systemen treten zusätzlich Variationen auf Grund von Geschlecht, Alter, Dialekt, Akzenten und persönlichen Eigenheiten auf.
- *Kontinuierliche Sprache*: Zum Erkennen kontinuierlich gesprochener Sprache muss das System Beginn und Ende der einzelnen Wörter ermitteln. Zusätzlich ist die Variabilität in kontinuierlicher Sprache deutlich größer, weil die Aussprache der einzelnen Wörter in diesem Fall wesentlich stärker von ihrer Umgebung beeinflusst wird.
- *Spontansprache*: Umgangssprache, wie sie unter Freunden oder in der Familie verwendet wird, ist schwerer zu erkennen als beispielsweise die Sprache von Nachrichtensprechern.
- *Gestörte Übertragung*: Störungen des Übertragungskanals, beispielsweise bei Telefon, Funk, oder Internet, erschweren ebenfalls die Spracherkennung.
- *Umgebungsgeräusche*: Der Einfluss von Umgebungsgeräuschen erschwert ganz erheblich die Spracherkennung. Die Qualität des verwendeten Mikrofons und der Abstand zum Sprecher können deshalb den Ausschlag für einen erfolgreichen Einsatz von Systemen zur Spracherkennung geben. Daher werden für Diktiersysteme, die ein großes Vokabular behandeln müssen, oft Headsets mit qualitativ hochwertigen Mikrofonen eingesetzt.

Die Schwierigkeit der Aufgabe wird von allen genannten Faktoren beeinflusst. Die erreichbare Erkennungsleistung hängt aber auch von der verfügbaren Prozessorleistung und dem Speicherplatz ab. Mit einem leistungsfähigen PC oder sogar Großrechner kann bei gleicher Schwierigkeit ein größeres Vokabular behandelt werden als mit einem Signalverarbeitungsbaustein in einem Handy.

Die Analyse des Sprachsignals

Die erste Aufgabe des Spracherkenners ist die akustische Analyse des Sprachsignals. Schall - also auch Sprache - ist die Ausbreitung von Schwingungen in der Luft. Treffen solche Schwingungen auf die Membran eines Mikrofons, werden sie in elektrische Spannung umgewandelt und für die weitere Verarbeitung digitalisiert.

Durch Abzählen der *Nulldurchgänge* wird der Verlauf der Sprachgrundfrequenz ermittelt, was der Sprachmelodie entspricht. Die zeitliche Änderung der *Amplitude* gibt Auskunft über die Veränderung der Lautstärke oder Dynamik. Am wichtigsten für die Spracherkennung sind jedoch die zeitlichen *Frequenzänderungen*, die im Spektrum des Sprachsignals erkennbar sind.

Alle stimmhaften Sprachlaute – etwa die Vokale oder gewisse Konsonanten wie m oder n – können als eine Überlagerung von Sinusschwingungen betrachtet werden (Fourier-Analyse). Im Spektrum des Sprachsignals ist außerdem die Frequenz, mit der die Stimmbänder schwingen, deutlich erkennbar (Sprachgrundfrequenz) und darüber hinaus die Vielfachen dieser Frequenz (Obertöne). Bei den Vokalen treten dann jeweils bestimmte Bereiche von Obertönen besonders hervor, die für den Klang und die Unterscheidung der Vokale untereinander charakteristisch sind. Diese hervortretenden Frequenzbereiche – genauer: die relativen Maxima in der Einhüllenden des Spektrums – haben einen eigenen Namen: Sie werden Formanten genannt. Auch stimmhafte Konsonanten weisen jeweils charakteristische Formanten auf. Zusätzlich enthalten die meisten Konsonanten auch rauschförmige Anteile; augenfällig ist dies etwa bei den so genannten Reibelauten oder Frikativen (s, f, ch). Verschlusslaute (p, t, k) erkennt man an einer stummen Schallphase, der ein plötzlicher Anstieg der Amplitude folgt.

In einem so genannten Sonagramm wird die spektrale Energieverteilung sowohl über der Zeit (in horizontaler Richtung) als auch über der Frequenz (in vertikaler Richtung) übersichtlich dargestellt. Bereiche hoher Energiedichte werden durch Schwärzung wiedergegeben.

Training von Spracherkennern

Die beschriebenen Verfahren der Spracherkennung setzen immer Referenzmuster für die Erkennungseinheiten (Einzelwörter, Phoneme) voraus, die vorab in einer Trainingsphase gewonnen werden. Um beispielsweise HMMs zur Spracherkennung zu verwenden, müssen zunächst die Übergangs- und Emissionswahrscheinlichkeiten bestimmt werden. Dies geschieht anhand von Sprachproben, für die die dazugehörige Wortfolge - und damit auch die Folge von Phonemen - bekannt ist.
Die Sprachproben werden mit den oben beschriebenen Techniken der Signalanalyse in Folgen von Merkmalsvektoren umgewandelt.
Dann muss festgestellt werden, welcher Teil des Sprachsignals welchem Phonem in der Verschriftung entspricht. Dieser Vorgang wird als *Alignment* bezeichnet. Nach diesem Schritt stehen für jedes Phonem beziehungsweise jede Folge von Phonemen eine Reihe von akustischen Mustern zur Verfügung.
Für jedes Phonem wird daraus ein HMM gebildet, das in der Regel aus drei Zuständen besteht - je einer für Anfang, Mitte und Ende des Phonems.
Im letzten Schritt werden die Übergangs- und Emissionswahrscheinlichkeiten bestimmt, die den Trainingsdaten am besten entsprechen. Dies geschieht in einem iterativen Verfahren, bei dem ausgehend von mehr oder weniger willkürlich gewählten Anfangswerten die Parameter solange angepasst werden, bis keine weitere Verbesserung mehr möglich ist.

Für die Berechnung der Wahrscheinlichkeiten von Wortfolgen werden große Textmengen verwendet, die der erwarteten Eingabe des Spracherkennungssystems entsprechen sollten. Die Bandbreite der verwendeten Daten reicht von sehr anwendungsspezifischen Daten, zum Beispiel transkribierte Benutzereingaben aus einem Dialogsystem, bis hin zu Daten, in denen die häufigsten Wortfolgen der allgemeinen Sprache enthalten sind, zum Beispiel große Mengen von Zeitungstexten.

➜ Bei *sprecherabhängigen* Spracherkennern spricht der Benutzer die jeweils gewünschten - und damit vorab identifizierten - Erkennungseinheiten in der Regel mehrmals vor, um die entsprechenden Referenzmuster zu generieren.

➜ Bei *sprecherunabhängigen* Spracherkennern müssen die Daten von sehr vielen Sprechern unterschiedlichen Geschlechts, regionalen Akzents und Alters gesammelt und verschriftet werden. Mit solchen Daten trainierte Systeme funktionieren für Sprecher einer bestimmten Bevölkerungsgruppe, zum Beispiel einer Nationalität.

➜ Bei *sprecheradaptiven* Spracherkennern geht man zunächst von Referenzmustern aus, welche ähnlich wie bei sprecherunabhängigen Systemen mit Hilfe umfangreicher Sprachdatenbanken gewonnen werden. Der Benutzer hat dann die Möglichkeit, diese Referenzmuster individuell zu optimieren, indem er vorgegebene und vorab verschriftete Texte vorliest.

Die Extraktion von Sprachsignalen

Zur weiteren Verarbeitung im Hinblick auf die automatische Spracherkennung wird das Sprachsignal in gleichlange Abschnitte von zehn bis 20 Millisekunden unterteilt, für die jeweils Sätze von Parametern bestimmt werden. Die in der Spracherkennung am häufigsten verwendeten Parameter sind so genannte Mel-Frequency Cepstral Coefficients (MFCCs), die Eigenschaften der Hörwahrnehmung des Menschen miteinbeziehen. Pro Signalabschnitt werden zirka zehn bis

Vergleich von Diktier- und Dialogsystemen

	Diktiersystem	Dialogsystem
Sprecherabhängigkeit	sprecherabhängig beziehungsweise sprecheradaptiv (muss für jeden Sprecher neu trainiert werden)	sprecherunabhängig
Größe des Vokabulars	bis 100.000 Wörter, die immer aktiv sind	einige tausend Wörter, von denen nur eine Teilmenge aktiv ist
Art der Eingabe	unbeschränkt, auch bei jedem Dialogschritt komplexe Sätze sind möglich	nur bestimmte Muster werden erkannt (abhängig von der Dialoggrammatik)

Bewertungskriterien für Spracherkennungssystemen

Maßeinheit	Bedeutung	Anwendung
Wortfehlerrate (word error rate)	Anteil gesprochener Wörter, die vom Benutzer als falsch erkannt werden	Beurteilung der Leistungsfähigkeit des Spracherkenners
Erfolgsrate (task success rate)	Anteil der Aufgaben, die der Benutzer erfolgreich zu Ende führt	Beurteilung der Benutzbarkeit von Dialogsystemen
Produktivität	für die Durchführung einer Transaktion benötigte Zeit	Beurteilung der Effizienz eines Dialogsystems für den Benutzer
Benutzer-zufriedenheit	subjektives Maß für die Zufriedenheit mit einem Sprachdialogsystem	Beurteilung des Gesamteindruckes eines Sprachdialogsystems

20 solcher Koeffizienten berechnet und dann in einem Vektor gespeichert. Im nächsten Schritt werden diese Vektoren mit Referenzvektoren aus einem Codebuch verglichen und dem ähnlichsten zugeordnet.

Umwandlung der Sprachsignale

Der zentrale Verarbeitungsschritt in der Spracherkennung ist dann die Umwandlung einer Folge von Merkmalsvektoren – der akustischen Repräsentation – in eine Symbolfolge, die ein einzelnes Wort (Einzelworterkennung) oder eine Folge von Wörtern darstellt.

Bei der Einzelworterkennung wird eine akustische Repräsentation mit einer Anzahl von gespeicherten Referenzmustern verglichen und das am besten passende ausgewählt.

Da sich die akustischen Repräsentationen durch unterschiedliche Sprechgeschwindigkeit in ihrem zeitlichen Verlauf stark unterscheiden, wird beim Vergleich eine entsprechende zeitliche Anpassung vorgenommen (dynamic time warping).

Für die Erkennung kontinuierlich gesprochener Sprache mit größerem Vokabular ist dieser direkte Mustervergleich nicht praktikabel, da es nicht möglich ist, für alle gewünschten Eingabewörter Referenzmuster zu speichern. Hier wird von kleineren Einheiten, typischerweise Phonemen, ausgegangen. Die Phoneme selbst werden durch so genannte Hidden-Markov-Modelle (HMM) repräsentiert.

Ein HMM ist ein Netzwerk, das aus Zuständen und Übergängen zwischen diesen Zuständen besteht. Aus jedem Zustand kann mit einer gewissen Wahrscheinlichkeit in andere Zustände gewechselt werden. Außerdem werden bei jedem Zustand mit einer bestimmten Wahrscheinlichkeit (Emissionswahrscheinlichkeit) beobachtbare Ereignisse ausgegeben. Im Falle der Spracherkennung sind diese Ereignisse die Merkmalsvektoren. Da nur die ausgegebenen Ereignisse für einen Beobachter sichtbar sind, nicht aber die durchlaufenen Zustände, spricht man von versteckten (hidden) Markov-Modellen.

Da bei HMMs die Möglichkeit besteht, einen Zustand mehrmals hintereinander zu durchlaufen, sind Variationen der Sprechgeschwindigkeit gut zu behandeln.

Die HMMs für einzelne Phoneme können dann zu größeren HMMs für ganze Wörter (Wortmodelle) zusammengesetzt werden. Als Grundlage dafür dient ein phonetisches Wörterbuch, in dem für jedes Wort die entsprechende Folge von Phonemen enthalten ist.

Das Problem der Spracherkennung lässt sich dann so formulieren: Finde jene Folge von Wörtern, die – für eine gegebene Folge von Merkmalsvektoren – die wahrscheinlichste ist.

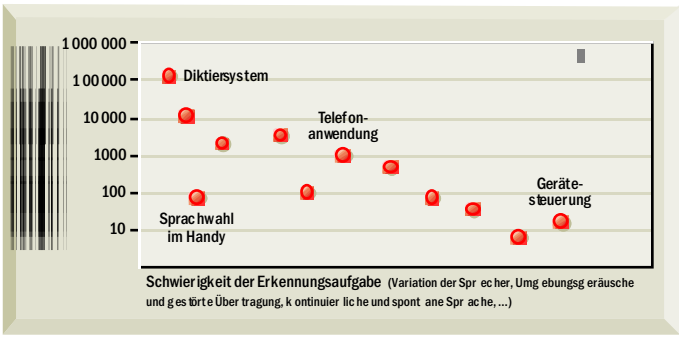
Um diese Folge von Wörtern zu ermitteln, wird unter Zugrundelegen von HMMs nach einer Folge von Zuständen gesucht:

- ➜ Bei der die ausgegebenen Merkmalsvektoren mit jenen übereinstimmen, die aus der Signalanalyse entstanden sind
- ➜ die einem Wortmodell beziehungsweise einer Folge von Wortmodellen entsprechen und
- ➜ bei der das Produkt aus Übergangs- und Emissionswahrscheinlichkeiten maximal ist.

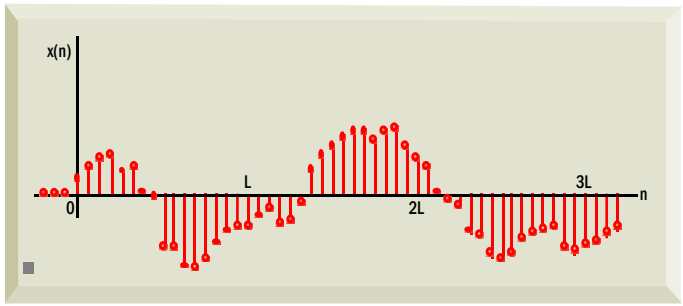
Dabei werden in der Regel nur die Zustandsfolgen mit den höchsten Wahrscheinlichkeitswerten bis zum Ende verfolgt.

Um der Tatsache gerecht zu werden, dass die akustische Realisierung eines Merkmalsvektors variieren kann, wird auch mit Wahrscheinlichkeitsdichtefunktionen an Stelle von diskreten Emissionswahrscheinlichkeiten gearbeitet. Da aufeinander folgende Phoneme sich gegenseitig in ihrer akustischen Realisierung beeinflussen, werden oft HMMs für Triphone - drei aufeinander folgende Phoneme - gebildet.

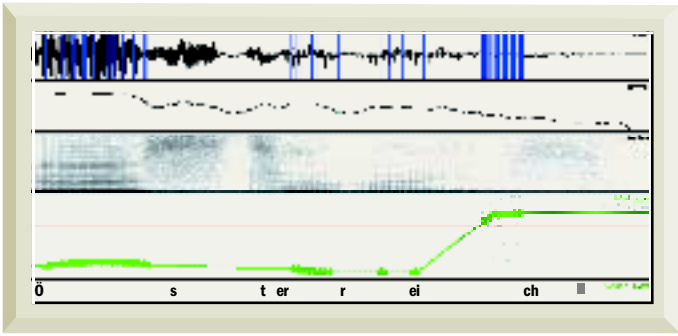
Bei der Analyse zusammenhängender Sprache wird auch die Wahrscheinlichkeit bestimmter Folgen von Wörtern berücksichtigt, um unter den akustisch möglichen Wortfolgen die wahrscheinlichsten auszusuchen. Hierfür werden in der Regel eben-

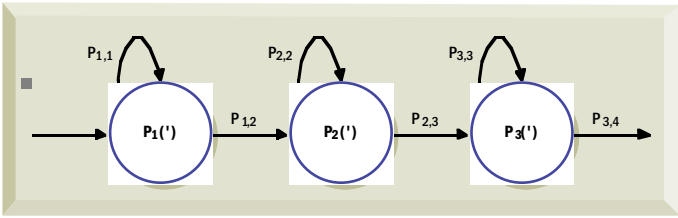


Zusammenhang zwischen Schwierigkeit der Erkennungsaufgabe und Vokabulargröße



Ein Sprachsignal nach der Digitalisierung





(oben)
Analyse eines Sprachsignals. Es ist das Wort Österreich dargestellt

(unten)
Hidden Markov Modell: die Zustände (mit Emissionswahrscheinlichkeiten) sind durch Kreise dargestellt, Zustandsübergänge (mit Übergangswahrscheinlichkeiten) durch Pfeile

falls statistische Modelle verwendet, welche die Wahrscheinlichkeit von Einzelwörtern (Unigramm), Folgen von zwei Wörtern (Bigramm) oder drei Wörtern (Trigramm) auswerten. Da nicht für alle möglichen Trigramme die Wahrscheinlichkeiten bekannt sind, werden auch die Wahrscheinlichkeiten der enthaltenen Bigramme und Unigramme mit berücksichtigt.

Diese Art von Wahrscheinlichkeitsberechnung für Wortfolgen wird vor allem bei Diktiersystemen verwendet, wo eine uneingeschränkte Eingabe erwartet wird.

Bewertung von Spracherkennungssystemen

Für die Bewertung von Spracherkennungs- und Sprachdialogsystemen gibt es verschiedene Maße, mit denen Eigenschaften des Systems beurteilt werden.

Bei Diktiersystemen ist die Wortfehlerrate das entscheidende Maß in Bezug auf die Spracherkennungstechnologie. Für die Beurteilung eines Gesamtsystems spielen allerdings auch andere Faktoren, etwa die Benutzerschnittstelle oder die Einbindung mit anderen Programmen, eine Rolle. Bei Dialogsystemen ist die Erfolgsrate ein wichtigeres Maß. Eine gute Wortfehlerrate führt nicht automatisch zu einer hohen Erfolgsrate. Denn mit einer schwachen Dialogführung können schon geringe Wortfehlerraten zu einem Scheitern des Dialogs führen. Umgekehrt können durch eine geschickte Dialogführung, beispielsweise Rückfragen, Fehler des Spracherkenners erkannt und kompensiert werden.

Die Benutzerzufriedenheit hängt nicht nur von Erfolgsrate und Produktivität ab. Andere Faktoren, wie zum Beispiel die Qualität der Systemausgabe, der Umgang des Systems mit Erkennungsfehlern und die Nachvollziehbarkeit von Systemreaktionen spielen auch eine große Rolle. *(CK)*