Computer-Spracherkennung

Abstract

Seit jeher wird versucht, den Mensch-Maschine-Dialog zu vereinfachen und mithilfe gewohnter Kommunikationsmittel zu ermöglichen. Die Sprach-Erzeugung ist dabei wesentlich einfacher zu realisieren als die maschinengestützte Sprach-Erkennung, die im folgenden etwas genauer betrachtet werden soll.

Probleme der Spracherkennung

- *Sehr viele verschiedene Aussprachemöglichkeiten*. Auch ein einzelner Sprecher kann ein Wort auf diverse Weisen in der Lautstärke, Artikulation, Sprechgeschwindigkeit etc. verfremden.
- *Ähnlichkeit verschiedener Worte.* Andererseits ändern bereits minimale Unterschiede den Aussagegehalt eines Ausspruches.
- *Wort-Separation im Fließtext*. Auch, wenn einzelne Worte zuverlässig erkannt werden, ist es schwer, in der natürlichen Sprachweise einzelne Worte zu separieren. Es sind heute noch deutliche Sprachpausen zwischen den einzelnen Worten nötig.
- *Verschleierung des Signales*. Hintergrundgeräusche oder andere Stimmen, die der Mensch problemlos ausfiltern kann, stören den Erkennungsprozeß.
- *Begrenztes Vokabular*. Mit der Größe des Wortschatzes steigt die Erkennungszeit stark an und die Erkennungsquote geht stark zurück.

Laut-Erzeugung

Beim Sprechen von Vokalen werden durch den Luftstrom die Stimmbänder in Sinusförmige, periodische Schwingungen versetzt. Es werden dabei ein Grundton mit den zugehörigen Obertönen erzeugt. Der *Sprechtrakt*, bestehend aus Mund, Lippen, Zunge und Gaumensegel, beeinflußt dieses Grundsignal. Einige Frequenzen werden verstärkt, andere abgeschwächt. Das Verhalten des Sprechtraktes kann durch eine *Übertragungsfunktion* dargestellt werden.

Einem wiedergegebenen Laut kann so ein Frequenzspektrum zugeordnet werden, welches aufzeigt, mit welcher Amplitude ein Sinus-Signal einer bestimmten Frequenz in dem Laut vorhanden ist.

Geflüsterte Vokale und Konsonanten besitzen keinen solchen periodischen Grundton, sie verlaufen aperiodisch. Dennoch gibt es ähnliche Übertragungsfunktionen für die verschiedenen Stellungen des Sprechtraktes.

Variationen der Spracherkennung

Zu unterscheiden sind *sprecherunabhängige* und *sprecherabhängige* Erkennungsverfahren. Dabei gibt es wieder starke Unterschiede zwischen der *Einzelworterkennung*, der *Wortkettenerkennung*, der *Wortsuche* und dem *Erkennen fließender Sprache*. Bei der Einzelworterkennung muß nur genau ein Wort erkannt werden, wobei Wortanfang und -ende angegeben sind. Bei Verwendung einer Wortkette muß der Sprecher zwischen den einzelnen Worten eine überakzentuierte Pause (50-100ms) einlegen, um eine saubere Separation zu erlauben. Unter der „Wortsuche“ versteht man den Prozeß, aus gesprochenem Text einige wenige Schlüsselworte auszufiltern und zu hoffen, damit den Sinn des Textes zu erfassen. Das „Erkennen fließender Sprache“ schließlich ist das menschenähnliche Verstehen und Interpretieren einer Aussage.

Mechanismen der Einzelworterkennung

Bei der elektrischen Aufnahme von akustischen Signale wird ein Signal aufgezeichnet, welches direkt mit dem kurzzeitigen Luftdruck am Mikrofon zusammenhängt. Dieses Signal soll nun mit *Referenz-Mustern* in Verbindung gebracht werden, um eine Ähnlichkeit zu bereits trainierten Worten zu finden. Diese Referenzmuster werden in einem Lernprozeß gewonnen, bei dem jedes Wort mehrfach „trainiert“ wird und die Maschine die Daten aufnimmt.

Nulldurchgangsanalyse. Es wird ermittelt, wie häufig das Eingangssignal in einem Zeitintervall zwischen dem positiven und dem negativen Spannungsbereich wechselt. Diese

Sommerakademie IV 1996, Molveno
„Die Sinne: Molekulare Mechanismen bei der Wahrnehmung“
Studienstiftung des deutschen Volkes

Quote entspricht ungefähr der Frequenz der Schwingung, die in diesem Signal dominiert. Die Änderung dieser Dominanz-Frequenz im Zeitverlauf wird mit vorher aufgenommenen Referenzmustern verglichen, etwa indem die Frequenzdifferenzen intervallweise aufsummiert werden. Jenes Referenzmuster mit der geringsten Summen-Abweichung wird dann ausgewählt.

Verfeinerung der Nulldurchgangsanalyse. Durch vorgeschaltete Bandpässe, die aus dem Signal nur bestimmte Frequenzbereiche ausfiltern, welche dann getrennt ausgewertet werden, ergeben sich mehrere Dominanz-Frequenzen in verschiedenen Bereichen.

Zeit-Amplituden-Verfahren. Das Signal wird durch mehrere Bandpässe in Frequenzbereiche zerlegt, deren durchschnittliche Amplitude während eines kleinen Zeitintervalles für die Bewertung herangezogen wird.

Fast-Fourier-Transformation (FFT). Auf schnellen Rechnern ist heute schon eine vollständige Zerlegung des Signales in Sinus-Bestandteile möglich. Es kann so das exakte Frequenzspektrum ermittelt und verglichen werden.

Lineare Prädikation. Statistisches Verfahren, bei dem aus dem gewichteten Mittelwert einiger Signal-Meßwerte der jeweils folgende Meßwert stochastisch vorberechnet wird. Auch hieraus kann ein Amplitudengang ermittelt werden.

All diesen bisher genannten Verfahren ist aber ein Problem gemeinsam: Wie sollen die Daten mit den Referenzmustern verglichen werden? Je nach Aufgabenbereich müssen oben angegebene Probleme möglichst vermieden werden. Alleine die Dynamik-Anpassung und die Geschwindigkeits-Skalierung bereiten Probleme. Um Geschwindigkeits-Schwankungen innerhalb eines Musters ausgleichen zu können, ist es sogar nötig, das Signal eventuell nur partiell zu skalieren. Dieses Prinzip der *Dynamischen Programmierung* erfordert einen sehr hohen Rechenaufwand. Es wurde daher von *Hidden-Markov-Modellen* abgelöst, die die Skalierung nicht mehr explizit durchführen. Doch auch hier ist eine Echtzeit-Verarbeitung nur bei sehr kleinen Wortschätzen (bis zu 300 Wörter) möglich.

Neuronale Netzwerke oder **Hybrid-Systeme**. Am fortschrittlichsten ist bis jetzt die computergestützte Simulation der Natur. Es werden die biologischen Interaktionen von Neuronen „simuliert“, man erhält lernfähige Programme. In Verbindung mit einem der obigen Verfahren ergeben sich hohe Genauigkeiten und Geschwindigkeiten. Ein Vertreter dieses Mechanismus ist das **Feature-Finding-Neural-Network**. In der Vorverarbeitung werden wie gewohnt Kurzzeitspektren des Signales berechnet. Durch mehrere Transformationsschritte werden die Amplitudenstufen der subjektiven menschlichen Wahrnehmung angepaßt, so daß das Spektogramm stark kontrastiert wird. In der *Lernphase* werden diese Spektogramme nach *invarianten Mustern* durchsucht, d.h. nach Kennzeichen, die sich bei mehrfacher Aussprache eines Wortes von einem oder mehreren Sprechern nicht ändern. Die „virtuelle Verschaltung“ der „virtuellen Neuronen“ wird dann den Erfordernissen angepaßt. In der *Erkennungsphase* wird das Spektogramm stückweise betrachtet und nach den bekannten invarianten Mustern durchsucht. Durch spezielle neuronale Programmierverfahren wird auch während der Erkennungsphase das neuronale Netzwerk weiter angepaßt. Viele Probleme werden mit diesem Verfahren umgangen: Es ist (fast) unabhängig von der Sprechgeschwindigkeit und Hintergrundgeräusche stören nur wenig. Es können sogar zwei von verschiedenen Sprechern gleichzeitig gesprochene Worte differenziert und erkannt werden!

Verbesserungsmethoden

Zusätzlich zur rein akustischen Information können Zusatzinformationen verwendet werden, etwa *grammatikalische Regeln* (Festlegung der gültigen Wortreihenfolgen) oder *linguistische Regeln* (Gültigkeit einer Silbenreihenfolge). Die Erkennungsquoten könne auf diese Weise stark verbessert werden.

Literatur

- **S. E. Levinson · M. Y. Liberman**, „Computer lernen hören“
  In: *Spektrum der Wissenschaft*, Juni 1981
- **T. Gramß**, „Worterkennung mit einem künstlichen neuronalen Netzwerk“
  Dissertation an der Georg-August-Universität zu Göttingen, 1992
- **H. Ney**, Ausarbeitung zur Vorlesung „Algorithmen der Spracherkennung“
  RWTH Aachen, Wintersemester 1994/95
- **E. R. Kandel · J. H. Schwartz · T. M. Jessell**, „Neurowissenschaften“
  Spektrum Akademischer Verlag, 1996
- **H. P. Zenner**, „Das Hören“
  Thieme-Verlag, 1994

Referent: Wilko Hein · Osteröder Str. 6b · 38678 Clausthal-Zellerfeld · Tel./Fax 05323 3828
email: Wilko.Hein@Heim2.TU-Clausthal.DE · WWW: http://www.heim2.tu-clausthal.de/~wilko