Folgenabschätzung von Informationstechnologien

SS 2001

Gegenüberstellung von zwei Produkten

Spracherkennungssoftware:

Viavoice Pro Millennium von IBM

vs.

Freespeech 2000 von Philips

von
Tymon Wiedemair
und
Balint Hegedüs

April 2001

Einleitung

Sprache als universelles „Eingabegerät“ ein Traum von dem Informatiker seit etwa 4 Jahrzehnten träumen. Bereits in den 60er Jahren beschäftigten sich Informatiker und Mathematiker mit dem Problem, dem Computer das Hören beizubringen. Damals schätze man, dass dieses Problem innerhalb eines Jahrzehnts gelöst werden könnte. Heute schreiben wir das Jahr 2001 und es gibt, trotz der für damals unvorstellbarer Rechenleitung die heute für jeden erschwinglich ist, noch immer keine universelle Spracherkennung, die die Interaktion mit dem Computer revolutioniert hat.

Die Anfang der kommerziellen Spracherkennung geht auf IBM zurück 1984 wurde ein Spracherkennungssystem vorgestellt, das mit Hilfe eines Großrechners in einem mehrere Minuten dauernden Rechenvorgang etwa 5.000 englische Einzelwörter erkennen konnte. Im Jahre 1986 haben Wissenschaftler des IBM Forschungslabors in Yorktown Heights, USA, den Prototyp TANGORA 4 für Englisch, entwickelt. Der Name wurde in Erinnerung an den Weltrekordhalter im Schreibmaschinenschreiben, Alberto Tangora, gewählt. Bei diesem System war es durch spezielle Mikroprozessoren möglich, die komplizierten Verarbeitungsschritte der gesprochenen Sprache auf einem Arbeitsplatzrechner in Echtzeit durchzuführen. Das bemerkenswerte an diesem System war, dass es bereits eine Kontextprüfung beinhaltete.
Zur CeBit 1994 wurde das erste für den Endkonsumenten erschwingliche Produkt das IBM Personal Dictation System, das kurze Zeit später in IBM VoiceType auf den Markt gebracht. Dieses Produkt war gleichzeitig die erste PC basierende Software in diesem Bereich. Während alle bisherigen Lösungen über 140.000 ATS gekostet hatten, wurde IBM Voice Type für unter 1.000€ bzw. $ verkauft.
Heute tümeln sich etliche Hersteller am heiß umkämpften Zukunftsmarkt der Spracherkennungssoftware für den PC. Einige Produkte die am Markt zu finden sind, sind Naturally Speaking Preferred 4., Viavoice Pro Millennium, Freespeech 2000 und Voice Xpress Professional 4.01.

Im Rahmen dieses Produktvergleich haben wir Viavoice Pro Millennium von IBM und Freespeech 2000 von Philips miteinander verglichen. Diese Produkte werden heutzutage für einige Spezialaufgaben wie z.B. das Erkennen medizinischer Diagnosen und anderer Texte nach Operationen erfolgreich angewendet.
Das große Marktpotential von Spracherkennungssoftware wurde von allen großen Software- und Technologiekonzernen erkannt, was zu einer verstärkten Forschung im Bereich der Spracherkennung führt.

Technische Grundlagen von Spracherkennungssoftware

Bei der Spracherkennung handelt es sich nicht um ein einheitliches System. Sie lässt sich in unterschiedliche Spracherkennungssysteme für unterschiedliche Einsatzmöglichkeiten einteilen.

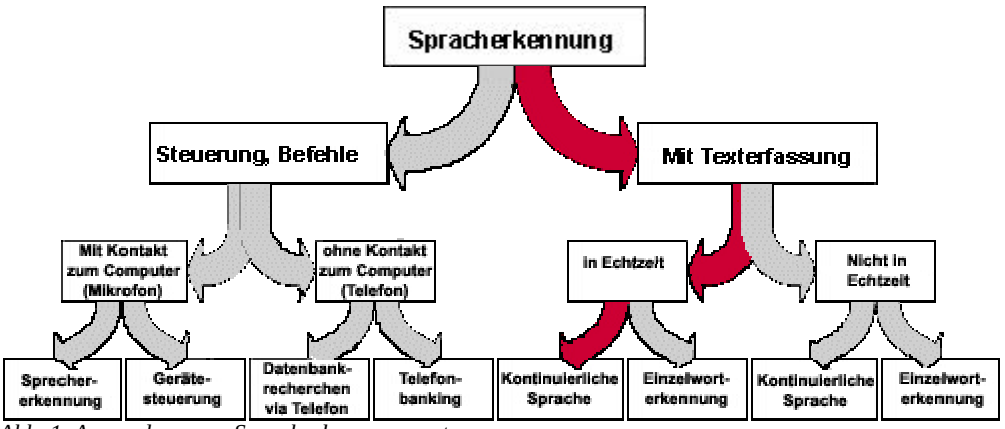

Abb. 1 Anwendung von Spracherkennungssystemen.

Wir werden uns im weitern dem in Abbildung 1 Rot markierten Bereich der Erkennung kontinuierlicher Sprache widmen.
Beim kontinuierlichen Sprechen sind fast alle Wörter lückenlos aneinandergereiht. Dem Menschen fällt es beim Zuhören leicht, die einzelnen Wörter zu unterscheiden. Für eine Maschine ist es um ein Vielfaches schwieriger, einen zusammenhängenden Redefluss zu strukturierten und in einzelne Wörter zu zerlegen.
Die akustischen Signale, die durch das Mikrofon aufgenommen werden, müssen durch das Spracherkennungssystem so verarbeitet werden, dass als Ergebnis ein geschriebener Text vorliegt. Die Verbindung der Akustik mit dem Text wird durch sogenannte Referenzmuster hergestellt.
Ein Spracherkennungssystem besitzt einen großen Vorrat an derartigen Referenzmustern, die sozusagen als "Schablone" für die akustisch aufgenommenen Wörter dienen.
Bei kontinuierlicher Spracherkennung besteht eine Schwierigkeit darin eine akustische Einheit in mehrere Referenzmuster zu zerteilen.

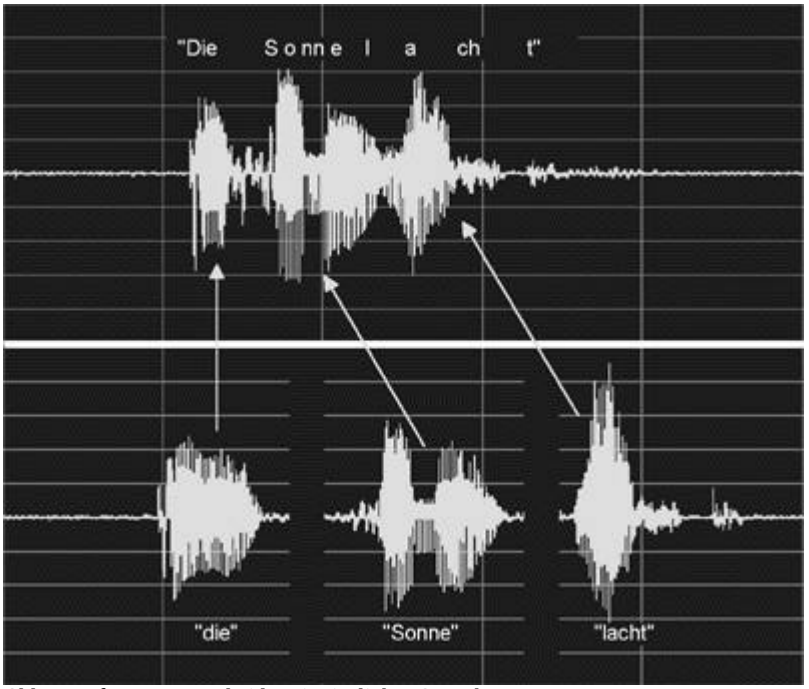

Abb. 2 Referenzmuster bei kontinuierlicher Sprache

Die größte Schwierigkeit bei der Spracherkennung besteht darin, dass ein und dasselbe Wort nie ein zweites Mal absolut identisch ausgesprochen werden kann, selbst wenn es der Sprecher versucht.

Es den eingangs Mustern werden sogenannte Merkmalsvektoren erstellt die durch geeignete Softwarealgorithmen den einzelnen Wörtern zugeordnet werden. Drei wichtige Verfahren die bei der Erkennung dieser Eingangsmuster angewendet werden sind:

- Dynamische Programmierung
- Darstellung in Form von -> "Hidden-Markov"-Modellen
- Künstliche Intelligenz

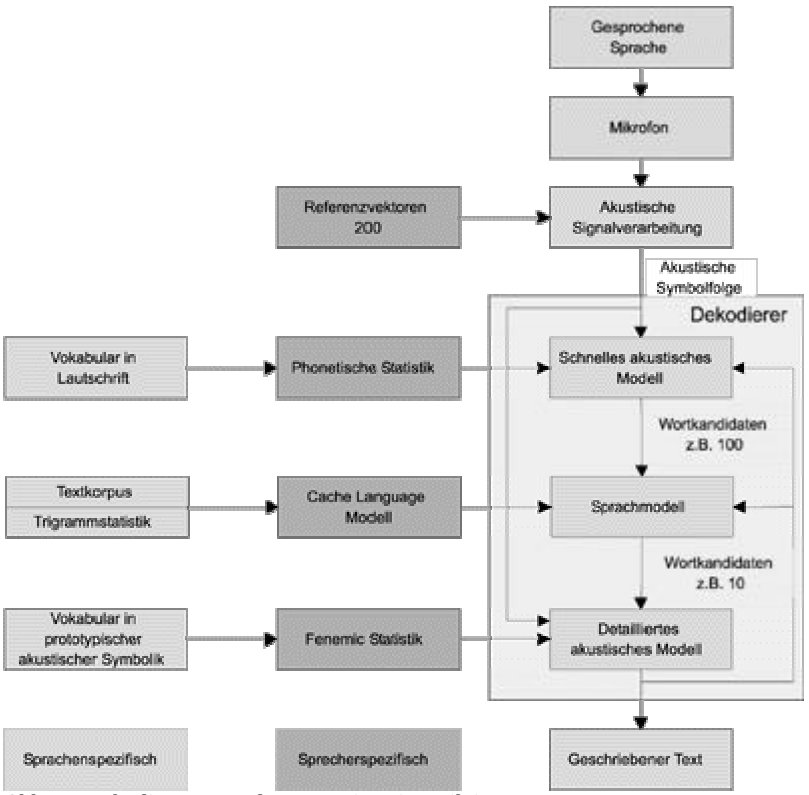

Abb. 11 Dekodierprozess des IBM ViaVoice Diktiersystems

FreeSpeech Produkt Familie
Version: Free Speech 2000 Deutsch

FreeSpeech ist das Softwareprodukt von der Firma Philips. Dieses wird vollständig in Österreich, in Wien entwickelt. FreeSpeech besteht aus drei verschiedenen Komponenten:

1) Browser-Erweiterung
2) Home& Office
3) Professional
4) Entwicklungspaket

Die Browser-Erweiterung kann man von der Homepage der Firma unentgeltlich runterladen. (siehe hierzu die Linksammlung am Ende des Dokuments). Dieses Paket beinhaltet eine Erweiterung für den Internet Explorer, mit deren Hilfe man den Browser steuern kann. Mögliche Interaktionen sind „Back“, „Forward“, „Stop“ usw.. Außerdem kann man auch Links, die sich auf einer Html-Seite befinden, verfolgen. Bei einfacheren Links geschieht das in dem man den Text vorliest und bei komplizierteren wird automatisch eine Nummer

zugeordnet, die dann über dem Link in Form von einer Sprechblase erscheint. Leider liegt das Produkt nur in der englischen Version vor, die mit deutschen Seiten ihre Schwierigkeiten hat.

Die Pakete „Home & Office“ und „Professional“ muss man beide erwerben und kosten ca. 1000 öS bzw. 2000 öS. Der Unterschied zwischen den beiden Version liegt bei der besseren Erweiterbarkeit des „Professional“ Version. Dies ist ganz besonders für diejenigen von Bedeutung, die dann auf spezielle Branchenerweiterungen zugreifen wollen. Es gibt beispielsweise für Juristen, Ärzte und Versicherungen zusätzliches Vokabular, das man sonst nur in mühevoller Arbeit selbst eingeben könnte. Die Erweiterungen umfasse je ca. 64000 Wörter bzw. Wortformen und erweitern so das Referenzlexikon, das ohnehin schon 400.000 Wörter enthält. Mit Hilfe dieser Lösung kann man diktieren oder aber auch gleich das ganze Betriebsystem bedienen.
Das Entwicklungspaket ist für Softwareentwickler gedacht, die Spracherkennung in ihre eigene Applikation integrieren wollen.

ViaVoice
Version: Millennium Edition 2 Deutsch

Bei ViaVoice von IBM gibt es wiederum mehrere Versionen:

1) Personal
2) Standard
3) Advanced
4) Pro
5) Entwicklungspaket

Die Unterschiede liegen vor allem in der Anzahl der vorhandenen Tools. Bei der Advanced und bei der Pro Version ist auch noch ein Headset mit dabei. In der Personal Version kann man nur in den Texteditor von IBM diktieren bei den andere in jedem beliebigen Windowsprogramm. IBM bietet auch eine Macintosh Version von dem Produkt an. ViaVoice kostet zwischen 1000 und 3000 öS, wobei Stundenten die Standard-Version verbilligt (500 öS) erwerben können.

Probleme

Auch wenn die Technik schon sehr ausgereift scheint, funktioniert doch noch nicht alles so, wie man sich das vorstellt. Einfach Installieren und gleich loslegen funktioniert bei keinem der auf dem Mark befindlichen Produkte. Ein Training des Systems, während dessen sich die Software auf die Eigenheiten des Users gewöhnen kann, ist unerlässlich. Beide getesteten Produkte lernen im Laufe der Zeit dazu und erhöhen hierdurch auch die Trefferquote. Nichtsdestotrotz ist es heutzutage noch so, dass sich das System nicht hundertprozentig auf den User einstellen kann, und aus dem Grund sich der User dem System anpassen muss, damit ein relativ reibungsloses Diktieren möglich ist.

Der User muss darauf achten, dass die Umgebungsgeräusche möglichst gering sind. Die Software kann zwar geringe Störungen ausfiltern. Mit steigender Lautstärke sinkt die Trefferquote rapide ab. Aus diesem Grund sollte man auch ein Headset tragen. Hierdurch werden die Umgebungsgeräusche leiser und außerdem stellt man sicher, dass die Entfernung zwischen Mikrophon und Mund, und somit auch der Pegel gleich bleibt. Die Versuche die Spracherkennung mit Hilfe eines billigen Kondensatormikrophons zu betreiben, sind klaglos

gescheitert. Die Software lässt sich zwar trainieren, die Erkennungsrate ist aber dermaßen gering, dass man den Text sogar mit einem Einfingersystem schneller tippen kann.
Auf der anderen Seite sind auch die oft mit den Produkten mitgelieferten Headsets mit Vorsicht zu genießen. Diese erlauben zwar eine recht gute Eingabe, der Tragekomfort lässt aber zu Wünschen übrig. An Headsets muss man sich generell gewöhnen und wenn diese noch drückt oder wenn man darunter schwitzt, wird man sich schweren Herzens an ein noch so tolles System heranwagen.

Auch wenn die Produzenten sehr große Konzerne sind, schaffen sie es nicht eine Software zu produzieren, die stabil funktioniert. Ganz besonders die FreeSpeech-Browsererweiterung stürzt nach ca. 5-10 minütigen Anwendung ab. Dabei wird nicht nur der Browser mitgerissen, sondern Teile des Programms bleiben im Speicher und ein Neustart wird unerlässlich. Abstürze kommen auch bei den anderen Produkten des öfteren vor.

Doch die Probleme, wie lange Trainingszeit, Tragen von Headsets oder auch Abstürzen, sind sicher zu vernachlässigen, solange die Trefferquote hoch bleibt und ein zugiges Arbeiten möglich ist. Laut verschiedenen Test ist man mit der Software in der Lage ca. zweimal so schnell zu diktieren, wie je eine SekretärIn schreiben könnte.

Erkennungsprobleme

Keine der Spracherkennungsprogramme ist in der Lage den Text eindeutig zu zuordnen, d.h. sie wissen aus dem Gesprochenen nicht welche Buchstaben einander folgen. Die Erkennung basiert vielmehr auf einen Vergleich der Silben, die man bei dem Training nach der Installation dem System angelernt hat.
Bei der Erkennung werden zuerst nach Wörter mit den erkannten Silbe gesucht. Es ist aber leider so, dass viele Wörter sehr ähnlich klingen, wie z.B. „mehr“ und Meer“. Dieser Unterschied kann die Software nicht erkennen. Aus dem Grund werden die Sätze analysiert und dabei werden die Wortgattungen bzw. deren Stellung im Satz beachtet, um an die fehlende Information zu kommen. Auch diese Analyse liefert nicht den gewünschten Erfolg weshalb noch der Kontext des vorangehenden Textteils beachtet wird. Die Wörter sind nach bestimmten Themengebieten geordnet, wodurch die Software ungefähr weiß um welchen Themengebiet es bei dem Text handelt. So können dann bestimmte Wörter in dem jeweiligen Gebiete mehr Prioriät bei dem Auswahl bekommen. Die Programme entscheiden dann willkürlich zwischen den möglichen Varianten. Es empfiehlt sicher aber nicht währen einem Text das Themengebiet zu wechseln, weil sonst die Fehlerrate drastisch ansteigen kann. Aus dem Grund eignen sich die Systeme meisten für Leute die sehr ähnliche Texte diktieren und nur ein eingeschränktes Vokabular verwenden.
Größere Probleme ergeben sich auch bei zusammengesetzten Wörter, wie „Weltneuheiten“ oder „Millionenschaden“. Diese werden überwiegend als einzelne Wörter erkannt.

Usability

Beim Testen haben wir feststellen müssen, dass die Erkennungsrate ziemlich hoch liegt, aber die Programme von einer 100% Erkennung weit entfernt sind. Fehler bei der Erkennung sind ganz besonders bei der Korrektur unangenehm. Entweder man macht Verbesserungen mit der Tastatur oder man löscht den gesamten, zu letzt erkannten Wortblock vollständig. Auf Anhieb haben wir uns mit den Programmen nicht anfreunden können und auch wenn diese Technologie sehr zukunftsträchtig ist, bleibt sie für uns vorerst einmal ein Spielzeug, mit dem man sich vielleicht gelegentlich beschäftigt, aber keineswegs arbeiten kann.

Auf der anderen Seite gibt es schon einige Anwender die regelmäßig mit solchen Systemen arbeiten. Hierzu gehören ja Ärzte oder Rechtsanwälte die meistens Befunde bzw. Berichte diktieren müssen, die ein beschränkte Vokabular aufweisen und aus dem Grund besser vom System erkannt werden. Es weiteres Problem beim ungeübten Diktierer ist es ja, dass man dazu neigt Sätze zu diktieren, die nicht ganz komplett sind du deswegen noch umformuliert werden müssen. Wie es schon vorher beschreiben, sind aber Korrekturen ohne Tastatur nicht sehr leicht durchzuführen; und wenn man dann schon sowieso die ganze Zeit mit der Tastatur die Formulierungen umschreibt, dann kann man doch gleich bei der Tastatur bleiben. Dabei wäre noch zu untersuchen wie weit die gesprochenen Fehler darauf zurückzuführen sind, dass man ein neues System verwendet.

Resümee des Vergleichs

Beide Systeme sind aus unserer Sicht vergleichbar. Die Installation verläuft bei beiden Programmen ohne Probleme und die Anleitungen sind auch sehr brauchbar. Bei der Erkennung haben wir keine größere Unterschiede bemerkt, wobei die Test ja nicht mit den gleichen Eingaben passiert ist. Leider erlaubte die Rahmenbedingungen nicht die Programme genauer zu testen. PC-Professional jedoch hat einen sehr genauen Test durchgeführt, auf den wir noch eingehen wollen.

PC-Professionell Test

Der Test ist in der März 2000 Ausgabe zu finden. Dabei wurden vier Produkte zwei Monate lang getestet.
Die nachfolgenden Ergebnisse zeigen auf, dass die Programme schon sehr gut funktionieren und dass die Erkennungsrate mit der Zeit noch steigt.

Kurzzeittest: Erkennungsrate im ersten Durchlauf

Produkt	Erkennungsrate
Naturally Speaking Preferred 4	93,9 Prozent
Viavoice Pro Millennium	90,0 Prozent
Voice Xpress Professional 4.01	89,5 Prozent
Freespeech 2000	88,9 Prozent

Kurzzeittest: Erkennungsrate nach Korrektur

Produkt	Erkennungsrate
Freespeech 2000	97,7 Prozent
Viavoice Pro Millennium	95,9 Prozent
Voice Xpress Professional 4.01	95,9 Prozent
Naturally Speaking Preferred 4	95,5 Prozent

Steigerung der Erkennungsraten durch Training

Produkt	Steigerung
Freespeech 2000	8,4 Prozent

Voice Xpress Professional 4.01	6,1 Prozent
Viavoice Pro Millennium	5,7 Prozent
Naturally Speaking Preferred 4	2,0 Prozent

Steigerung der Erkennungsraten während Langzeittest

Produkt	**Steigerung**
Viavoice Pro Millennium	1,4 Prozent
Voice Xpress Professional 4.01	0,3 Prozent
Freespeech 2000	0,1 Prozent
Naturally Speaking Preferred 4	0,1 Prozent

Zukünftige Visionen

Die Einführung der Spracherkennung ist in gewisser Weise mit der Integration der Maus vergleichbar: Als die ersten Computer mit einer Maus ausgerüstet wurden, gab es kaum Software, die durch eine Maus bedient oder gesteuert werden konnte. Auch die Integration in das Betriebssystem verursachte zunächst Probleme. Heute ist praktisch jede Anwendung mit der Maus zu bedienen, und in den modernen Betriebssystemen mit grafischer Oberfläche wird die Maus standardmäßig voll unterstützt. In einer ähnlichen Situation befindet sich derzeit die Spracherkennung auf dem PC. Zunächst noch skeptisch betrachtet und kaum in einer Anwendung integriert, wird in Zukunft die Bedeutung der Spracherkennung stark zunehmen und bald nicht mehr aus dem täglichen Arbeiten mit dem Computer wegzudenken sein. Doch nicht nur auf allen PCs wird diese Lösung realisiert. Spracherkennung ist auch für viele andere Geräte denkbar.
Stellen wir uns einmal vor, wir könnten Ihrem elektronischen Notizbuch Ihre Texte einfach diktieren, ohne einen PC zu benutzen. Oder wir bräuchten die Bedienungsanleitung des Videorecorders nicht mehr auswendig zu lernen, weil wir ihm sagen können, wann er einen Film aufnehmen soll.
Eine Vision ist auch das vernetzte Auto bei dem fast alles mit der Sprache zu bedienen und zu steuern sein wird. Beim Autofahren sind Hände, Augen und Beine beschäftigt. Die Sprache ist das einzige freie Medium, dem keine Grenzen gesetzt sind.

Interessante Links

Die Produkte:

Free Speech von Philips
www.speech.philips.com

ViaVoice von IBM
www.ibm.com/viavoice

Veranstaltungen zum Themengebiet Sprachsynthese
auf der Universität Wien:
Institut für Medizinische Kybernetik und Arificial Intelligence
www.ai.univie.ac.at

Skriptum Spracherkennung und Sprachsynthese

http://www.ai.univie.ac.at/~hannes/lv_inhalt1/inhalt1.html
von Mag. Hannes Pirker

Artikel im PC-Professionell 03/2000
http://www.zdnet.de/produkte/artikel/sw/200003/speech03_00-wc.html