

Versuch i-09-a

Sprachein- und -ausgabe

Steve Moser, Markus Ganzenmüller

29. März 2003

Inhaltsverzeichnis

1	Prinzipien der Sprachein- und -ausgabe	2
1.1	Funktionsweise der Spracherkennung	2
1.2	Sprachmodell	2
1.3	Akkustikmodell	3
2	Anwendungsgebiete für Sprachein- und -ausgabe	3
3	Sprachsynthese	4
4	Sprechergebundene und Sprecherunabhängige Spracheingabe	5
5	VoiceXML	5
6	S.A.L.T.	6
7	Praktikumerfahrungen	6

1 Prinzipien der Sprachein- und -ausgabe

1.1 Funktionsweise der Spracherkennung

Es gibt mehrere Möglichkeiten eine Sprache zu analysieren. DragonDictate nimmt dazu ein Akkustik und ein Sprachmodell. Das Akkustikmodell besteht aus tausenden von Sprachmustern. Funktionen aus der hohen Mathematik erlauben eine Annäherung an das Sprachmuster und ermöglichen somit eine enge Auswahl an möglichen Wörtern die in Frage kommen.

Das Sprachmodell besteht aus tausenden Dokumentproben und liefert eine Aussage darüber in welchem Zusammenhang und wie häufig ein bestimmtes Wort vorkommt. Quelle: [5]

Ausgangssituation einer jeden Sprachanalyse ist ein vorliegendes Sprachmuster in digitalisierter Form. Dieses wird z.B durch Spracheingabe über ein Mikrofon erzeugt. Das Spracherkennungsprogramm führt dann eine Spektralanalyse durch. Aus der Spektralanalyse werden einzelne Kennwerte ermittelt und zu einem Merkmalsvektor zusammengefasst, der wiederum in einer Tabelle abgelegt wird bzw. mit einer vorhandenen Tabelle verglichen wird um das gesprochene Wort zu rekonstruieren.

1.2 Sprachmodell

Das Sprachmodell liefert die linguistische Wahrscheinlichkeit eines Satzes. Abhängig vom vorgegebenen Vokabular und seines Einsatzgebietes liefert das Sprachmodell eine sehr hohe Wahrscheinlichkeit für einen gesprochenen Satz der in dieses Umfeld passt.

Zum Beispiel würde man ein solches Programm als telefonisches Buchungssystem bei der Deutschen Bahn einsetzen würde der Satz: "Hauptbahnhof Leipzig, Abfahrt 16.30" eine höhere Wahrscheinlichkeit erhalten als "Der Maler geht Freitag ins Grüne". Das Sprachmodell ermittelt seine Sätze anhand von Trigrammen. Das heißt es wird eine Folge von drei aufeinanderfolgenden Wörtern hergenommen. Diese werden dann laut Wahrscheinlichkeitstabelle auf mögliche Nachfolger geprüft. Zum Beispiel das Wort "bis" taucht bei dem Satzgebilde "von neun bis" sehr wahrscheinlich auf. Man hat nun solche Abschätzungen für eine sehr große Anzahl von Trigrammen ermittelt und daraus eine Wahrscheinlichkeitstabelle erstellt. Hier nun die ersten 10 Trigramme.

1. ich rufe an: 80%
2. von neun bis: 77%
3. Anfang nächster Woche: 69%

4. Ihnen das recht: 67%
5. lassen sie uns 61%
6. am Freitag den: 59%
7. wenn Ihnen das: 56%
8. rufe an wegen: 54%
9. ich weiss nicht: 54%
10. es geht um: 49%

[6]

1.3 Akkustikmodell

Das Akkustikmodell kennt man vom Erlernen einer Fremdsprache in der Schule. Bei einem guten Vokabelbuch stehen neben den Bedeutungen der Fremdwörter auch die Lautsprache. Zum Beispiel [d] und [t] für "Daumen" und "Tante". Diese kleinen Lautunterschiede werden in der Fachsprache Phoneme genannt. Die Deutsche Sprache wird z.B mit ca. 40 Phonemen beschrieben.

Folglich entsteht bei der Lautanalyse eine Art Lautfolgebaum. Dadurch reduziert sich die Suche nach einem Wort auf den Faktor 1.5 bis 6.

Kleines Beispiel: "[ae]->[r]->[l]->[aa]->[ih]->[n] == Airline."

Es gibt noch weitere Möglichkeiten ein Sprachsignal zu analysieren. Doch jede Spracherkennungsmethode für sich genommen ist keine Ideallösung. Wie so oft macht die Mischung die Lösung. Die geschickte Kombination aus einzelnen Analysetechniken ist das vorläufige Problem der Forschung.

2 Anwendungsgebiete für Sprachein- und -ausgabe

Hauptsächlich wird das TTS (Text to Speech) System in Betrieben eingesetzt, die versuchen Personal für einfache Aufgaben einzusparen. Folgende Beispiele finden bereits in der Praxis Anwendung:

- Telefonbanking

- SMS-Übertragungen aufs Festnetztelefon
- Sprachdurchsagen auf Flughäfen und Bahnhöfen
- Text2Speech-Systeme aller Art
- Vorlesen von Internetseiten (z.B. für blinde Menschen)

Bereiche die nur einen begrenzten Wortschatz benötigen, wie z.B. Auskünfte, sind sehr beliebt um Spracherkennungs- und Sprachausgabeprogramme einzusetzen.

3 Sprachsynthese

Im ersten Schritt erfolgt eine Transkription, die Übersetzung eines Textes in die entsprechende Lautschrift. Die meisten Verfahren arbeiten hier mit einem Lexikon, das aus einer großen Menge von Wörtern oder auch nur aus Silben oder Lautgruppen besteht. Die Erstellung einer solchen Bibliothek erfordert einen sehr hohen Aufwand, wobei die Qualität über die interaktive Kontrolle des Benutzers kontinuierlich verbessert werden kann: Der Anwender erkennt einen Mangel einer solchen Transkriptionsformel. Er verbessert die Aussprache manuell, wodurch seine Erkenntnisse Bestandteil des Lexikons werden. Hier kann man sich sowohl eine individuelle Implementierung als auch ein gemeinsam genutztes Lexikon vorstellen. Im zweiten Schritt wird die Lautschrift in ein akustisches Sprachsignal umgewandelt. Hier kann dann beispielsweise eine Verkettung im Zeit- oder Frequenzbereich erfolgen. Während der erste Schritt fast immer ausschließlich eine Software-Lösung darstellt, werden im zweiten Schritt neben Signalprozessoren auch dedizierte Prozessoren verwendet.[6]

Neben der Problematik der Koartikulation und der Prosodie ist hier die mehrdeutige Aussprache zu beachten. Eine Aussprache kann oft nur mit zusätzlichem Wissen des Inhalts korrekt erfolgen; sie ist semantikabhängig. Ein Beispiel ist das Wort Wachstube. Dies kann entweder eine Stube mit Wachpersonal oder eine Tube mit Wachs sein. Die Aussprache ist vollkommen verschieden. Diese Problematik kann nur durch zusätzliche Informationen des Kontextes gelöst werden.

4 Sprechergebundene und Sprecherunabhängige Spracheingabe

Ein sprecherunabhängiges System kann bei gleicher Zuverlässigkeit wesentlich weniger Worte erkennen als ein sprechergebundenes System. Dafür muß jedes sprechergebundene System vorab "trainiert" werden. Hierzu werden meistens vorgegebene Sprachsequenzen nachgesprochen. Man geht heute in vielen Spracherkennungssystemen von einer Trainingsphase aus, die weniger als eine halbe Stunde dauert. Sprecherabhängige Systeme können mehr als 25.000 Worte erkennen. Die Erkennungsrate sprecherunabhängiger Systeme liegt in der Größenordnung bis ca. 1.000 Worte. Diese Werte sind allerdings nur als grobe Richtlinie zu verstehen. Quelle: [5]

Bei einem konkreten Vergleich müssen die Randbedingungen sehr genau bekannt sein (Wurde die Messung im schalltoten Raum vorgenommen? Hat sich der Sprecher an das System anzupassen, um beispielsweise die Zeitnormierung zu vereinfachen?).

5 VoiceXML

Voice XML bedeutet "Voice Extensible Markup Language" und ist eine Erweiterung der Sprache XML. Es ermöglicht grob gesagt die Steuerung von Webseiten mittels Spracheingabe. Zumindest sah dies die erste Spezifikation so vor. Laut Entwurf soll Voice XML in der zweiten Version die Fähigkeiten, audio output, audio input, Präsentationslayer, Kontrollfluß und einfaches Eventhandling beherrschen. Quelle: [1] Die Grundlagen von Voice XML wurden in den Labors von AT&T, IBM, Lucent Technologies und Motorola erdacht. Inzwischen hat sich daraus ein Zusammenschluß von etwa 150 Firmen gebildet, welche die Weiterentwicklung unterstützen. Die Spezifikationen der zur Verfügung stehenden Tags wurden im März 2000 von den beteiligten Firmen abgesegnet und im Mai 2000 vom W3C-Konsortium als Voice XML 1.0 Standard angenommen. Quelle: [2] Durch die konsequente Weiterentwicklung von Voice XML ergeben sich bisher ungeahnte Möglichkeiten. Vor allem behinderte Menschen erhalten die Chance von diesem Fortschritt zu profitieren. Blinde oder motorisch behinderte sind nicht mehr vom Informationsangebot des World Wide Web ausgeschlossen. Natürlich sind in diesem Zusammenhang noch eine Flut von Problemen zu lösen, aber die ersten Schritte waren erfolgreich und der Ausblick auf die Zukunft ist vielversprechend.

6 S.A.L.T.

SALT, die Speech Application Language Tags, sind eine Entwicklung von Microsoft, Intel, Cisco und Philips Electronics, welche zusammen mit den Sprachspezialisten von SpeechWorks und Comverse das SALT- Forum gegründet haben, in dem über die Gestaltung von SALT beraten wird. Schon bei der Gründung setzte man sich von Voice XML ab: SALT solle ein Standard für alle sein und mit XML, HTML und XHTML multimodal funktionieren, also nicht allein mit Sprache arbeiten. Multimodal bedeutet, dass ein Anwender, der mit seinem PDA spricht, gleichzeitig mit dem Stift auf dem Bildschirm eine Aktion auslösen kann. Ziel der gesamten Entwicklung ist, dem Benutzer die Interaktion mit bestimmten Anwendungen auf mehrere Arten zu ermöglichen. Erregte die Gründung des SALT-Forums wenig Aufsehen, so schafften dies 18 neue Mitglieder, die dem Forum beigetreten sind. Unter ihnen befinden sich PDA- Hersteller wie Compaq, aber auch Anbieter von Telefonanlagen wie Alcatel, Siemens und Tenovis oder Softwarefirmen wie VoiceGenie, die sich unter anderem auf intelligente Küchengeräte spezialisiert hat. Was anfangs wie der Sonderweg einer Handvoll Firmen aussah, hat sich zu einem Forum gemauert, in dem, entgegen aller Erwartungen, eine offene Diskussion möglich scheint. Ende März 2002 konnte eine Arbeitsgruppe die erste SALT-Spezifikation fertigstellen, die schon im Sommer 2002 dem Web- Standardisierungsgremium W3C vorgelegt wurde. Quelle: [3]

Mittlerweile ist SALT ein lizenzfreier, plattformunabhängiger Standard. Microsoft hat vor kurzem die zweite Betaversion des .NET Speech SDK 1.0 (Software Developer Kit) herausgebracht. Sie steht zum Download bereit, eine CD-ROM-Version ist ebenfalls erhältlich. Mit dem SDK sollen sich Anwendungen erstellen lassen, die eine Verarbeitung von Sprachbefehlen in Web-Sevices ermöglichen. Dadurch sollen Anwender zum Beispiel über das Handy oder ein an den PC angeschlossenes Mikrofon auf Dienste zugreifen können. Quelle: [4]

7 Praktikum Erfahrungen

Im Praktikum mussten wir feststellen dass das Programm DragonDictate einfacher und schneller zu handhaben war. Die Lernphase, also die Sprechphase des Vokabulars (das Stimmenlernprogramm), war wesentlich ausgereifter und erkannte das Vorgelesene sehr schnell. Sie dauerte nur etwa 5 Minuten. Zudem hat man die Möglichkeit die Lernphase, bzw. das gesprochene Vokabular jederzeit zu erweitern. Das Diktieren war auf Anhieb zu etwa 75 Prozent erfolgreich, die Quote verschlechterte sich jedoch bei steigendem Umgebungslärm beträchtlich. Trotzdem eine gute Ausbeute im Vergleich zu Voice Office. Voice Office zeigte sich sehr aufwändig bei der Konfiguration. Auch das Aufsprechen des Vokabulars erwies sich als sehr mühsam. Wir benötigten 25 Minuten um den Text korrekt

aufzusprechen. Beim darauffolgenden Diktat zeigte sich nicht mal eine Übereinstimmung von 50 Prozent. Der Einsatz dieses Programms erfordert wesentlich mehr Übung und genauere Spracheingabe als DragonDictate. Auf den ersten Blick ist Voice Office eher eine Abschreckung, falls man sich entschlossen hat, sich mit der Spracheingabe zu beschäftigen. DragonDictate lässt einen jedoch hoffen, die Ansätze sind lobenswert und können bei einem Neuling schnell Gefallen finden. Ein positiver Punkt ist uns beim Testen von VoiceOffice aufgefallen. Es bietet die Möglichkeit einfache Bedienungskommandos über die Spracheingabe abzusetzen. Z.B. "Schaltfläche Start", oder den Browser zu steuern. Im großen Ganzen kann man einem Neuling DragonDictate als Einstieg empfehlen. Es macht einen schnell mit den grundlegendsten Dingen der Spracherkennungselemente vertraut und ist zudem einfach zu bedienen.

VoiceOffice ist versierten Benutzern zu empfehlen, die sich schon vorher mit Spracherkennungssoftware auseinandergesetzt haben. VoiceOffice bietet weitere Features, z.B. die Steuerung vieler Büroprogramme und ist auch im Funktionsumfang ergiebiger. Eine intensivere Einarbeitung ist jedoch vorausgesetzt.

Literatur

- [1] www.fh-sbg.ac.at/~theistra/etc/lectures/mms/seminararbeiten2002/Brennsteiner_voicexml.pdf
- [2] www.glossar.de
- [3] www.nzz.ch/netzstoff/2002/2002.02.22-em-article7Z9H1.html
- [4] [http://www.contentxxl.de/\(0k2a2ouvg2iaorjs20i3wkbr\)/DesktopDefault.aspx/tabid-283/692_read-5788/](http://www.contentxxl.de/(0k2a2ouvg2iaorjs20i3wkbr)/DesktopDefault.aspx/tabid-283/692_read-5788/)
- [5] <http://www.et-online.fernuni-hagen.de/lehre/k02415.ws/ivz/ivz.htm>
- [6] <http://www.spracherkennung.de>
- [7] http://umwelttechnik.mnd.fh-wiesbaden.de/stud/berger/speech_r.html