Planung für die Vorlesung Spracherkennung und integrierte Dialoganwendungen Sommersemester 2006 am Lehrstuhl Medieninformatik des Institut für Informatik der LMU München

Prof. Dr. Marcus Spies
Institut f. Informatik
LMU München

24. April 2006

Kapitel 1

kurze Einführung

- SpE Systeme sind seit Anfang der 90er Jahre als Produkte im Einsatz
- Anwendungsgebiete waren zunächst Diktiersysteme, zunehmend dann Sprachsteuerung. Derzeit stehen vor allem sprachbasierte Dienste im Fokus, die SpE mit Sprachsynthese zu Dialogsystemen verbinden.
- Technologisch setzte die Produktreife von SpE-Systemen mehrere Entwicklungen voraus, darunter besonders in den 80ern die statistische Modellierung von Sprachdaten anhand von Worthypothesen durch Hidden Markov Modelle.
- Ergänzend wurden auf der Basis neuer Ergebnisse in Psychoakustik und DSP bes. in den 90ern erheblich verbesserte Vorverarbeitungen der Sprachsignale möglich. Damit konnte man den neuen Anforderungen, SpE etwa in die Telephonie zu integrieren und sprecherunabhängige Systeme aufzubauen, besser begegnen.
- Ferner gelang es in den 80ern und 90ern, anhand von Textkorpora für bestimmte Textdomänen „typische“ Wortfolgen statistisch zu modellieren
- Federführend in der Entwicklung der SpE-Technologie zur Produktreife war das IBM-Research Lab mit der bis 1994 von F. Jelinek geleiteten Forschergruppe (Leitung heute M. Picheny). Diese Gruppe arbeitete mit IBM-Wissenschaftszentren weltweit (für DE: WZ Heidelberg) zusammen, so daß die IBM die ersten SpE-Produkte ab 1991 sofort in mehreren Sprachen verfügbar machen konnte. In Heidelberg wurden Textkorpora u.a. aus den Bereichen Pathologie und Nachrichtentexte (Mannheimer Morgen-Korpus) gepflegt. Ferner wurden dort Sprechertrainings für Referenzmodelle der jeweils aktuellen Version der akustischen Vorverarbeitung durchgeführt.
- In den späten 80ern fanden dann aber mit gleichem Gewicht Forschergruppen von Philips, Lernot&Hauspie sowie Dragon Dictate den Weg in die Produktreife. Auch wurde von diesen Firmen in Zusammenarbeit mit zahlreichen Hochschulen die Forschung mit entscheidenden neuen Impulsen angeregt.

- Ebenso in den späten 80ern begründete die CMU das Sphinx-Projekt. Ziel war eine Referenz-Umgebung für SpE sowie die Bereitstellung von Benchmarking Daten. Dank der in Sphinx eingegangenen weltweiten Forschungskooperation hat sich eine weitgehend einheitliche Basisarchitektur für SpE-Systeme herausgebildet.

- eine open source Implementierung entsprechend dieser Architektur steht durch das Sphinx-Projekt der CMU zur Verfügung.

- Die ersten SpE-Produkte setzten Einzelworteingabe (Sprechen mit Wortpausen), Vokabulare um 20K Worte (Flexionsformen zählen als eigene Worte!), sowie umfangreiches Training voraus.

- Seit Mitte der 90er ist kontinuierliches Diktieren möglich; heute werden Vokabulare um140K Wortformen oder größer unterstützt. Ferner ist in Bereichen mit geringem Wortschatz (bes. Sprachsteuerung) sprecherunabhängige Erkennung mit hoher Erkennungsgenauigkeit möglich.

Kapitel 2

Übersicht über die Themen der Vorlesung

2.1 Eigenschaften gesprochener Sprache

Sprachsignale sind Realisierungen phonetischer Folgen, die sich anhand der Schriftform gesprochener Worte nur teilweise voraussagen lassen.

Sprachsignale haben u.a. folgende für SpE relevante Eigenschaften:

- stimmhafte Abschnitte: Grundfrequenz der Stimmbänder variabel durch Intonation und abhängig von Sprechern (Kinder, Frauen, Männer)
- stimmhafte Abschnitte, Formanten: Vokale sind annähernd durch bestimmte dominierende Frequenzbereiche im Resonanzraum des Sprechapparates zu kennzeichnen
- stimmlose Abschnitte: Konsonanten lassen sich teilweise (!) weder in der zeitlichen Ausdehnung noch durch Frequenzbänder oder Frequenzmuster beschreiben; hier spielt der Kontext anderer Phoneme eine wichtige Rolle bei der Erkennung
- Koartikulation – in gesprochener Sprache werden akustische Kennzeichen phonetischer Elemente verschliffen; Sonderfall *vocalic reduction*
- Prosodie Variation der Grundfrequenz der Stimmbänder führt zu Veränderungen der Spektra stimmhafter Segmente
- Lombardisierung

2.2 Auditive Wahrnehmung

SpE-Systeme nutzen für das *front end* (s.u.) Erkenntnisse aus der Forschung zur auditiven Wahrnehmung und Psychoakustik.

Besondere Merkmale menschlicher Sprachwahrnehmung

- frequenzselektive Wahrnehmung durch Verarbeitung des Schallsignals im Innenohr – 2 wichtige Implikationen:
- Trennung von simultanen Sprachmustern mehrerer Sprecher (sog. Cocktal-Party Effekt),
- Unterdrückung von Umweltgeräuschen
- räumliche Wahrnehmung, Eliminierung von Echos etc.

2.3 Komponenten spracherkennender Systeme

Die Komponenten (jeweils mit engl. Bezeichnung entspr. Sphinx-Architektur) sind

akustische Vorverarbeitung (*front end*)

Generierung und Aktualisierung von Erkennungshypothesen (*decoder*)

Verwaltung von Vokabularen (*dictionary*)

Verwaltung von Aussprachemodellen (*acoustic model*)

Bereitstellung und Verwaltung von Sprachmodellen (*language model*)

Anwendung die SpE nutzt, z.B. für Spracheingabe von Befehlen, Diktat

Die nachfolgend genannten Unterpunkte werden im Lauf der Vl erläutert.

2.3.1 Statische Objekte je Erkennungsaufgabe

Vokabular

Aussprachemodelle für wahrscheinliche akustische Realisierung von phonetischen Elementen im Kontext der Erkennungsaufgabe

Sprachmodelle für wahrscheinliche Wortfolgen im Kontext der Erkennungsaufgabe (n-Gramm-Modelle oder spezielle Grammatiken)

Sprecherspezifische Parameter (*voice data*)

2.3.2 Dynamische Objekte je Erkennungsaufgabe:

akustische Daten in verschiedenen Verarbeitungsstufen, optional zusätzliche Daten (z.B. F_0-Tracking, cross-modale Daten wie etwa von Lippenbewegungen)

Sequenz von Merkmalsvektoren

Suchgraph: Graph mit speziellen Strukturen für Worthypothesen und / oder Wortfolgehypothesen

2.3.3 Grundlagen: Aufgaben der Spracherkennung – Grundansatz nach Bayes

Die Komponenten ergeben sich in natürlicher Weise durch einen theoretischen Ansatz zur SpE entsprechend dem Bayesschen Theorem (Bayesscher Klassifikator).

2.4 Modellierung des Sprachsignals

Ziel für spracherkennende Systeme ist Merkmalsextraktion und Reduktion der Information auf ca. 100 Merkmalsvektoren /sec:

- DFT über windowed speech
- Cepstrum (homomorphe Signalverarbeitung; speziell MFCC, PLPCC)
- Aufbau von Merkmalsvektoren (Clusteranalyse, Nutzung zusätzlich von Delta und DeltaDelta Cepstrum)
- Optimierung der Trennungsleistung von Merkmalsvektoren (LDA)

Spezialthema: Multiresolution in der zeitlichen Analyse des Sprachsignals (Wavelets)

2.5 Aussprachemodelle

Grundlegender Ansatz: Hidden Markov Modelle (HMM)

Analyse beobachteter Sequenzen von Merkmalsvektoren in Bezug auf latente Sequenzen phonetischer Elemente –

- Wahrscheinlichkeit einer beobachteten Sequenz
- Ermittlung der wahrscheinlichsten latenten Sequenz anhand einer beobachteten Sequenz
- Schätzung der Parameter eines HMM anhand von Daten

2.6 Phonologische Modellierung

- Zuordnung wahrscheinlicher phonetischer Elemente zu Elementen der Schriftform von Worten
- klassischer phonetischer Ansatz: Morphologische Analyse der Worte des Vokabulars – dieser Ansatz wurde in den 90er Jahren zunehmend durch einen statistischen Ansatz verdrängt:

- Entscheidungsbäume für letter to soundRegeln anhand von Sprachdaten, die das Vokabular repräsentieren
- Weiterentwicklung durch sprecherabhängige, sog. phänonische Modelle

2.7 Sprachmodelle

- Klassische n-Gramm Modelle, die vor allem für Diktieranwendungen genutzt werden. – Hier sind zwei Probleme interessant: Zum einen die Schätzung der Wahrscheinlichkeiten für n-Gramme, die nicht im Textkorpus beobachtet werden (was i.a. die Mehrzahl der n-Gramme ausmacht!), zum anderen die Bereitstellung der sehr zahlreichen Parameter zur Laufzeit des SpE-Systems beim Aufbau und bei der Aktualisierung von Suchgraphen
- Grammatikbasierte Sprachmodelle, die vor allem für Sprachsteuerung und Dialogsysteme genutzt werden.
- Spezialthema: Probabilistische Kontextfreie Grammatiken (SCFG) – interessant, wenn man die Grammatik eines Sprachdienstes aus den Daten erschließen möchte. Evtl. Kandidat zur Ergänzung klassischer n-Gramm Modelle für Diktiersysteme.
- Spezialthema: Sprachmodelle für Wortkomposita. Patent des Autors dieser Vl, in Zusammenarbeit mit der SpE-Gruppe am (ehem.) Wissenschaftszentrum Heidelberg der IBM.

2.8 Suchalgorithmen für die Dekodierung

- Erkennungsalgorithmen bis zur Wortebene
- Erkennungsalgorithmen bis zur Satzebene

2.9 Grundlagen von TTS-Systemen

2.10 Dialogsysteme

- Anwendungsbeispiele für sprachbasierte Dienste
- VoiceXML
- Speech Recognition Grammar