Grundlagen der Spracherkennung

Stefan Petrik
Institut f. Signalverarbeitung & Sprachkommunikation
Technische Universität Graz

Überblick

Einführung

Merkmalsextraktion
- von Waveform bis Cepstrum

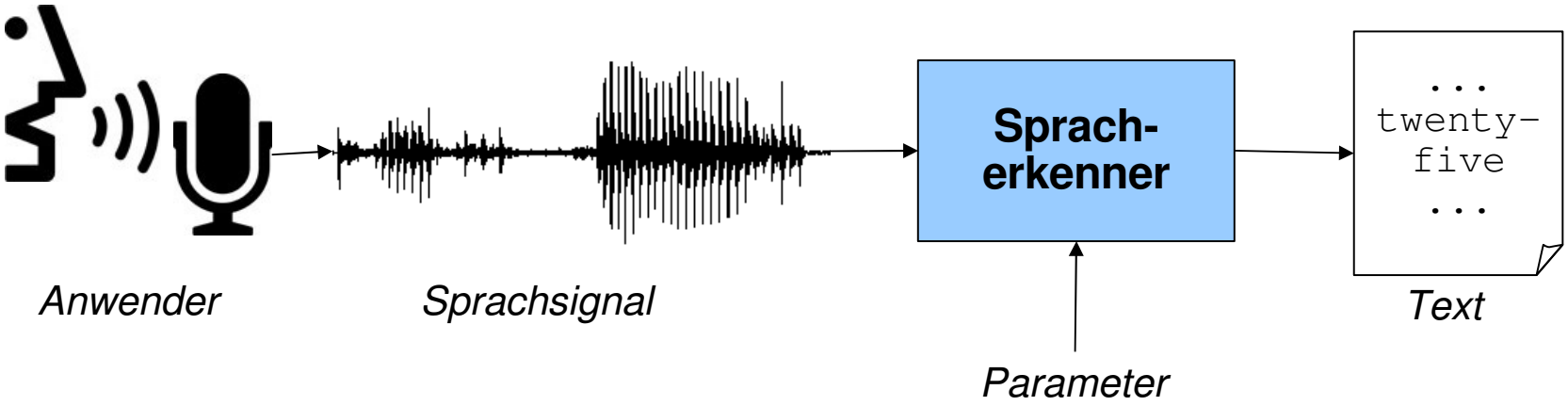
Akustische Modellierung
- Modellierungseinheit für Sprache
- Hidden Markov Modell (HMM)

Sprachmodell
- Regelbasierte Grammatik vs. statist. N-Gramm Sprachmodell

Ein Anwendungsbeispiel

Sprachdatenbanken

Einführung

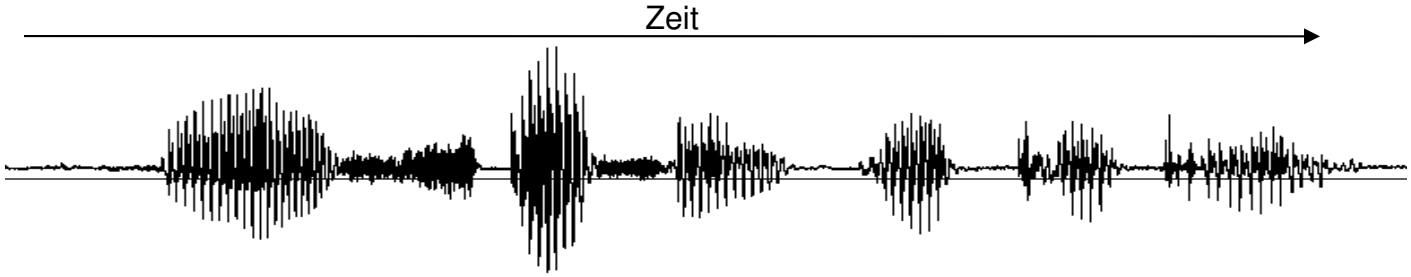


Ressourcen

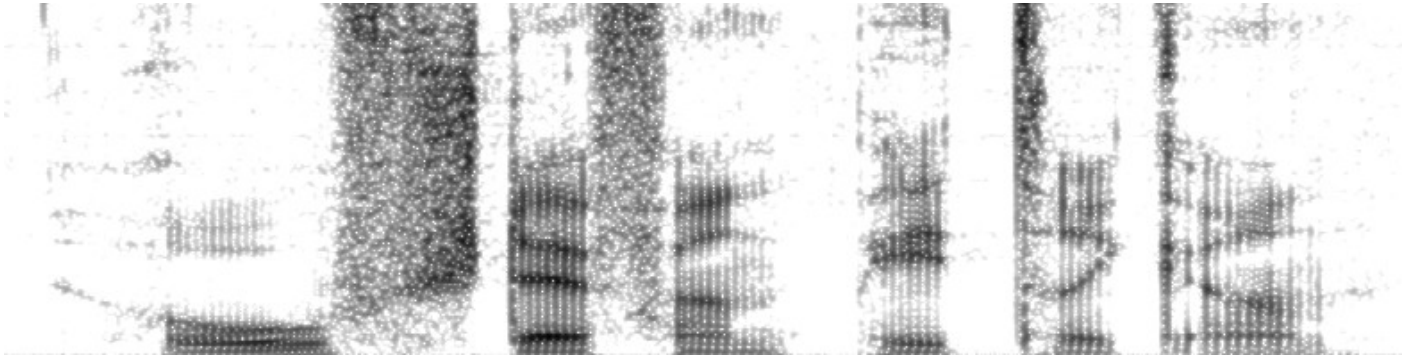
- Lexikon
- Grammatik
- Trainingsdaten
  - Sprachdaten
  - Transkriptionen
- Wissen zur Bildung eines statistischen Modells

Merkmalsextraktion

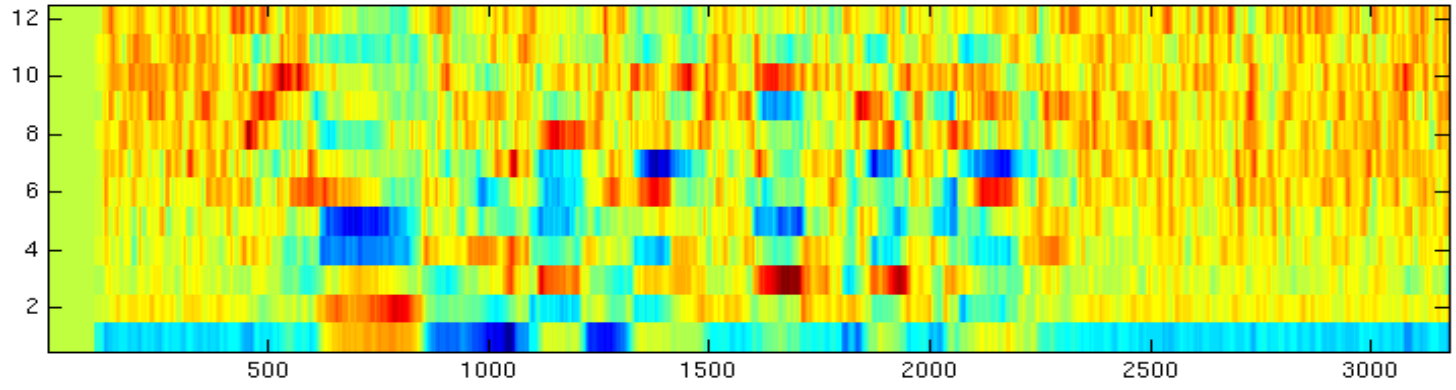
Wellenform



Spektrogramm



Cepstrum:
MFCC
Koeffizienten



Akustische Modellierung

Modellierungseinheit für Sprache

- Ganzwort
  - akkurat
  - viele Trainingsbeispiele nötig f. jedes Wort (i.e. > 10000)
- Silbe
  - weniger Silben als Worte im Wortschatz (< 1000)
  - schwierige Darstellung in Trainingsdaten
- **Phonem**
  - kleinste Einheit mit stationärem Verhalten
  - modular
  - begrenzte Anzahl (ca. 40 Phoneme f. Deutsch)
  - Problem: Koartikulation
  - Lösung: Kontext einbeziehen

Lexikon bildet Phonemsequenzen auf Worte ab

- Wortschatz des Erkenners
- Aussprachevarianten inklusive (Variationen)

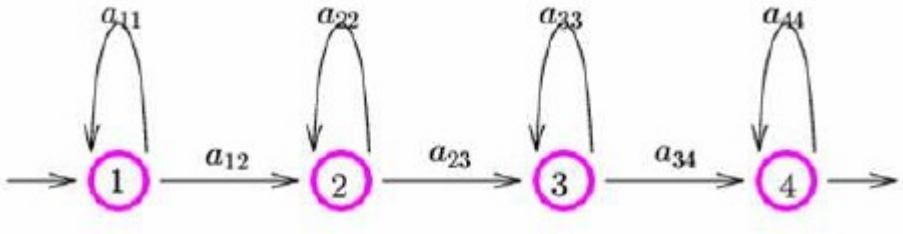
Akustische Modellierung

Problem: Zeitlich veränderliche Sprache

- schnelle vs. langsame Sprecher
- ein Referenzbeispiel pro Phonem ist zu wenig

Lösung: Hidden Markov Modell (HMM)

- modelliere Phoneme als Folge von Zuständen & Übergangswahrscheinlichkeiten
- 1 HMM pro Phonem, 3-5 zustände pro HMM



- Lernen: Anpassung der Modellparameter anhand von Trainingsbeispielen
- Dekodieren: finde die wahrscheinlichste Zustandsfolge, die die vorgegebene Ausgabesequenz erzeugt hat.

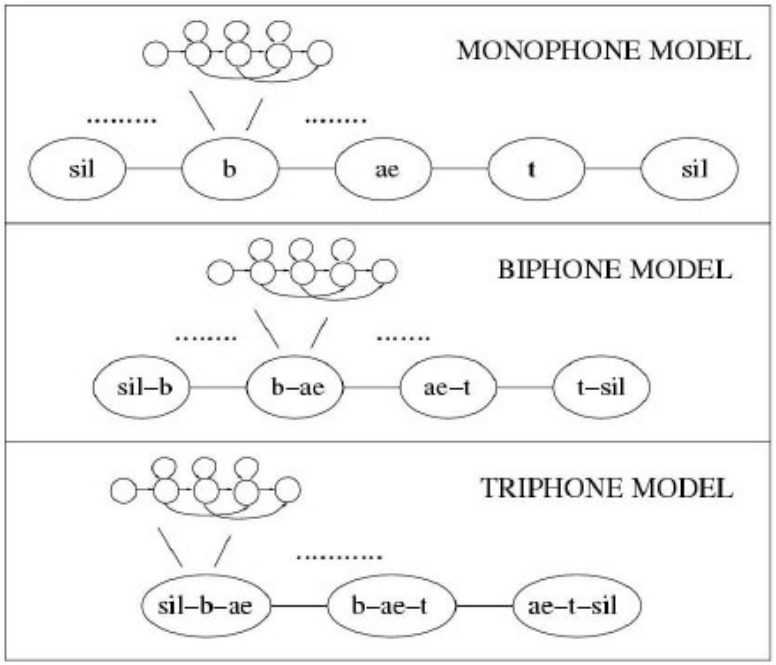
Akustische Modellierung

Problem: Koartikulation

- Verschmelzen von Phonemen am Wortanfang und Ende, z.B. “Roß und Reiter”

Lösung: Triphonmodelle statt Monophonmodelle

- Zusammenfassen v. 3 benachbarten Phonemen
- wort-intern bzw. wortübergreifend
- Koartikulationseffekte gelöst
- mehr Triphonmodelle als Monophonmodelle



Sprachmodell

Regelbasierte Grammatik

- definiere mögliche Wortfolgen anhand von Regeln
- z.B. “Satz → SUBJEKT – PRÄDIKAT – OBJEKT.”
- nicht alles Gesprochene ist auch grammatikalisch
- Komplexität zu hoch für spontansprachliche Eingaben

Statistisches N-Gramm Modell

- Vorhersage des nächsten Wortes anhand der letzten N Worte.
- z.B. “Mein Hund hat ...”
  - gebellt
  - Flöhe
  - Steuererklärung
- lerne Wahrscheinlichkeiten anhand von Beispieltexten (große Mengen an Text, ca. 1 Milliarde Worte)
- “Wahrscheinlichkeit eines Satzes”

$$
\begin{aligned}
P(w_1 w_2 \dots w_T) &= P(w_1) \\
&\cdot P(w_2 \mid w_1) \\
&\cdot P(w_3 \mid w_1 w_2) \\
&\cdot P(w_4 \mid w_1 w_2 w_3) \\
&\cdot \dots\dots\dots\dots \\
&\cdot P(w_T \mid w_1 \dots w_{T-1}) \\
&= \prod_{t=1}^{T} P(w_t \mid w_1 \dots w_{t-1})
\end{aligned}
$$

- $P(w_t \mid w_1 \dots w_t)$ ist die ***bedingte*** Wortwahrscheinlichkeit
- die Wortkette $w_1 \dots w_{t-1}$ heißt ***Vergangenheit*** oder ***Kontext*** des aktuellen Wortes w_t

Ein Anwendungsbeispiel

Sprachgesteuertes Telefon

- Telefonnummern wählen
- Adressbucheinträge wählen

Software

- Hidden Markov Model Toolkit (HTK) von Universität Cambridge
  http://htk.eng.cam.ac.uk
- Soundeditor: Wavesurfer
- Texteditor

Ressourcen

- selbst aufgenommene Sprachdateien, orthografisch transkribiert
- regelbasierte Grammatik
- einfaches Lexikon

Datensammlung

EU-Projekt SPEECHDAT

http://www.speechdat.org/

- 15 Länder in Europa
- Telefonsprache: Festnetz / Mobilfunk
- Kategorien:
  - Befehlswörter und -ausdrücke
  - Ziffernfolgen
  - Zahlen
  - Geldbeträge
  - Datums- und Uhrzeitangaben
  - Buchstabierungen
  - Auskunftsdienst (geograph. Bezeichnungen, Eigennamen)
  - ja/nein Antworten
  - phonetisch reiche Wörter/Sätze
- mind. 1000 Sprecher / Sprache
  - 50% männlich bzw. weiblich
  - gleichverteilt auf 5 Altersklassen
  - regional balanciert