Transkription von Radiospots

Stand der Forschung

Kristoffer Witt - AW2 – HAW-Hamburg

Gliederung

I. Einleitung
   I. Motivation
   II. Überblick: Funktionsweise von Spracherkennung
II. Stand der Forschung
   I. Historie
   II. Aktuell
   III. Ähnliche Arbeiten
III. Einordnung der eigenen Arbeit
IV. Zusammenfassung
V. Quellen

Einleitung

- Motivation
  - AdVision, Webportal für Werbung
  - Transkription von Audiodaten
    - Anlegen eines Suchindex
    - Navigation innerhalb der Daten
    - Unterstützung der Erfassungsmitarbeiter
    - *„Wissen ist Macht“* (Francis Bacon, Religiöse Betrachtungen)

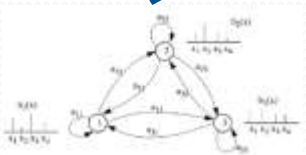
Funktionsweise von Spracherkennung

Audiomodell

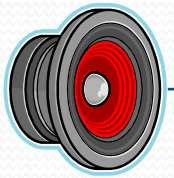
Sprachmodell

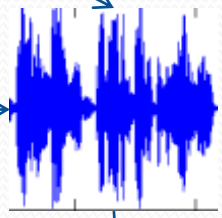
Musik (Lärm)

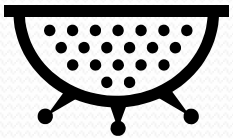
Phonem-Erkennung



Sprachmodellierung







Sprach quelle

Filterung

Merkmals-Extraktion

Digitalisierung / Diskretisierung

TRANSKRIPT

Stand der Forschung

Historie

- 1969 – Linear Predictive Coding (LPC)
- 1975 – Hidden Markov Models
- 1980 – Mel-frequency cepstrum
- 1980er – Language Models und Neurale Netzwerke
- …

Advance	Date	Impact
Linear predictive coding	1969	Automatic, simple speech compression
Dynamic time warping	1970s	Reduces search while allowing temporal flexibility
Hidden Markov models	1975	Treat both temporal and spectral variation statistically
Mel-frequency cepstrum	1980	Improved auditory-based speech compression
Language models	1980s	Including language redundancy improves ASR accuracy
Neural networks	1980s	Excellent static nonlinear classifier
Kernel-based classifiers	1998	Better discriminative training
Dynamic Bayesian networks	1999	More general statistical networks

(O'Shaughnessy, 2008, s. 2967)

Stand der Forschung

Aktuell

- Aktuelle Herausforderungen
  - "For practical applications, errors must be reduced further.“ (O'Shaughnessy, 2008, s. 2797)
- Aktuelle Forschungsfelder (Baker und weitere, Mai 2009, s.76ff):
  1. **„Everyday Audio“**
  2. **Selbstlernende ASR-Technologie**
  3. **Erkennung von seltenen Ereignissen**
  4. **Verbesserung der ASR Infrastruktur**
  5. An die menschliche Wahrnehmung angepasste Erkennung.
  6. Zusammenfassung von gesprochener Sprache
  7. Wissenstransfer für weitere Sprachen

Stand der Forschung

Ähnliche Arbeiten

- **AUTOMATIC TRANSCRIPTION OF GENERAL AUDIO DATA: EFFECT OF ENVIRONMENT SEGMENTATION ON PHONETIC RECOGNITION**
  (*Spina und Zue, 1997)*
  - Erkennung von Radio-Nachrichten
  - Segmentierung in „saubere“/“verrauschte“/“musikhinterlegte“ Sprache
- Ergebnisse:
  - Mehrere Modelle für einzelne Kategorien verbessern erkennung.
  - Z.T. bessere Ergebnisse für verrauschte Sprache mit einem Erkenner für „saubere“ Sprache

Stand der Forschung

Ähnliche Arbeiten (2)

- Ohtsuki und andere, 2006:
  Automatic Multimedia Indexing:
  Combining audio, speech, and visual information to index broadcast news
- Segmentierung in Sprache, Musik, Lärm und Stille
- verschiedene Audio-Modelle für Spracherkennung
  - Single: Ein Einzelnes (geschlechtsunabhängiges) Modell
  - Parallel: mehrere unterschiedliche Modelle, das mit der höchsten Wahrscheinlichkeit wird verwendet.
  - Select: ein kurzes Stück am Anfang wird getestet und das mit der besten Leistung wird verwendet.

Stand der Forschung

Ähnliche Arbeiten (2)

- Ergebnis der ASR
  - word error rate (WER)
  - real time factor (RTF)

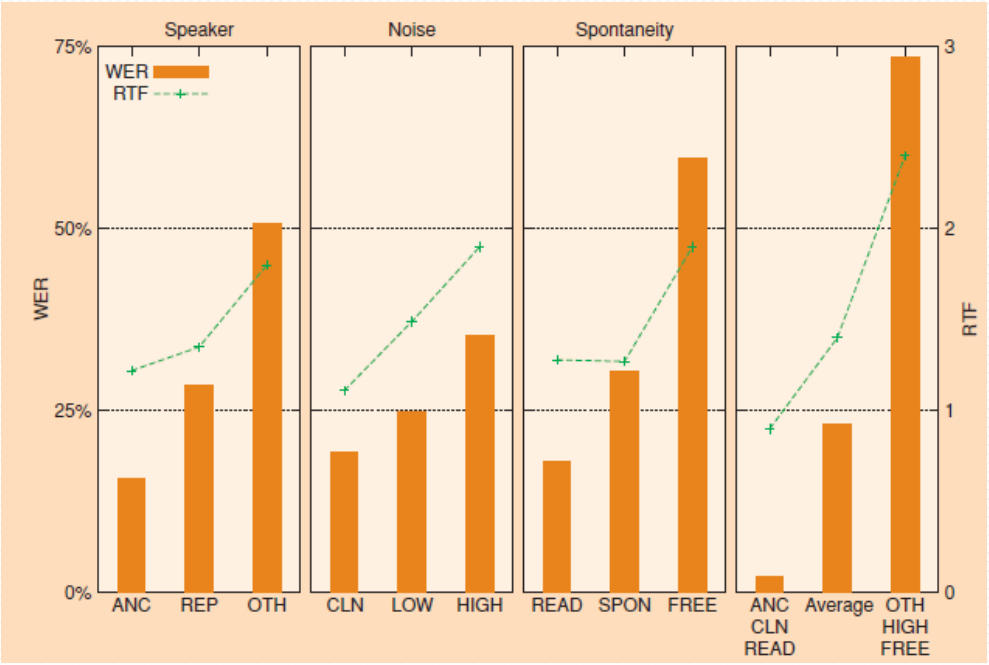


Fig 8, s. 76

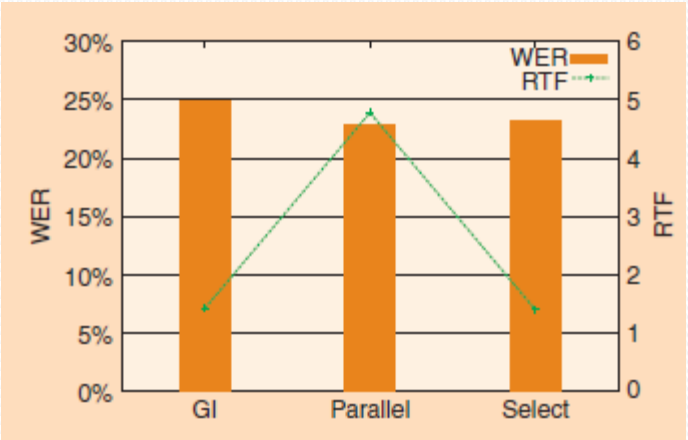


Fig 7, s.75

Grafiken aus Ohtsuki und andere, 2006

Einordnung der eigenen Arbeit

- Eigenschaften
  - Größtenteils deutliche Sprache
  - Hohe Rauschvarianz
  - Unterschiedliche Prosodie für Kontextherstellung
  - Hoher Anteil von OOV (Marken- und Produktnamen)

- Deshalb Erforderlich
  - Validierung vorhandener Modelle, und
  - Erstellung neuer Modelle (Audio- / Sprachmodell)
    - Nutzung vorhandener TV-Werbespotdaten für OOV
    - Aufbau des Sprachmodells aus WWW-Daten bzw. OCR-Daten.
  - Klassifizierung der Daten für besseres ASR Matchings

Zusammenfassung

- Umfangreiches und aktives Forschungsgebiet
- Der Eigenen Arbeit ähnliche Ansätze (ASR von Radio-News-Übertragungen) vorhanden.
- Aktueller Technologieeinsatz zum Teil (Sprecher unabhängige Erkennung) nicht zufriedenstellend
  - => Erstellung eines fehlerfreien Transkripts äußerst unwahrscheinlich.

Quellen

- Magazine
  - EURASIP Journal on Audio, Speech, and Music Processing (http://www.hindawi.com/journals/asmp/contents.html)
  - ACM Transactions on Speech and Language Processing (http://tslp.acm.org/)
  - Multimedia Tools and Applications, Springer Netherlands
  - Signal Processing Magazine, IEEE
- Konferenzen
  - ICASSP, IEEE Conference on Acustic, Speech and Signal Processing (http://www.icasspo9.com/)
  - Association for Computational Linguistics and Joint Conference on Natural Language Processing (http://www.acl-ijcnlp-2009.org)
  - IEEE workshop on Automatic Speech Recognition and Understanding (http://www.asru2009.org/)

Quellen (2)

- Papers
  - *O'Shaughnessy, 2008:*
    Invited paper: Automatic speech recognition: History, methods and challenges, Pattern Recognition, Volume 41, Issue 10 (October 2008),Pages 2965-2979
  - *Baker und weitere, Mai 2009:*
    Baker, Li Deng, Glass, Khudanpur, Chin-hui Lee, Morgan and O'Shaughnessy
    Developments and directions in speech recognition and understanding, part 1
  - Nguyen, 2009:
    IEEE Signal Processing Magazine, Volume 26, Issue 3, May 2009, **TechWare: Speech Recognition Software and Resources on the Web**
  - *Spina und Zue, 1997:*
    AUTOMATIC TRANSCRIPTION OF GENERAL AUDIO DATA: EFFECT OF ENVIRONMENT SEGMENTATION ON PHONETIC RECOGNITION, Proc. Eurospeech 97, pp. 1547-1550, Rhodes, Greece, September 1997.
  - Cook und weitere,1996:
    Real-time recognition of broadcast radio speech
    Cook, G.D. Christie, J.D. Clarkson, P.R. Hochberg, M.M. Logan, B.T. Robinson, A.J. Seymour, C.W. Dept. of Eng., Cambridge Univ.;
    Acoustics, Speech, and Signal Processing, 1996. ICASSP-96, Volume 1 (May 1996) pp 141-144

Quellen (3)

- Ohtsuki und andere, 2006:
  Katsutoshi Ohtsuki, Katsuji Bessho, Yoshihiro Matsuo, Shoichi Matsunaga, and Yoshihiko Hayashi:
  Automatic Multimedia Indexing:
  [Combining audio, speech, and visual information to index broadcast news]
  IEEE SIGNAL PROCESSING MAGAZINE, pp. 69
  MARCH 2006

ENDE

VIELEN DANK FÜR DIE AUFMERKSAMKEIT.

Gibt es noch Fragen?