

Gebärdenerkennung

Modellierung der Sprache und Simultaneität

Christian Vogler, University of Pennsylvania

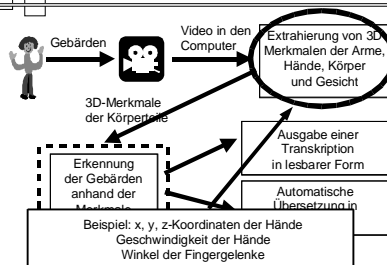
In Zusammenarbeit mit Dimitris Metaxas,
Rutgers University

Überblick

- **Problemstellung**
- Modellierung von amerikanischer GS (ASL)
 - Phonembasierte Modellierung
 - Simultane Ereignisse
- Erkennungssystem
 - Hidden Markov-Modelle
 - Erweiterung zu parallelen Hidden Markov-Modellen
- Experimente
- Ausblick

2

Was ist Gebärdenerkennung?




3

Warum ist es so schwierig?

- Komplexität in der Berechnung und der Modellierung
- ASL ist eine stark inflektierte Sprache
- Zum Beispiel: Das Verb GIVE (GEBEN)
 - Subjekt-Kongruenz
 - Objekt-Kongruenz
 - Simultaneität
- Viele verschiedene Erscheinungsbilder für eine einzelne Gebärde
- Zu viele, um alle einzeln zu erfassen


4



Komplexität

- Ein Rechenbeispiel:
- Nimm durchschnittlich 10 Erscheinungsbilder pro Gebärde an (**vorsichtige Schätzung**)
 - In etwa 6000 ASL Gebärden im Lexikon
 - $6000 \times 10 = 60.000$ Hidden Markov-Modelle
 - Jedes HMM braucht mindestens 10-15 Beispiele für das Training
 - 600.000-900.000 Beispiele insgesamt
- Hat jemand Lust, die alle aufzunehmen?


5



Überblick

- Problemstellung
- **Modellierung von ASL**
 - Phonembasierte Modellierung
 - Simultane Ereignisse
- Erkennungssystem
 - Hidden Markov-Modelle
 - Erweiterung zu parallelen Hidden Markov-Modellen
- Experimente
- Ausblick


6



Phonembasierte Modellierung

- Laßt uns von Spracherkennung abgucken:
 - Zerlegt Wörter in Phoneme
- Die Hauptidee:
 - Die Anzahl der Phoneme in einer Sprache ist **begrenzt**
 - Im Gegensatz zu praktisch unbegrenzter Flexion
 - Baue Wörter aus Phonemen zusammen
- Geht das Gleiche mit ASL-Erkennung?
 - Radio Eriwan: Im Prinzip ja, aber ...

7



Phoneme: Schwierigkeiten

- Diese Idee hat 3 große Probleme:
 - Keine Einigung über ein phonologisches Modell für ASL
 - Linguisten streiten munter darüber!
 - Einige Aspekte existierender Modelle ungeeignet für Computer-Modellierung
 - Linguistik und Informatik haben unterschiedliche Anforderungen
 - Simultane Ereignisse
 - Spracherkennung kann alle Ereignisse in eine Reihenfolge abstrahieren
 - ASL-Erkennung kann das nicht!

8

Das Movement Hold-Modell

- Wir wählen das Movement Hold-Modell
 - Liddell & Johnson (1989)
- Vorteile:
 - Großes Gewicht auf Segmenten
 - Großes Gewicht auf sequentiellm Kontrast
- Segmente ideal für HMM-basierte Erkennung
 - HMMs in ein Netzwerk zusammenkoppeln
= Segmente in eine Gebärde zusammenkoppeln

9

Beispiel: Movement-Hold

Sequentielle Struktur von FATHER

10

Einsatz von Movement-Hold

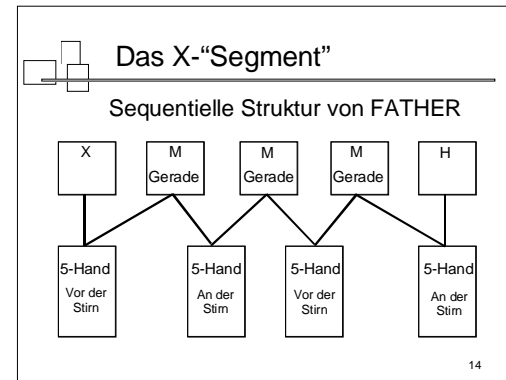
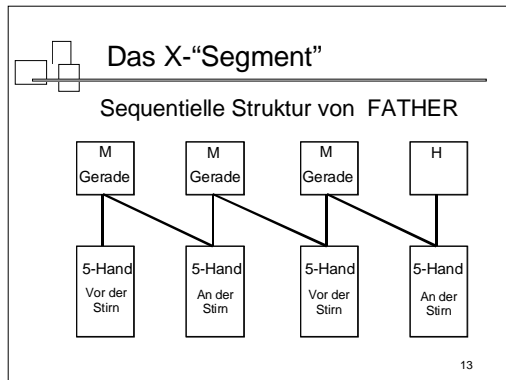
- Relativ geradlinig für die Ausführungsstelle und die Handbewegungen
- Hold: die Hand bewegt sich für kurze Zeit nicht
 - Gut, um die Ausführungsstelle zu erfassen
 - Ein HMM pro Ausführungsstelle
- Movement: die Hand bewegt sich
 - Gut, um Bewegungsrichtung und Geschwindigkeit zu erfassen
 - Ein HMM pro Typ und Richtung von Bewegung

11

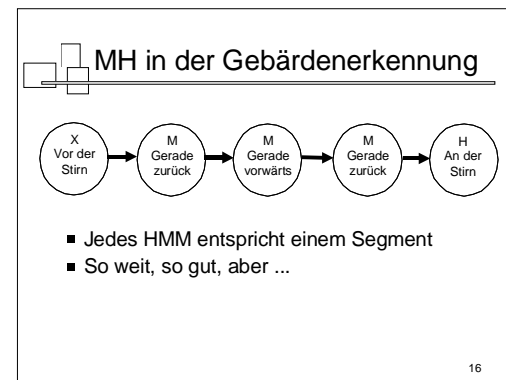
Ein Problem mit MH ...

- Viele Gebärden beginnen **nicht** mit Hold
 - FATHER (VATER), SIT (SETZEN)
- D.h., das System erfäßt die Ausführungsstelle erst am Ende solcher Gebärden
- Nicht gut für die Erkennungsrate
- Erfinde ein neues "Segment" namens **X**
- Ähnlich wie Hold, aber die Hand braucht nicht still zu halten
 - "Schnappschuß" der Ausführungsstelle

12



- ### Na und? Was bringt das alles?
- Wenn wir uns auf Ausführungsstelle und Handbewegungen beschränken:
 - ~ 40 Modelle für Holds
 - ~ 100 Modelle für Movements
 - ~ 40 Modelle für X-“Segmente” (wie Hold)
 - Insgesamt: 150–200 Modelle
 - Vergleiche mit 60.000 vom naiven Ansatz
 - Große Vokabulare scheinen plötzlich **machbar!**
- 15



Simultane Prozesse

- ... wie modelliert man die simultanen Ereignisse in ASL?
 - Zweihändige Gebärden
 - Handbewegungen + Handformänderungen
 - Handstellungsänderungen
- MH packt sie in die artikulatorischen Merkmale
- Wie können wir diese mit dem Computer erfassen?

17

Simultaneität und HMMs

Können wir nicht einfach diese Information zu den HMMs hinzufügen?

18

Komplexität und Simultaneität

- Leider geht das nicht!
- Eine Menge möglicher Kombinationen simultaner Ereignisse
- Eine grobe Schätzung:
 - 30 Handformen, 20 Ausführungsstellen, 8 Orientierungen, 40 Bewegungen
 - $30 \times 20 \times 8 \times 40 \times 30 \times 20 \times 8 \times 40$
- Das sind **37 Milliarden**

19

Komplexität und Simultaneität

- Selbst, wenn nur jede 100. Kombination gültig ist, sind das immer noch viel zu viele
- Das Movement Hold-Modell hilft hier gar nicht
 - Konflikt zwischen Anforderungen der Linguistik und der Informatik
- Wir brauchen eine Methode, um simultane Ereignisse voneinander zu entkoppeln

20

Eine typische Informatikerlösung

- Oder: "Der billige Weg aus der Schlinge"
- Nimm an, daß alle simultanen Ereignisse **unabhängig** voneinander sind
- Zerteilung in **unabhängige Kanäle**:
 - Können unabhängig voneinander gemessen werden
 - Können sehr leicht kombiniert werden
 - **Können unabhängig voneinander trainiert werden**

21

Unabhängige Kanäle

- Unabhängige Kanäle
- Beispiel: BROTHER (BRUDER)

22

Vor- und Nachteile der Annahme

- Vorteile:
 - Anzahl der Modelle ist nur $30 + 20 + 8 + 40 \times 2$
 - Anstelle von $30 \times 20 \times 8 \times 40^2$
 - Reduzierung der Komplexität um **6 Größenordnungen**
 - Unscharfe Grenzen sind gut geeignet, um Antizipation zu modellieren
 - Kombinationen können laufend erstellt werden
- Nachteile:
 - Kanäle oft **nicht wirklich** unabhängig

23

Billige Ausrede

- Billige Informatikerausrede:
- Hauptsache, es funktioniert!

24

Überblick

- Problemstellung
- Modellierung von ASL
 - Phonembasierte Modellierung
 - Simultane Ereignisse
- **Erkennungssystem**
 - Hidden Markov-Modelle
 - Erweiterung zu parallelen Hidden Markov-Modellen
- Experimente
- Ausblick

25

Hidden Markov-Modelle

- Statistisches Modell mit Zuständen
 - System ist immer in einem Zustand
 - Nach einem diskreten Zeitschritt nimmt es eine Transition zu einem anderen Zustand
- Transitionen sind stochastisch
- Jeder Zustand hat Ausgabewahrscheinlichkeit
 - Oftmals eine Gauß'sche Kurve
 - Stellt Wahrscheinlichkeit dar, daß HMM diese Ausgabe in diesem Zustand erstellt hat

26

Beispiel: HMM

The diagram illustrates a Hidden Markov Model (HMM) with five hidden states represented by circles in a horizontal line. Arrows indicate transitions between adjacent states, and a curved arrow shows a transition from the second state to the fifth state. Below each state, an arrow points to a specific point on a continuous signal waveform plotted on a coordinate system with X and Y axes. The waveform is a smooth, wavy line that passes through the points indicated by the arrows.

27

Anwendung von HMMs

- HMMs erstellen ein Signal
- Für die Erkennung kehre dies um:
 - Was ist die Wahrscheinlichkeit, daß ein HMM ein bestimmtes Signal erstellt hat?
 - Welche Zustandsfolge hat es erstellt?
- Diese Fragen definieren das kontinuierliche Erkennungsproblem
- HMM-Wahrscheinlichkeiten werden trainiert

28

Kontinuierliche Erkennung

- Koppele HMMs in ein Netzwerk zusammen
- Vergleiche Netzwerk mit Signal
- Finde die wahrscheinlichste Zustandsfolge durch Netzwerk

29

Kontinuierliche Erkennung

30

Kontinuierliche Erkennung

31

"Token passing"-Algorithmus

32

Formal: "Token passing"

- Der "Token passing"-Algorithmus findet

$$\max_{\mathcal{Q}} \pi_{\mathcal{Q}_1} b_{\mathcal{Q}_1} O_1 \prod_{i=2}^T a_{\mathcal{Q}_{i-1}, \mathcal{Q}_i} b_{\mathcal{Q}_i} O_i$$

$$= \max_{\mathcal{Q}} P(\mathcal{O}, \mathcal{Q} | \lambda)$$
- Das ist ein Produkt unabhängiger Zufallsvariablen

33

Parallele HMMs

- Erweiterung auf mehrere unabhängige Kanäle
 - Maximiere die Gesamtwahrscheinlichkeit
 - Multipliziere Kanalwahrscheinlichkeiten
- Diese Erweiterung formalisiert PaHMMs
- Maximiere jetzt:

$$\max_{\mathcal{Q}^1, \dots, \mathcal{Q}^c} \sum_{c=1}^c \log P(\mathcal{O}^c, \mathcal{Q}^c | \lambda^c)$$

34

Wahrscheinlichkeitskombination

- Im Prinzip: Durchsuche HMM-Netzwerk parallel in jedem Kanal
- Wann findet die Multiplikation statt?
 - Teile das Signal in gewichtete Beiträge von jeder einzelnen Gebärde

$$\max_{\mathcal{Q}^1, \dots, \mathcal{Q}^c} \sum_{c=1}^c \log P(\mathcal{O}^c, \mathcal{Q}^c | \lambda^c)$$

$$= \max_{\mathcal{Q}^1, \dots, \mathcal{Q}^c} \sum_{w=1}^W \sum_{c=1}^c \omega_w^c \log P(\mathcal{O}_w^c, \mathcal{Q}_w^c | \lambda^c)$$

35

Wahrscheinlichkeitskombination

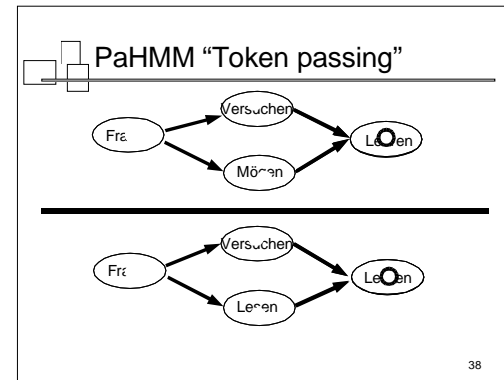
- Was bedeutet das konkret?
 - Wir können die Teilwahrscheinlichkeiten nach jeder Gebärde (oder Phonem) multiplizieren
 - Hilft, die parallele Suche zu beschränken
- Wir brauchen noch eine Einschränkung:
 - Wege durch die Netzwerke müssen in allen Kanälen übereinstimmen
 - D.h., sie müssen die gleiche Gebärdenfolge aufweisen

36

Einschränkung der Kombination

- Einschränkung durch Pfadnummern
 - Erteile jedem Token eine eindeutige Nummer
 - Tokens haben die gleiche Nummer genau dann, wenn sie die gleiche Gebärdenfolge aufweisen
 - Kombiniere nur die Wahrscheinlichkeiten von Tokens mit gleicher Nummer
- Aber: Die beste Gesamtwahrscheinlichkeit maximiert nicht die Randwahrscheinlichkeiten
 - Benutze mehrere Hypothesen pro Zustand
 - d.h., mehrere Token pro Zustand

37



Überblick

- Problemstellung
- Modellierung von ASL
 - Phonembasierte Modellierung
 - Simultane Ereignisse
- Erkennungssystem
 - Hidden Markov-Modelle
 - Erweiterung zu parallelen Hidden Markov-Modellen
- Experimente
- Ausblick

39

Datensammlung

- Vokabular von 22 Gebärden
- 400 Sätze für Training
- 99 Sätze für Testen
- 3D-Daten von einem "Motion Star"-System
 - Magneten liefern Position und Orientierung
- Fingerstellungen von einem Datenhandschuh

40

Merkmalsvektoren

- Wir füttern folgendes an die HMMs:
 - Ausführungsstelle und Bewegungen von dominanter/nondominanter Hand
 - 3D-Positionen, Geschwindigkeit
 - Globale Merkmale: Charakterisierung von Linien und Ebenen
 - Handform von dominanter Hand
 - Winkel der Fingeransätze
 - Messung der Fingerkrümmung

41

Experimente

- Experiment 1
 - Nur Bewegungen und Ausführungsstellen:
 - Dominante Hand, ohne Zerlegung in Phoneme
 - Dominante Hand, mit Zerlegung in Phoneme
 - Welchen Einfluß hat Zerlegung in Phoneme auf die Erkennungsrate?

42

Experiment 1

Bedingung	Erkennungsrate in %
Kontrollexperiment: Keine Zerlegung in Phoneme, keine globalen Merkmale	92.95
Zerlegung in Phoneme, keine globalen Merkmale	90.06
Zerlegung in Phoneme, mit globalen Merkmalen	93.27

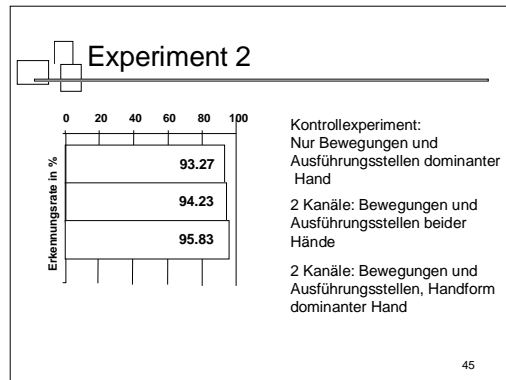
Nur dominante Hand

43

Experimente

- 2 Kanäle mit Zerlegung in Phoneme:
 - Experiment 2a
 - Kanal 1: Dominante Hand, Bewegungen und Ausführungsstellen
 - Kanal 2: Nondominante Hand, Bewegungen und Ausführungsstellen
 - Experiment 2b
 - Kanal 1: Dominante Hand, Bewegungen und Ausführungsstellen
 - Kanal 2: Dominante Hand, Handform
- Wie gut funktioniert Annahme stochastischer Unabhängigkeit in der Praxis?

44



Video

- Zeige Video hier

46

Überblick


- Problemstellung
- Modellierung von ASL
 - Phonembasierte Modellierung
 - Simultane Ereignisse
- Erkennungssystem
 - Hidden Markov-Modelle
 - Erweiterung zu parallelen Hidden Markov-Modellen
- Experimente
- **Ausblick**

47

Wie geht es von hier weiter?

- Überprüfung der Methoden mit größerem Vokabular
- Überprüfung mit Leuten, deren Muttersprache ASL ist
- Mehr Kanäle (Orientierung, Mimik)


48



Wie geht es von hier weiter?

- Bessere phonetische Modellierung:
 - Können wir die stochastische Unabhängigkeit einschränken?
 - Ja, vielleicht: Die Anzahl der **gültigen** simultanen Kombinationen ist viel weniger
 - Stark eingeschränkt durch linguistische Prinzipien
 - Problem: Wie zählen wir sie alle auf?
 - Hier ist die Gebärdensprachforschung gefragt


49



Wie geht es von hier weiter?

- Weitere Forschung in der Bilderkennung-und Verarbeitung
 - Dringend notwendig, bevor Gebärdenerkennung eine Chance in der Praxis hat
 - Datenhandschuhe u.ä. sind einfach zu kompliziert
 - Wir brauchen Erkennung direkt vom Video
- Enorm schwieriges Gebiet

50



Mehr Infos

- cvogler@gradient.cis.upenn.edu
- <http://www.cis.upenn.edu/~cvogler/>

51