HMM-Synthese (Grundzüge)

Uwe Reichel
Institut für Phonetik und Sprachverarbeitung
Ludwig-Maximilians-Universität München
reichelu@phonetik.uni-muenchen.de

9. Januar 2017

Inhalt

- HMM-Grundlagen
- HMM und Phonemerkennung
- HMM-Synthese

HMM-Grundlagen

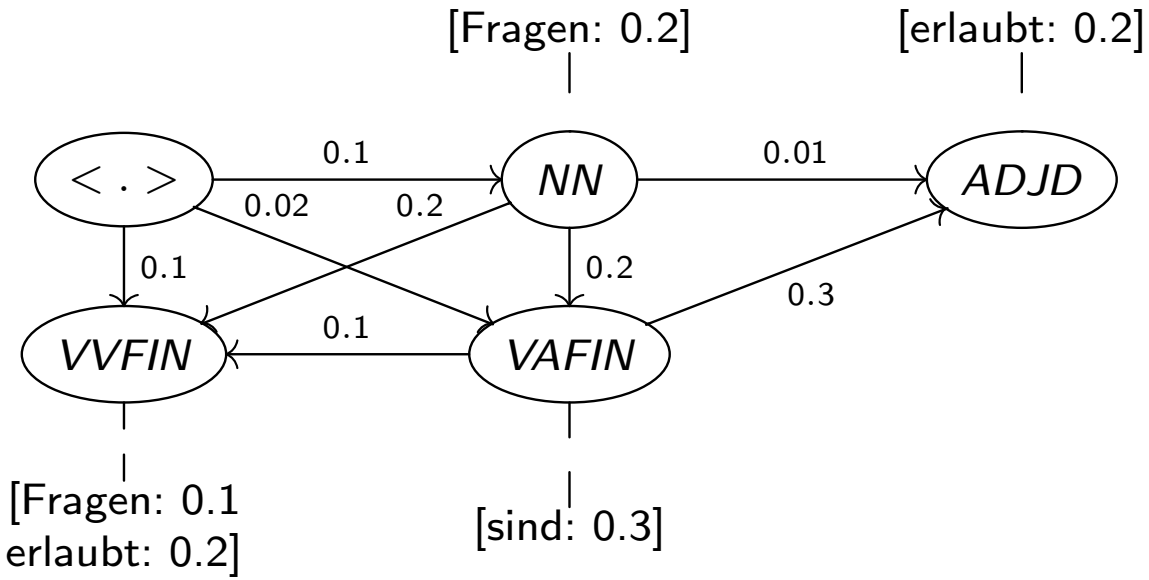
HMM: $< Q, K, S, A, B >$

- $Q = \{q_i\}$: Menge von Zuständen
- K: Ausgabealphabet
- $S = \{s_i\}$: Startwahrscheinlichkeiten, dass man sich zu Beginn in Zustand q_i befindet
- $A = \{a_{ij}\}$ Transitionswahrscheinlichkeiten von Zustand q_i nach q_j
- $B = \{b_{jo}\}$ Emissionswahrscheinlichkeiten für Beobachtung o in Zustand q_j

HMM-Grundlagen

Symbolverarbeitung

- Beispiel Part-of-Speech-Tagging (vgl. POS-Folien)
- $B = \{b_{jo}\}$ sind **eindimensional**: $P(\text{Wort}|\text{POS})$
- Emission ist **kategorial** (hier: ein Wort)



HMM-Grundlagen

Signalverarbeitung

- Emissionswahrscheinlichkeiten $B = \{b_{jo}\}$ sind **mehrdimensional**: $P(e_1, e_2, \ldots | q_j)$
- e_i: Signalcharakteristika (Dimensionen), z.B. Gesamtenergie, Mel-Cepstralkoeffizienten
- mögliche Modellierung als **mehrdimensionale Gaußglocken**
- Emissionen e_i sind **kontinuierlich**

HMM in der Phonemerkennung

HMM_x für Phonem x

- **Zustände** Q: Teilsegmente des Phonems x
- **Emissionswahrscheinlichkeiten** B: Wahrscheinlichkeiten für akustische Ausprägungen von x in den entsprechenden Teilsegmenten

HMM in der Phonemerkennung

Phonemerkennung

- **Aufgabe:**
  - klassifiziere Sprachsignal s
  - finde dasjenige HMM_x, durch das s mit der größten Wahrscheinlichkeit erzeugt wird
- **Lösung:**
  - berechne für jedes HMM_x: $P(s|\mathrm{HMM}_x)$ mittels **Viterbi** (*vgl. POS-Folien*)
  - Zielphonem $\hat{x} = \arg\max_x P(s|\mathrm{HMM}_x)$

HMM in der Sprachsynthese

Training

- **Full-Context-HMM**: je ein $\text{HMM}_{x,c}$ für ein Phonem x in einem bestimmten Kontext c
- c: *Phonemumgebung, prosodischer Kontext, Emotion, ...*
- Emissionen **kontinuierlich**, Emissionswahrscheinlichkeiten B **mehrdimensional**
  - Mel-Cepstral-Koeffizienten
  - F0
  - Segmentdauer
  - Intensität

HMM in der Sprachsynthese

- **Sparse-Data-Problem**
  - Trainingsmaterial reicht nicht aus, um für jedes Phonem x in jedem Kontext c die Parameter für $\text{HMM}_{x,c}$ verlässlich zu schätzen
  - Reduzierung der Anzahl unterschiedlicher Kontexte durch Clustering nach akustischer Ähnlichkeit
- Erstellung eines **Entscheidungsbaums**, der in Abhängigkeit von Phonem x und Kontext c das passende $\text{HMM}_{x,c}$ auswählt

HMM in der Sprachsynthese

Anwendung

- **Input:** segmentale, prosodische etc. **Zielspezifikationen** wie in Unit-Selection-Synthese
- Auswahl des passenden HMM mittels Entscheidungsbaum
- **kontextabhängige Phonemsequenz**
  $< x_1, c_1 > \ldots < x_n, c_n >$:
  Verkettung von $\mathrm{HMM}_{x1,c1} \ldots \mathrm{HMM}_{xn,cn}$
- Generierung des Signals über den **wahrscheinlichsten Pfad** durch die HMM-Kette (mittels **Viterbi**)

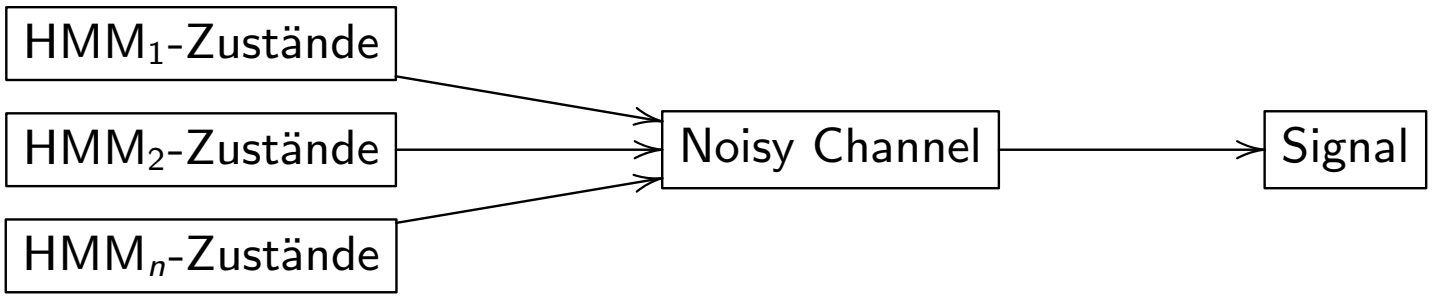
HMM in der Sprachsynthese

Gemeinsamkeiten zwischen Phonemerkennung und Synthese

- Es werden **kontinuierliche Werte** emitiert.
- Emissionswahrscheinlichkeitsverteilungen sind **mehrdimensional**.

HMM in der Sprachsynthese

Unterschiede zwischen Phonemerkennung und Synthese



- **Erkennung:**
  - ein HMM$_x$ je **Phonem** x
  - **bekannt: Kanalausgabe** (das akustische Signal)
  - **gesucht: Kanalinput**, d.h. dasjenige HMM$_x$, das dem Signal am wahrscheinlichsten zugrundeliegt

HMM in der Sprachsynthese

- **Synthese:**
  - **Full-Context:** ein $\text{HMM}_{x,c}$ je **Phonem** x **und Kontext** c
  - **bekannt: Kanalinput**, d.h. das zu den segmentalen, prosodischen etc. **Zielspezifikationen** passende $\text{HMM}_{x,c}$
  - **gesucht: Kanalausgabe**, d.h. das durch $\text{HMM}_{x,c}$ am wahrscheinlichsten generierte akustische Signal

HMM in der Sprachsynthese

Modularisierung

- z.B. **Trennung von Quelle und Filter**
  - **Training** von $\text{HMM}_{q,c}$ und $\text{HMM}_{f,c}$
  - q Quelle: Parameter des Anregungssignals
  - f Filter: Filterparameter
  - **Anwendung:** getrennte Steuerung von Quell- und Filterparametern
  - Erhöhung der Flexibilität

HMM in der Sprachsynthese

Vorzüge

- **geringer Speicheraufwand:** HMM-Parameter statt Datenbank mit Sprachsignalen
- **hohe Flexibilität:**
  - kontinuierliche Steuerung der akustischen Parameter
  - getrennte Modellierung von Quelle und Filter
  - Generierung neuer Stimmen

Nachteile

- **Vollsynthese** $\longrightarrow$ derzeit noch schlechtere Qualität als bei Unit-Selection