Akustisch-phonetische Wortmodellierung

Franz Pernkopf

Institute of Communications and Wave Propagation
University of Technology Graz
Inffeldgasse 16c , 8010 Graz, Austria
Tel: +43 316 873 4431
E-Mail: pernkopf@inw.tugraz.at

Bayes-Regel

$$\boldsymbol{w}^* = \underset{\boldsymbol{w} \in \mathcal{W}^*}{\operatorname{argmax}} P(\boldsymbol{w} \mid \boldsymbol{X})$$

mit den ***a posteriori Wahrscheinlichkeiten***

$$P(\boldsymbol{w} \mid \boldsymbol{X}) = \frac{\overbrace{P(\boldsymbol{X} \mid \boldsymbol{w})}^{\text{ASM}} \cdot \overbrace{P(\boldsymbol{w})}^{\text{LSM}}}{P(\boldsymbol{X})}$$

GEGEBEN:

- Eine Wortfolge $\boldsymbol{w} = w_1 \ldots w_n$

GESUCHT:

- Ein Markovmodell $\boldsymbol{\lambda}(\boldsymbol{w})$ mit $P(\boldsymbol{X}|\boldsymbol{\lambda}(\boldsymbol{w})) \approx P(\boldsymbol{X}|\boldsymbol{w})$

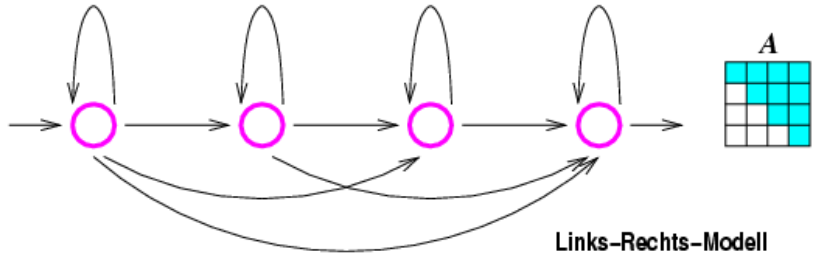
Praktisch gibt es eine unbeschränkte Anzahl von Wortfolgen =>
Sequentielle Verkettung von Wortmodellen

$$\boldsymbol{\lambda}(\boldsymbol{w}) = \boldsymbol{\lambda}(w_1) \circ \boldsymbol{\lambda}(w_2) \circ \ldots \circ \boldsymbol{\lambda}(w_n)$$

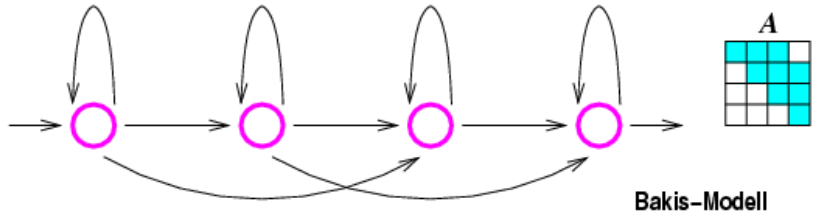
- Erzeuge Wortmodelle $\boldsymbol{\lambda}_l = \boldsymbol{\lambda}(W_l)$ mit der Eigenschaft

$$P(\boldsymbol{X}|\boldsymbol{\lambda}_l) \approx P(\boldsymbol{X}|W_l)$$

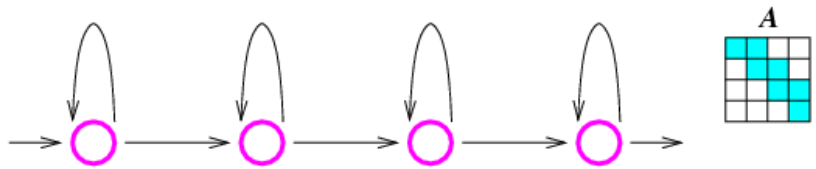
Unterschiedliche Modelltopologien:



Links–Rechts–Modell



Bakis–Modell



Lineares Modell

Modelldimension

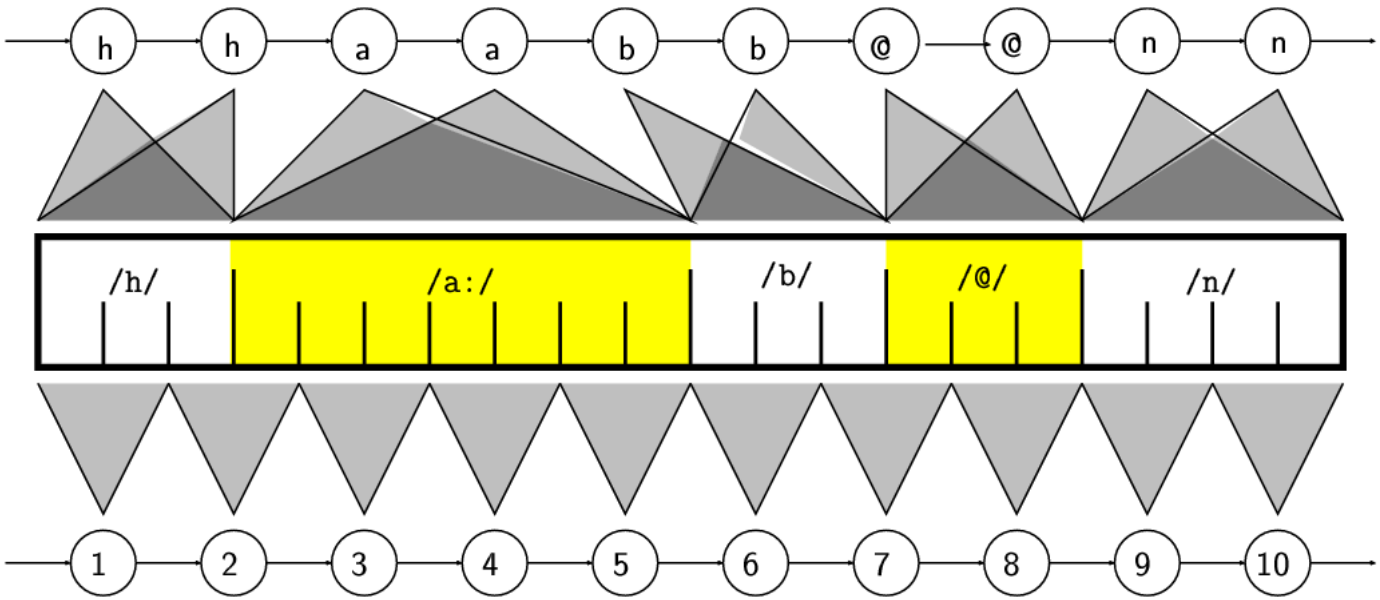
- Anzahl N_l der HMM-Zustände des Modells $\boldsymbol{\lambda}_l$
- Ziffernwörter ⇨ $N_l = 5$
- $N_l = N(W_l) \propto$ „Anzahl der Phoneme von W_l“ Individuelle Wörter: Wortabhängige Modelldimension

Parameterinitialisierung

Vorbesetzen der HMM-Parameter zu Beginn des Baum-Welch-Trainings

- Anfangs- und Übergangswahrscheinlichkeiten π_i, a_{ij} unkritisch
- Diskrete Ausgabeverteilung / Mischungskoeff. b_{jk}, c_{jk} uniform
- Kontinuierliche - und Mischverteilungsparameter $\boldsymbol{\mu}_{jk}$, $\boldsymbol{\Sigma}_{jk} = ?$

Phonetische Segmentierung



Lineare Zeitverzerrung

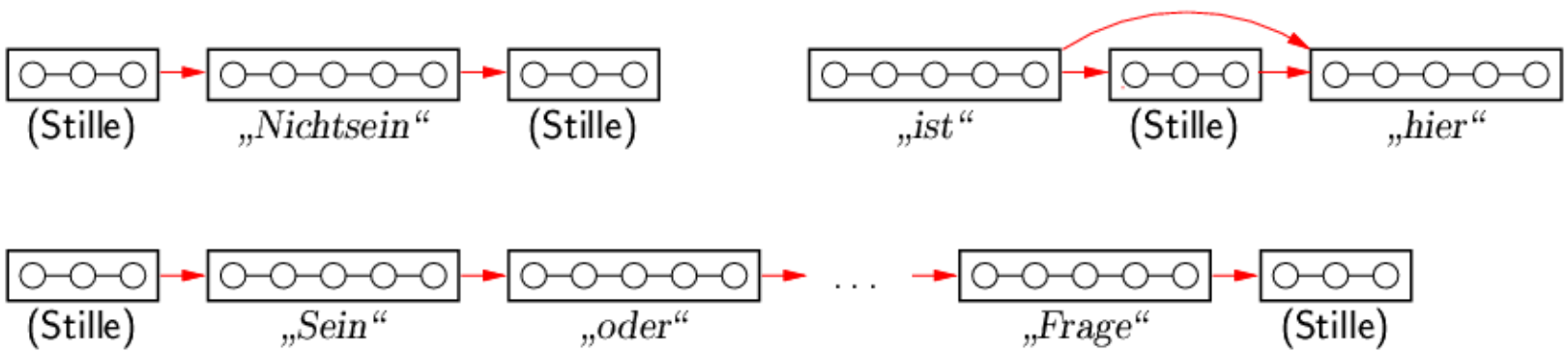
Aufbau der Lernstichprobe

- ***Ein Sprecher:***
  mindestens 10, besser 50 oder 100 Aussprachebeispiele/Wort
- ***Viele Sprecher:***
  ausgewogene Population von mindestens 100 SprecherInnen
  (Geschlecht, Anatomie, Dialekt, Ideolekt, Soziolekt)

⇨ Kompensation von ***Anzahl*** & ***Ergiebigkeit*** von Sprechern

Eingebettetes Lernen

- ***Einzelwortprobe:***
  Wortrealisierungen liegen in Sprechpausen eingebettet vor
- ***Verbundwortprobe:***
  Wortrealisierungen liegen in komplette Sätze eingebettet vor
- Wortfolgenrealisierungen sind u.U. durch *Stillebereiche* unterbrochen



Ganzwortmodelle ?

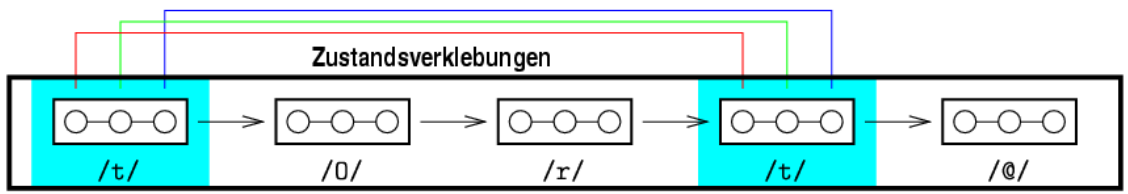
Warum ist das Wort keine geeignete Modellierungseinheit für die autom. Spracherkennung?

- erforderliches Trainingsmaterial $\propto$ Wortschatzumfang L
- enorme Zahl freier HMM-Parameter
  - ***labile Schätzwerte & hoher Rechen- und Speicheraufwand***
- geringe Flexibilität monolithischer Modelle
  - ***Trainingsvokabular = Erkennungsvokabular***
- keine Berücksichtigung wortübergreifender Ausspracheverschleifungen

Analyse–durch–Synthese

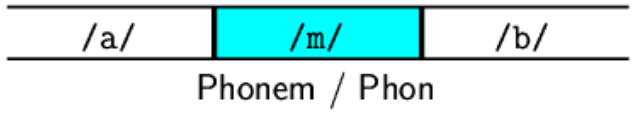
Beispielwort: „*Torte*“

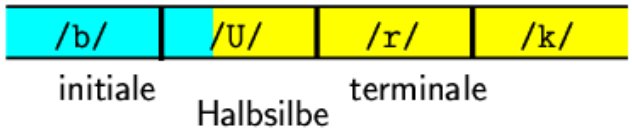
$$\boldsymbol{\lambda}(/\texttt{tOrt@}/) = \boldsymbol{\lambda}(/\texttt{t}/) \circ \boldsymbol{\lambda}(/\texttt{O}/) \circ \boldsymbol{\lambda}(/\texttt{r}/) \circ \boldsymbol{\lambda}(/\texttt{t}/) \circ \boldsymbol{\lambda}(/\texttt{@}/)$$

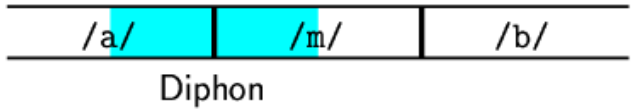


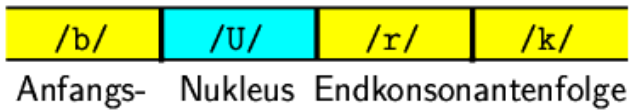
Entwurfskriterien für geeignete Wortuntereinheiten (WUE)

- ***Präzision***
  die WUE ist *hochspezialisiert* und folglich *trennscharf*
- ***Robustheit***
  großer *Trainingsmaterialvorrat* & wirksame *Glättungsmaßnahmen* ➡ gute *Schätzwerte*
- ***Modularität***
  festes, *endliches Inventar* moderaten Umfangs für alle potentiellen Sprechakte
- ***Transfer***
  Synthese neuer Wortmodelle aus vorhandenen WUE









Phoneme	je nach Sprache 20–60 *(+modular, −präzis)*
Silben	20 000 im Englischen, 100 im Japanischen; Koartikulation primär innerhalb *(+präzis, −modular)*
Halbsilben	800/2560 initiale/terminale im Deutschen *(+trennscharf, ±modular)*
Sylparts	47 AKF, 20 Nuklei, 159 EKF im Deutschen *(guter Kompromiß)*
Diphone	1000–1500 Einheiten (engl./ital.), ungünstige Nahtstellen *(besser: Transeme)*

Kontextabhängige Wortuntereinheiten

GRUNDIDEE:

- Segmentelle Basisspracheinheit **kurzer Dauer**
  — Modellierung **kontextabhängig**
- Basiseinheit kontextueller Modellierung ist das **Phonem**

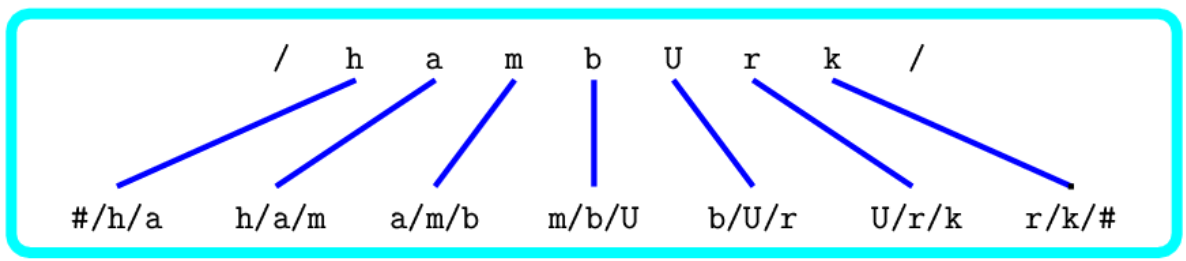
⇨ *„[...] Aufgliederung eines Phonems in allophonische Klassen mit scharf umrissenen artikulatorischen oder akustischen Eigenschaften unter Kontrolle seiner unmittelbaren lautlichen Nachbarschaft [...]“*

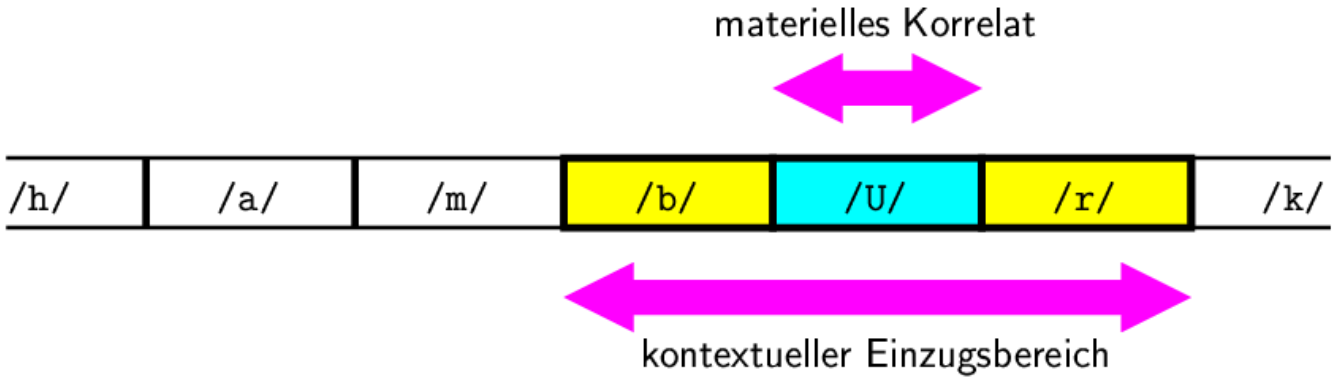
Schreibweise für **Triphone**

`r` in `hambUrk` ⇨ `U/r/k`

und für **rechte / linke Biphone** und **Monophone**

`r` ⟶ `/r/k` `r` ⟶ `U/r/` `r` ⟶ `/r/`





Training von Triphon-HMM's

1. Initialisierung / Training gewöhnlicher **Monophonmodelle**
2. Training der **Biphonmodelle** (⇦ Monophonparameter)
3. Training der **Triphonmodelle** (⇦ Biphonparameter)
4. Konstruktion der Wortmodelle des Erkennungsvokabulars:
   - ***Rückgriffstrategie***:

     $\boxed{\lambda(\mathrm{b/U/r})} \rightsquigarrow \boxed{\lambda(\mathrm{/U/r})} \rightsquigarrow \boxed{\lambda(\mathrm{b/U/})} \rightsquigarrow \boxed{\lambda(\mathrm{/U/})}$
   - ***Interpolation***:

Verallgemeinerte Triphone

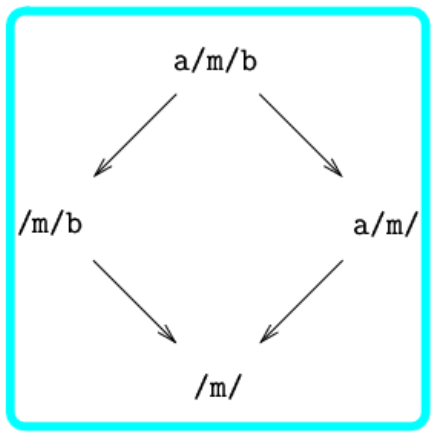
PROBLEM:

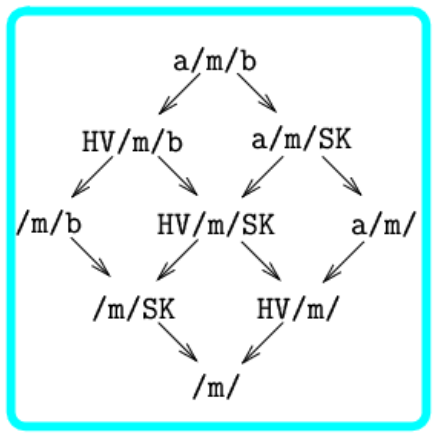
- Seltene Triphone ➡ statistisch labile Markovmodelle

LÖSUNG:

- **Bündeln** geeigneter Gruppen **kerngleicher Triphone** mit verwandten akustischen Eigenschaften

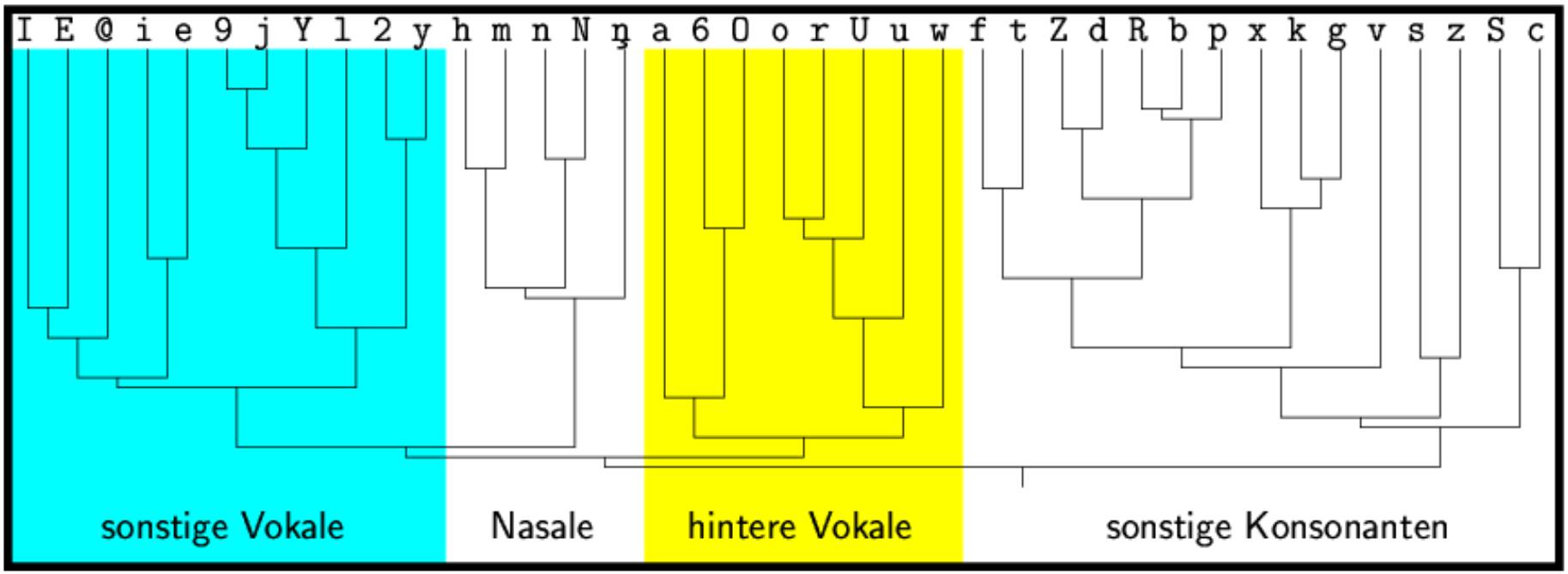
Resultierende Generalisierungsbeziehungen:



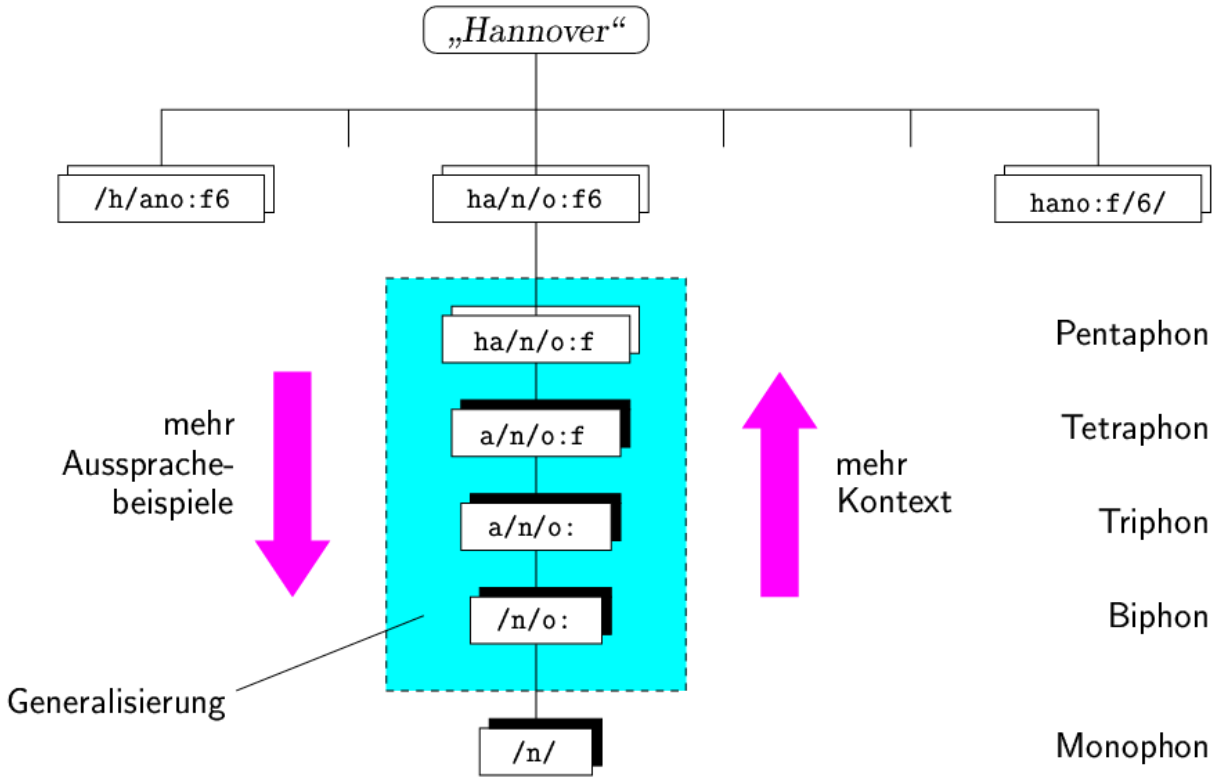


Umschriftgetriebene Generalisierung

Phonetisches Klassensystem für Nachbarphoneme



Polyphone



- häufige Spracheinheiten sind modellierungs**fähig**
  - ⇨ robuste Schätzwerte
- häufige Spracheinheiten sind modellierungs**bedürftig**
  - ⇨ Redundanz & Verschleifung
- Phoneme in **beliebig breitem** R/L Kontext

Wortübergreifende Verschleifungseffekte:

„in München“

/In/ + /mYnc@n/ ⇒ /ImYnc@n/

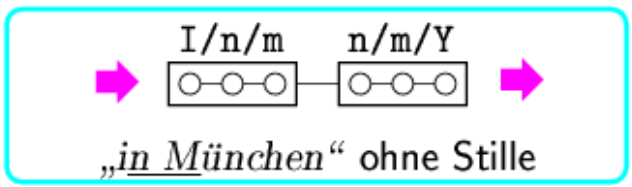
„Verstümmelung“ kurzer Funktionswörter:

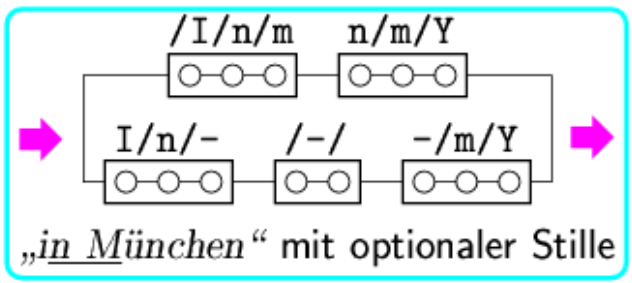
„Roß und Reiter“

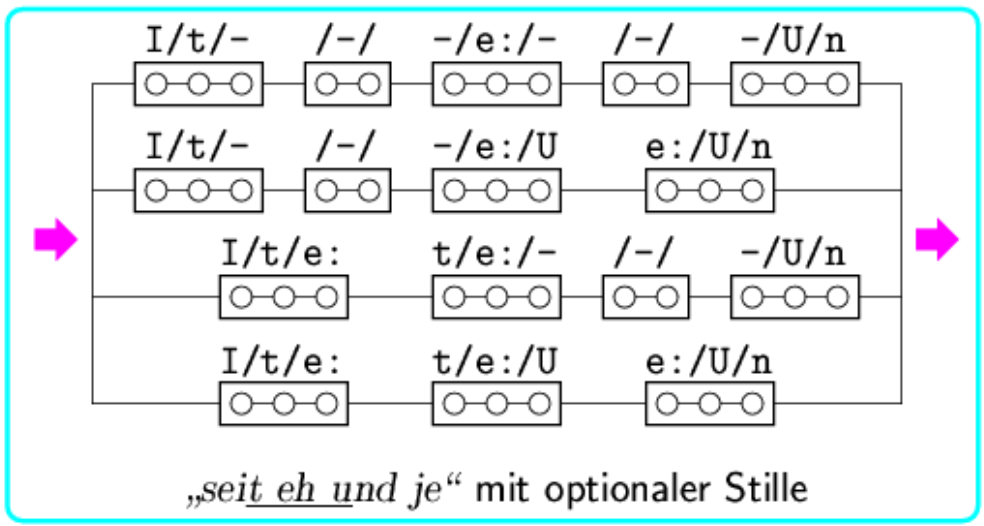
/rOs/ + /Unt/ + /raIt6/ ⇒ /rOsnraIt6/

Wortgrenzenübergreifende Triphonmodelle

für die Lernphase des Spracherkenners







„Performanzlücke“

FEHLERFREI GELESENER TEXT

—

SPONTAN PRODUZIERTE SPRACHE

- ***Unbekannte Wörter***
  - *Wörter außerhalb des Erkennungswortschatzes*
- ***Außerlexikalische Einheiten***
  - *Ungefüllte Pausen*
  - *Gefüllte Pausen („äh“, „mmh“, ...)*
- ***Nichtverbale Realisierungen***
  - *Räuspern, Husten*
  - *Lachen*
  - *Atemgeräusche, Schmatzlaute*
- ***Nichtartikulatorische Störproduktionen***
  - *Türenschlagen, Rascheln, Klopfen, ...*

Referenzen

- Duda, R.O. and Hart, P.E. , *Pattern Classification and Scene Analysis*. Wiley&Sons, Inc., 1973.
- Bolter R., *Bildverarbeitung und Mustererkennung,* Vorlesung ICG Graz, 2000.
- S. Bengio, *An Introduction to Statistical Machine Learning – EM for GMMs,* Dalle Molle Institute for Perceptual Artificial Intelligence.
- E.G. Schukat-Talamazzini, *Automatische Spracherkennung,* Vieweg-Verlag, 1995.
- F. Pernkopf, *Automatic Visual Inspection of Metallic Surfaces,* PhD Thesis, Leoben 2002.
- C.R. Houck, J.A. Joines, G.M. Kay, *A Genetic Algorithm for Function Optimization: A Matlab Implementation,* North Carolina State University.
- M. Obikito, *Introduction to Genetic Algorithms,* Hochschule für Technik und Wirtschaft Dresden, 1998.