Rheinisch Westfälische Technische Hochschule Aachen
Lehrstuhl für Informatik VI
Prof. Dr.-Ing. Hermann Ney

Seminar „Speech Recognition and Language Processing“ im SS 2003

Speaker Adaptation using Eigenvoices

Daniel Schneider

30.07.2003

Literatur

- R. Kuhn, J.C. Junqua, P. Nguyen, N. Niedzielski:
  "Rapid Speaker Adaptation in Eigenvoice Space",
  IEEE Transactions on Speech and Audio Processing, Vol. 8, Nr. 6, S. 695-707, 2000.

- H. Botterweck:
  "Very Fast Adaptation for Large Vocabulary Speech Recognition Using Eigenvoices",
  Proc. Int. Conf. on Spoken Language Processing, Vol. 4, S. 354-357, Bejing, Oktober 2000.

- M.J.F. Gales:
  "Cluster Adaptive Training of Hidden Markov Models",
  IEEE Transactions on Signal and Audio Processing, Vol. 8, Nr. 4, S. 417-420, 2000.

Gliederung

Speaker Adaptation using Eigenvoices

- Einführung und Motivation
- Modell-basierte Sprecheradaption
- Sprecheradaption mit Eigenvoices
- Ergebnisse
- Zusammenfassung

Gliederung

Speaker Adaptation using Eigenvoices

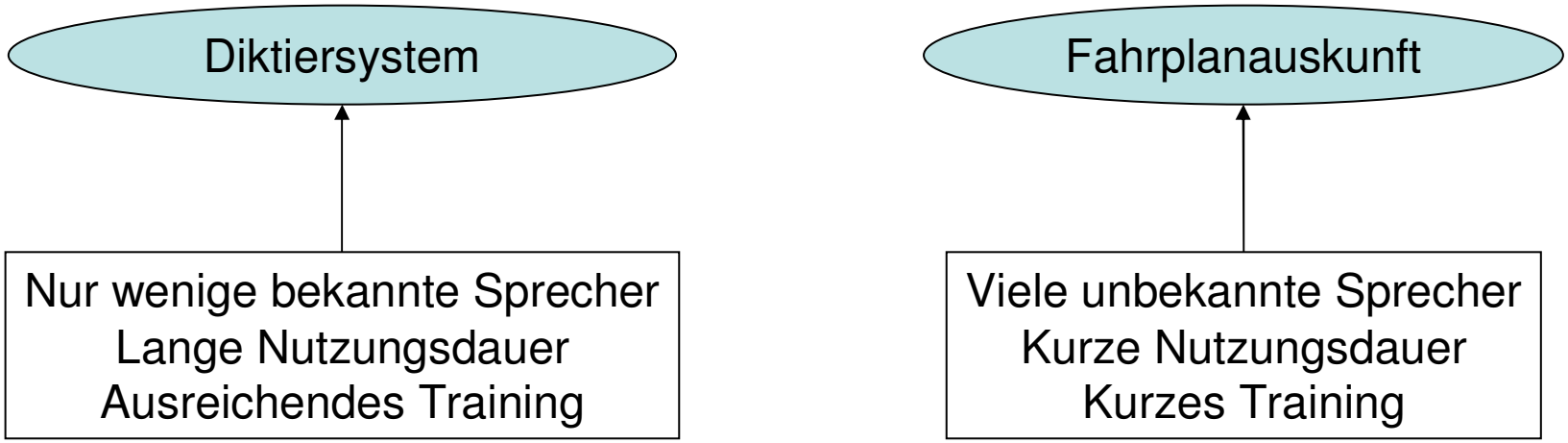
- **Einführung und Motivation**
- Modell-basierte Sprecheradaption
- Sprecheradaption mit Eigenvoices
- Ergebnisse
- Zusammenfassung

Einführung und Motivation

- Aufgabe: Spracherkennung mit unbekanntem Sprecher
  - Telefonauskunft, Archivierung von Nachrichtensendungen,…
- nur wenig Trainigsdaten vom Sprecher
  - Anrufer toleriert kein langes Training,…
  - Vgl. Diktiersystem

Einführung und Motivation

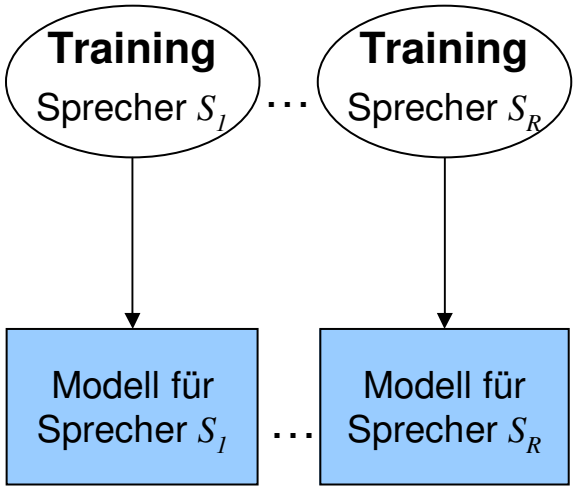
Zielkonflikt: Sprecher-Unabhängigkeit ⟺ Fehlerrate



Fehlerrate eines Spracherkenners ⟺ „Information“ über den Sprecher

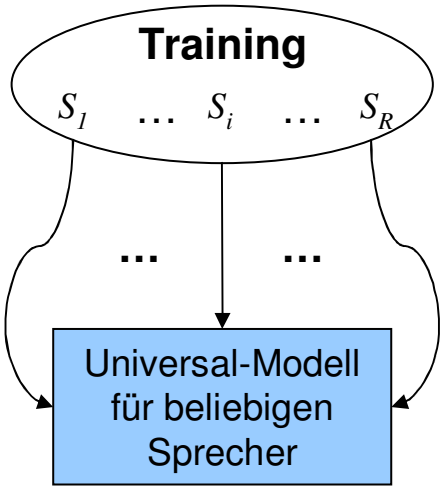
Einführung und Motivation

Sprecher-abhängig



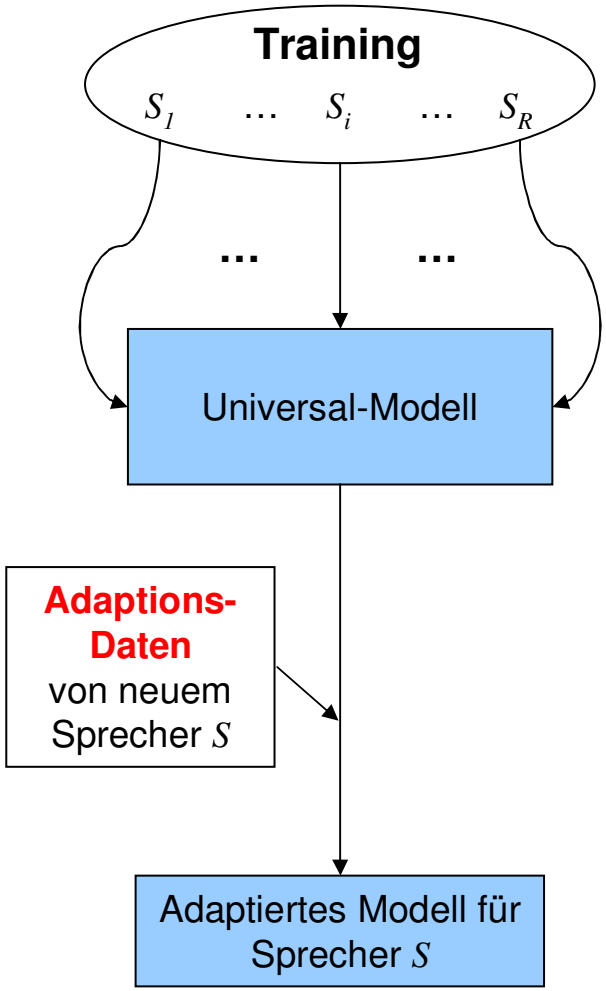
Bsp: Diktiersystem

Sprecher-unabhängig



Bsp: Telefonauskunft

Sprecher-Adaption



Einteilung Sprecher-Adaption

Textkenntnis

- Überwacht *(supervised)*
  - Transkription der Adaptions-Daten bekannt
- Nicht überwacht *(unsupervised)*
  - Keine Transkription der Adaptions-Daten vorhanden
  - Zuerst Erkennung der Adaptions-Daten selbst nötig (z.B. mit sprecher-unabhängigem Modell)

Dynamik

- Statisch
  - Erst Adaption, dann Erkennung
  - 2-Pass: mit dem gleichen Sprachsignal (vgl. Nachrichtensendung)
- Inkrementell
  - Modelle werden nach jeder Äußerung neu adaptiert
  - Mehr Adaptions-Daten ⇒ Verringerung der Fehlerrate

Einordnung

...

4. Speaker Adaptation using MLLR
5. **Speaker Adaptation using Eigenvoices**

} Modell-basierte Sprecher-Adaption

6. Vocal Tract Length Normalization
7. Histogram Normalization

} Adaption im Merkmals-Raum

...

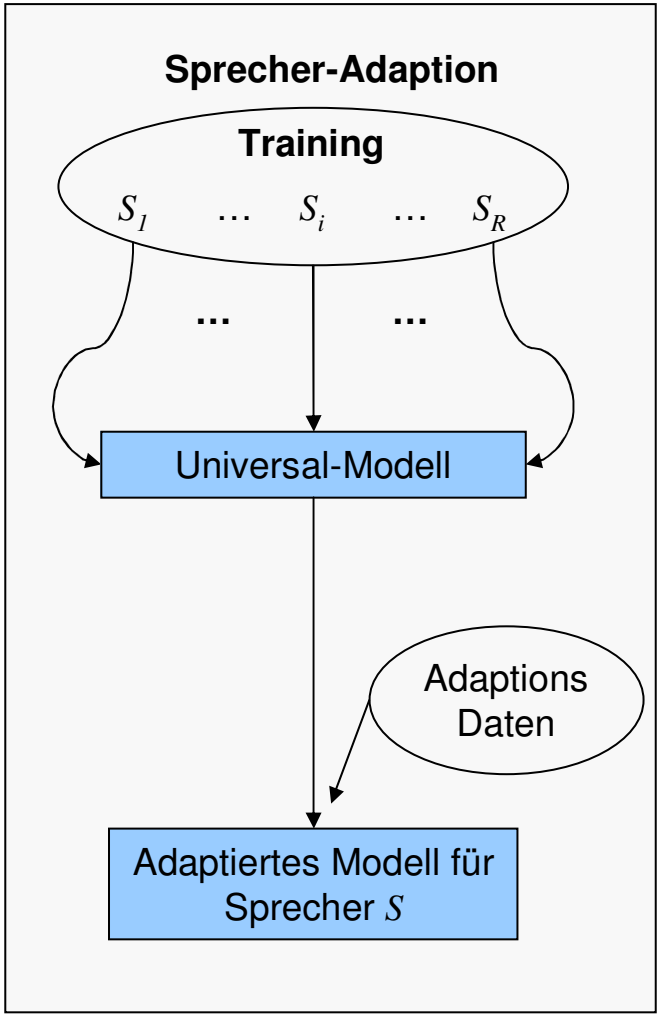
Gliederung

Speaker Adaptation using Eigenvoices

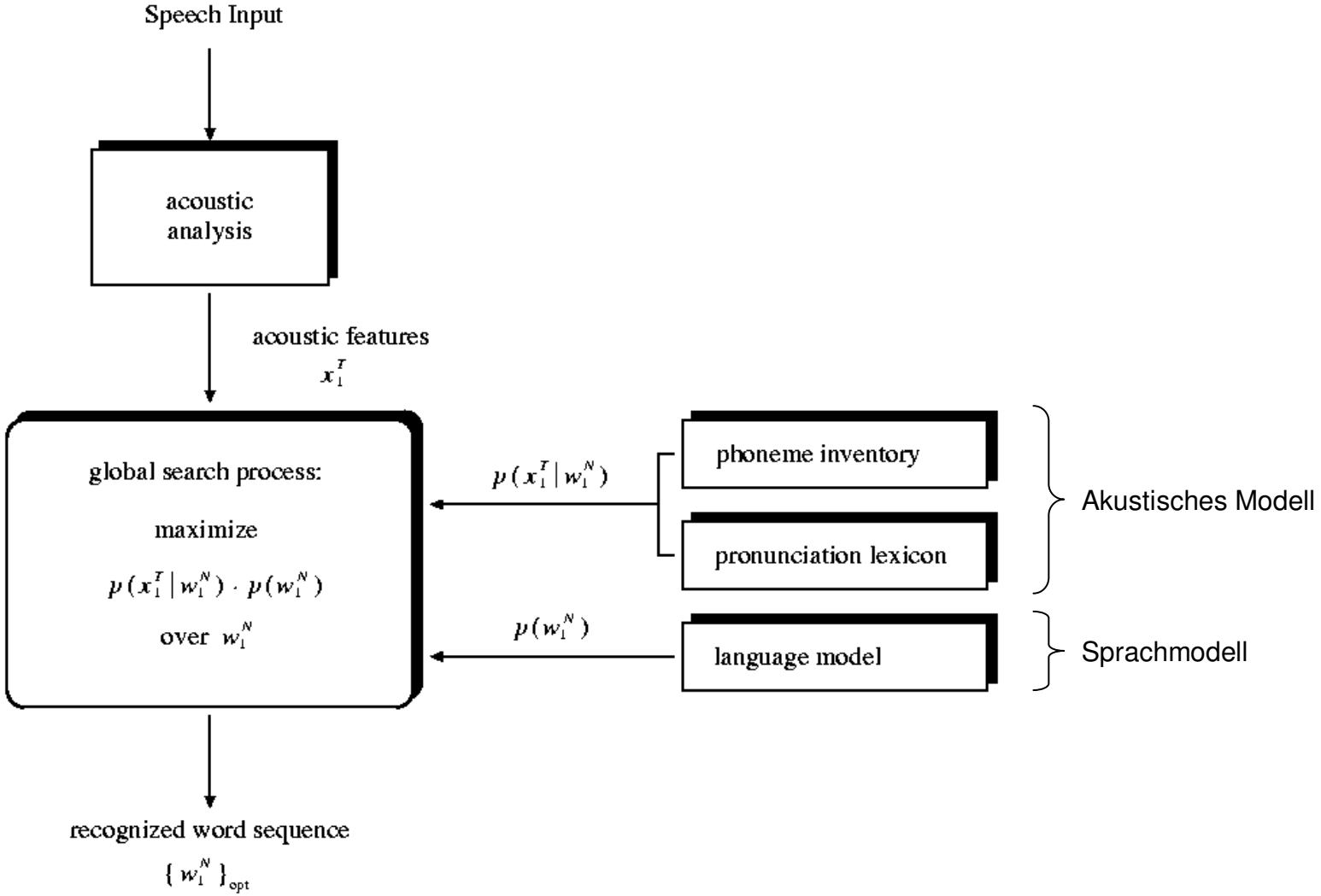
- Einführung und Motivation
- **Modell-basierte Sprecheradaption**
- Sprecheradaption mit Eigenvoices
- Ergebnisse
- Zusammenfassung

Überblick

- Prinzip der Spracherkennung
- Modellierung von Wörtern
  - Hidden Markov Modelle
- Modell-Parameter
  - Welche Werte beschreiben das Modell?
  - Adaption: Anpassen dieser Parameter an einen unbekannten Sprecher!



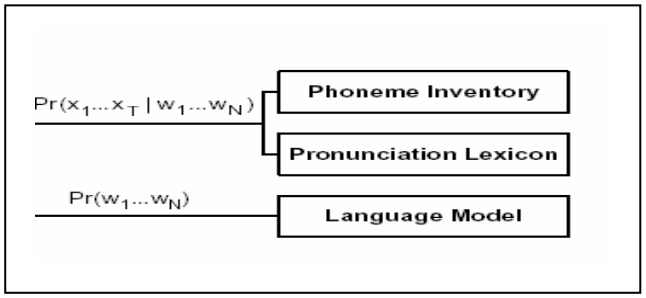
Prinzip der Spracherkennung – statistischer Ansatz



Ziel: Finde „optimale“ Wortfolge $w_1, \ldots, w_n$ (bzw. Wort w) zu Vektorfolge $x_1,\ldots x_T$

Benötigte Modelle

- Sprach-Modell:
  - $Pr(w_1,\ldots,w_N)$
  - Wahrscheinlichkeit für Wortfolge
  - „Musik hören“ ↔ „Musik essen“
  - sprecherunabhängig

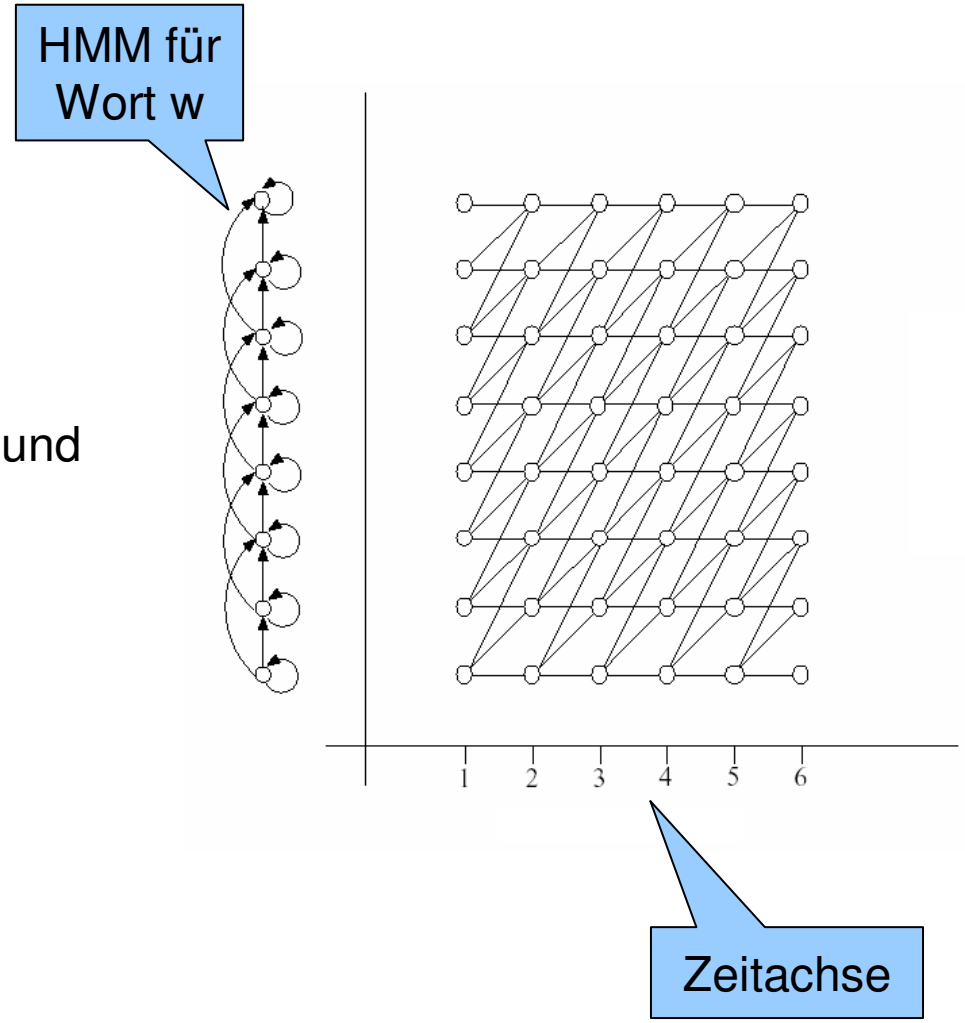


- Akustisches Modell:
  - $Pr(x_1,\ldots,x_T \mid w_1,\ldots,w_N)$
  - Wahrscheinlichkeit, dass die Wortfolge $w_1,\ldots,w_N$ dem akustischen Signal $x_1,\ldots,x_T$ entspricht
  - sprecherabhängig

Wie kann das akustische Modell beschrieben werden?

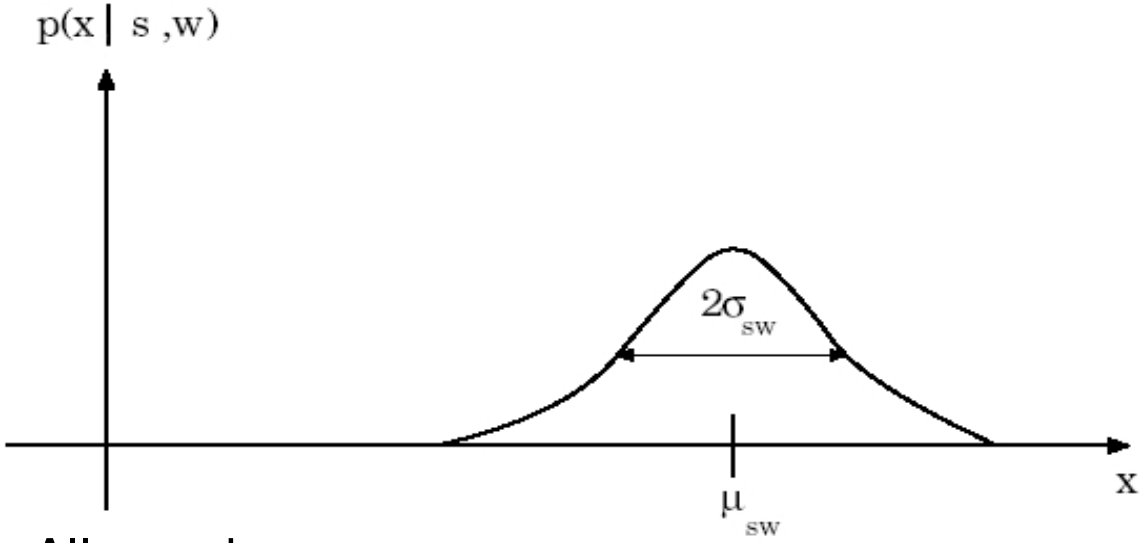
Das akustische Modell: Hidden Markov Modelle (HMMs)

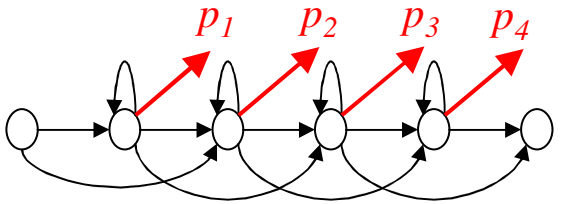
- Gesucht:
  Modell für $Pr(x_1 \cdots x_T | w)$
- Zuordnung: Wort – HMM
- HMM besteht aus Zuständen s und Übergängen (s,s‘)
- Bestimme Pfade der Länge T:
  $s_1, \ldots, s_T$
- Ordne jedem Pfad W‘keit zu:
  $p(x_1 \cdots x_T, s_1 \cdots s_T | w)$



HMMs: Emmisions-W‘keit

- W‘keit, dass Zustand s des Wortes w Output x produziert:
  - Emissions-W‘keit $p = p(x \mid s,w)$
- Annahme: p Gauss-verteilt





μ	Mittelwert
σ	Varianz

- Allgemein:

$$p(x \mid s,w) = \frac{1}{\sqrt{\det(2\pi \Sigma_{sw})}} e^{-\frac{1}{2}\left((x-\mu_{sw})^T \Sigma_{sw}^{-1}(x-\mu_{sw})\right)}$$

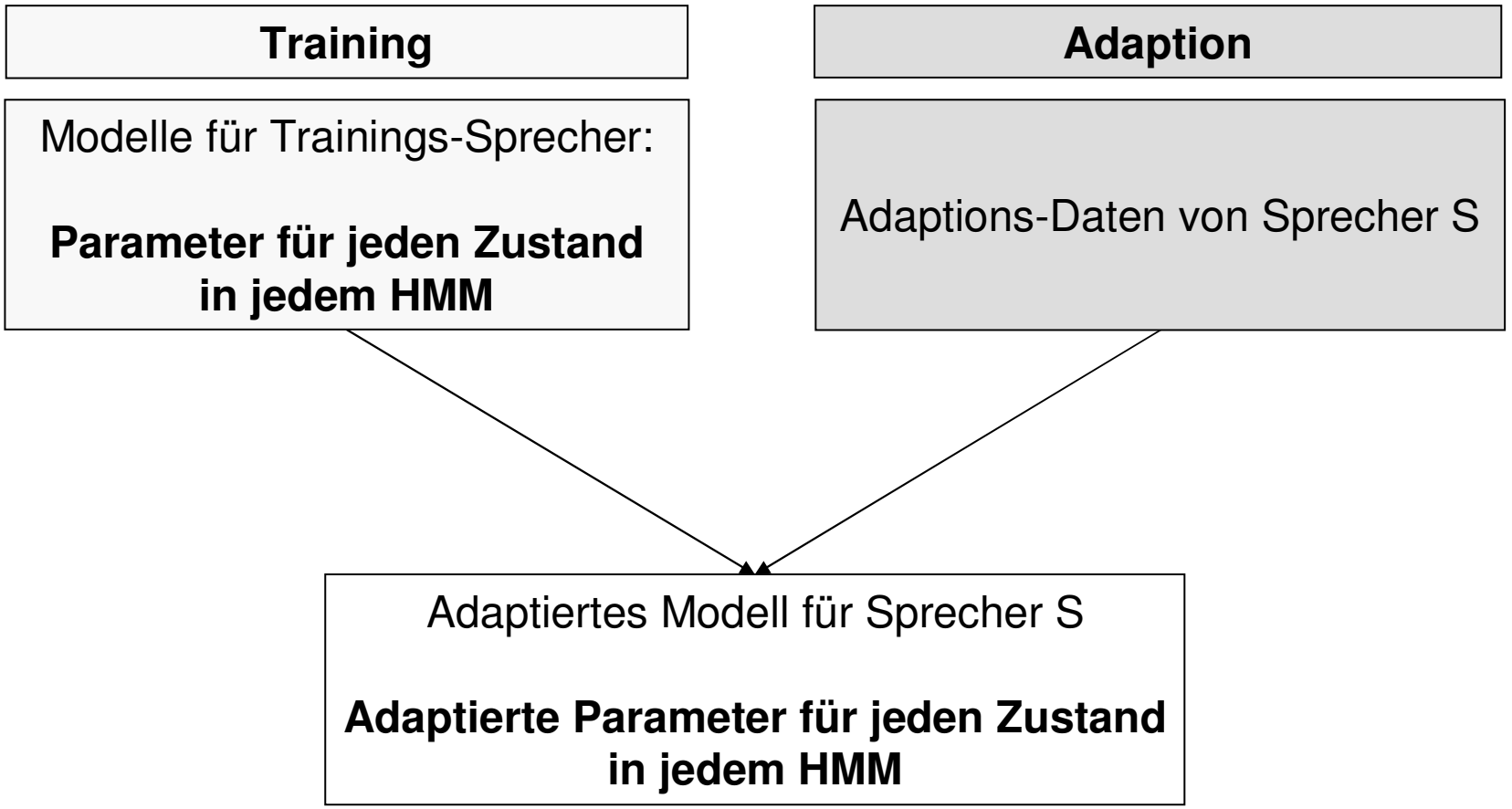
Benötigte Parameter

- Parameter, die das akustische Modell beschreiben:

μ_{sw}	Mittelwert
Σ_{sw}	Varianz

- Parameter werden für jeden Zustand in jedem Wort benötigt!
- Realität: Verteilungen besitzen mehrere Zentren
  - Berücksichtige Mischverteilungs-Komponente für jeden Zustand

Modell-basierte Sprecheradaption: Idee



Adaption der Modell-Parameter an den unbekannten Sprecher

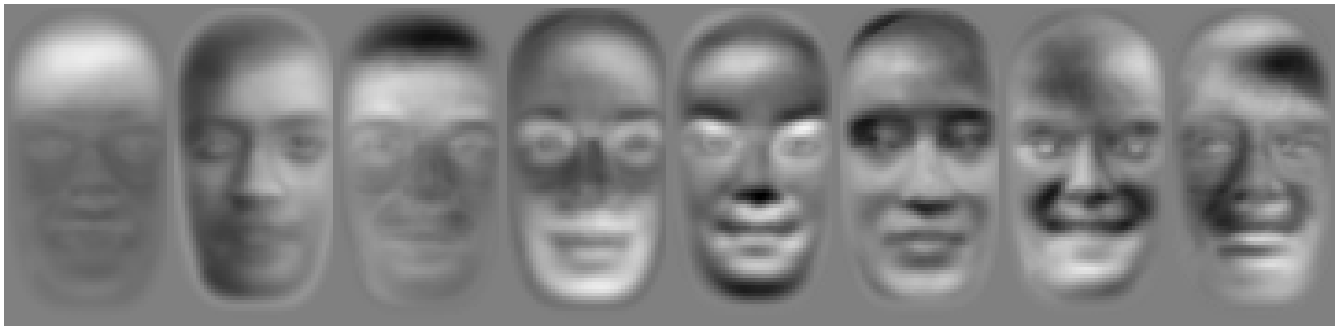
- Nur kleiner Teil aller Wörter in Adaptions-Daten
  - Viele Wörter kommen nicht vor
  - HMM-Parameter dieser Wörter können direkt nicht adaptiert werden

- Ansätze:
  - Speaker Clustering/Speaker Spaces (Eigenvoices)
  - Maximum a posteriori (MAP)
  - Lineare Transformationen (MLLR)

Ursprung der Eigenvoice-Technik: Eigenfaces

- Bilderkennung / Gesichtserkennung
- Matching Gesicht – Datenbank

- Dimension der Unterschiede zw. Gesichtern niedrig
  - Frisur, Kopfform, Augenpartie,…
- Modellierung:
  - **Eigenfaces** = mathematische Repräsentation für Bilder in der Datenbank
  - jedes Gesicht: Kombination dieser Eigenfaces



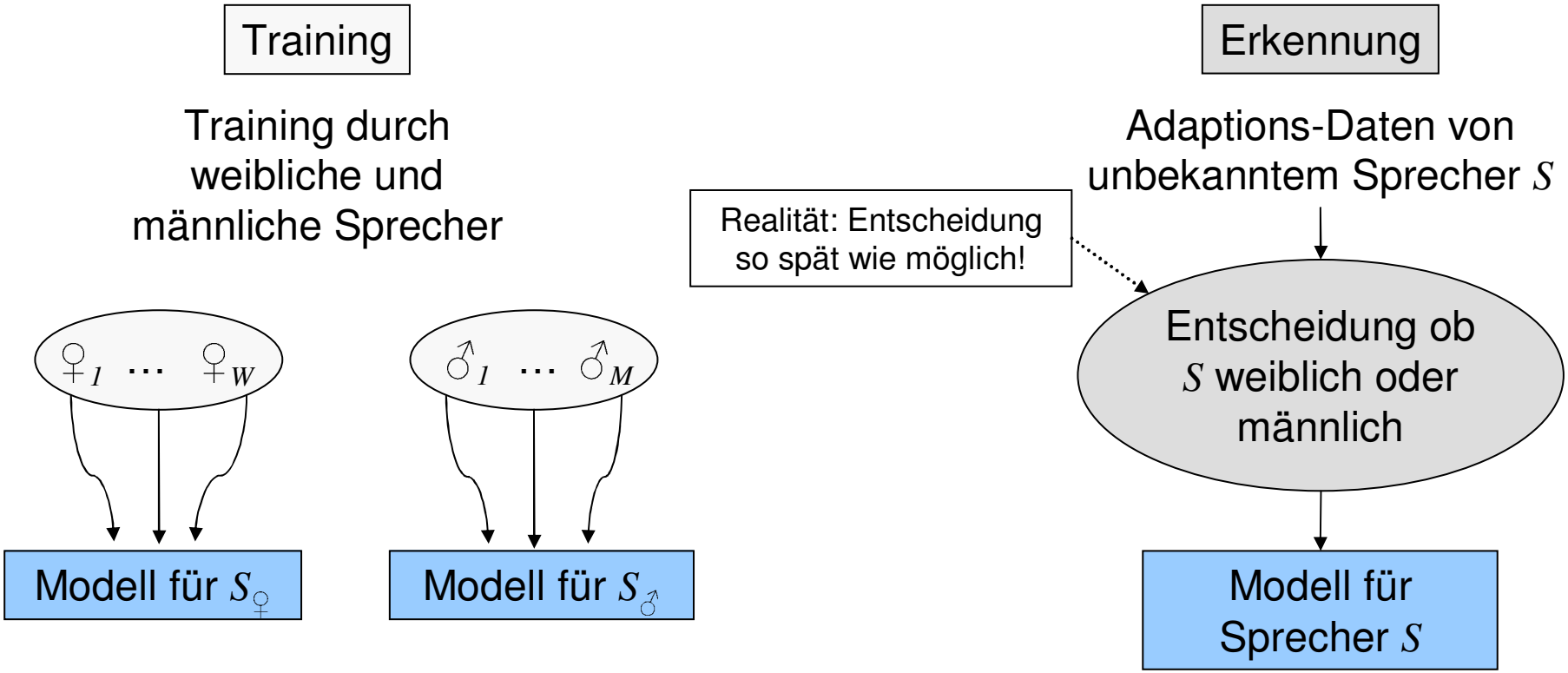
Gliederung

Speaker Adaptation using Eigenvoices

- Einführung und Motivation
- Modell-basierte Sprecheradaption
- **Sprecheradaption mit Eigenvoices**
- Ergebnisse
- Zusammenfassung

Zurordnung: Speaker Clustering

Einfaches, konstruiertes Beispiel:

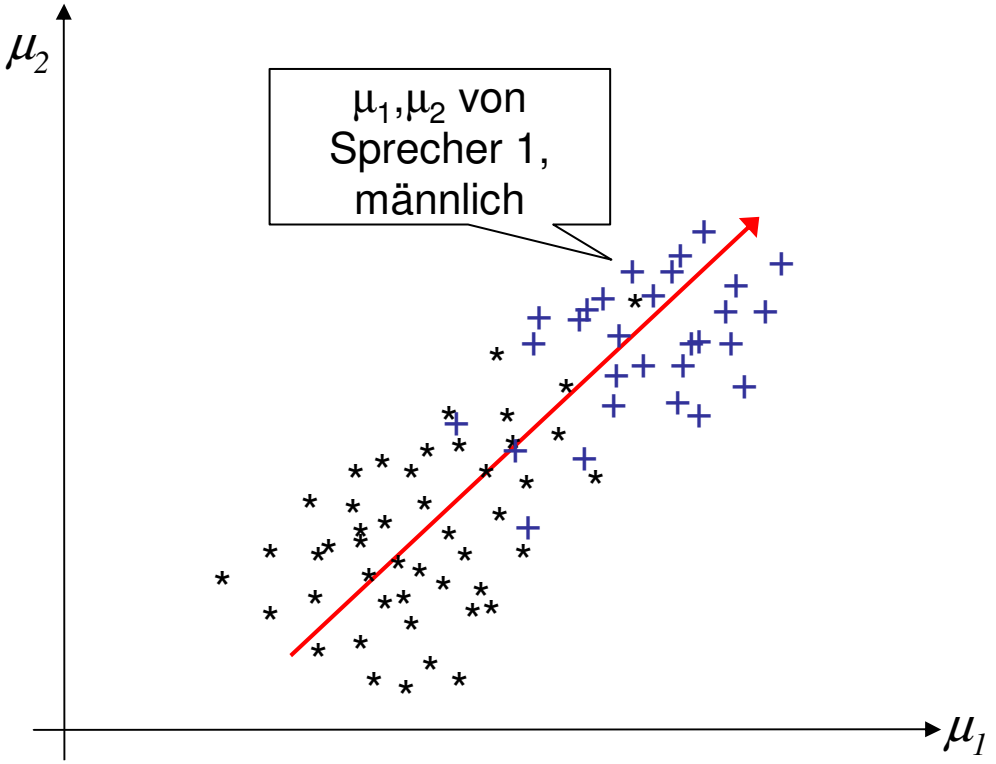


Speaker Clustering

- Erweiterung: mehr Cluster (Alter, Herkunft,…)
  - Problem: a priori Entscheidung über Cluster-Struktur
  - Trainingsdaten ungleichmässig verteilt
  - Manche Cluster erhalten wenig Trainigsdaten, Modelle ungenau
  - Unbekannter Sprecher passt nicht zu einem der Cluster
- Ansatz:
  - automatische Cluster-Bildung
  - neuer Sprecher = Kombination der Cluster
- Cluster Adaptive Training (Gales)
  - Iterative Bestimmung der Cluster während des Trainings
  - Grundlage für Adaption an neuen Sprecher
  - Modell des neuen Sprechers: gewichtete Summe der Cluster
  - Iterative Adaption (siehe Eigenvoices)
- Eigenvoices (Kuhn)

Eigenvoices: Idee

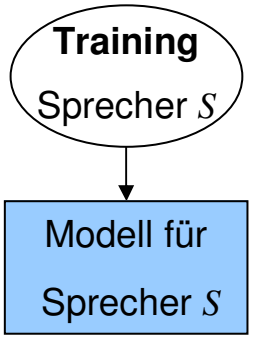
- Haupt-Streuungsrichtungen der Parameter



\+ Männlich

* Weiblich

Eigenvoices: Vorbereitungen



- Gesucht: Mathematische Repräsentation der trainierten Modelle
- Modelle bestimmt durch Parameter $(\mu, \Sigma, \textit{mixture weight})$
- Betrachte nur Mittelwerte μ

- Für jeden Sprecher wird ein **Supervektor** $\hat{\mu}_r$ konstruiert:
  - μ jedes Zustandes in jedem HMM in einen Vektor aufnehmen
  - Beispiel: N HMMs mit Zuständen $1,\ldots,S(w)$ für ein HMM w

$$\hat{\mu}_r = \begin{pmatrix} \mu_1^1(r) \\ \vdots \\ \mu_1^{S(1)}(r) \\ \vdots \\ \mu_N^1(r) \\ \vdots \\ \mu_N^{S(N)}(r) \end{pmatrix}$$

Mittelwert vom ersten Zustand des ersten HMM

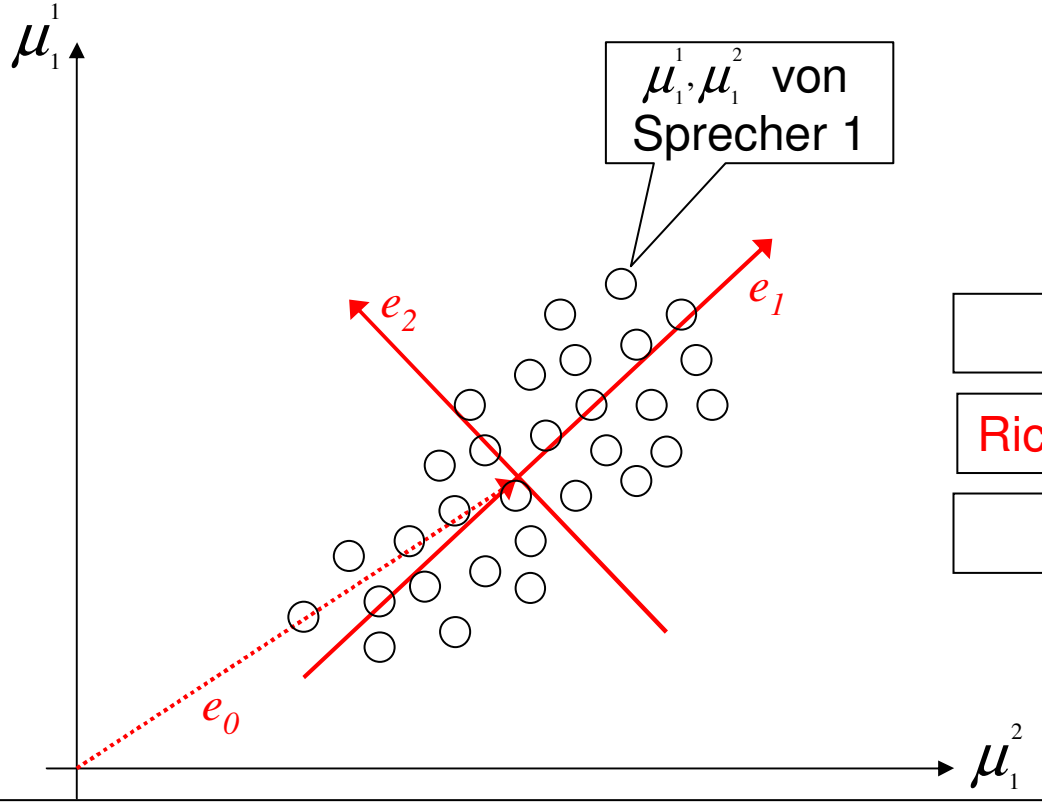
Mittelwert vom letzten Zustand des ersten HMM

Mittelwert vom ersten Zustand des letzten HMM

Mittelwert vom letzten Zustand des letzten HMM

Eigenvoices: Idee

- Anpassung des Parameter-Raumes an Streuung: **Hauptachsentransformation**
- Beispiel: ein Wort, Modell: ein HMM mit 2 Zuständen
  - 2 Parameter μ_1^1, μ_1^2
- 30 Trainings-Sprecher ⇒ 30 Supervektoren $\hat{\mu}_1 = \begin{pmatrix} \mu_1^1(1) \\ \mu_1^2(1) \end{pmatrix}, \ldots, \hat{\mu}_{30} = \begin{pmatrix} \mu_1^1(30) \\ \mu_1^2(30) \end{pmatrix}$



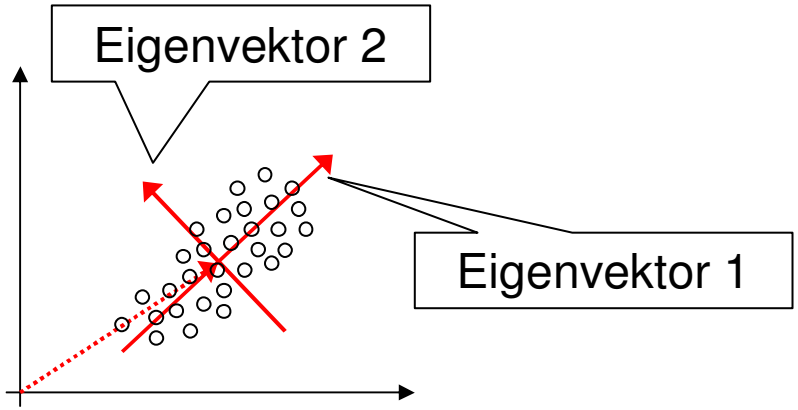
Mittelwert e_0

Richtung der grössten Streuung e_1

Richtung e_2 orthogonal e_1

Hauptachsentransformation: Bestimmung der Richtungen

- Betrachte Haupt-Streuungsrichtungen: **Eigenvektoren**



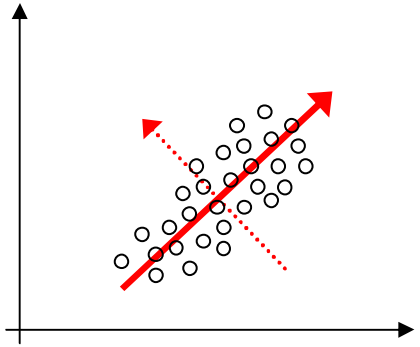
- Ansatz: Hauptachsentransformation
  - Anpassen des Koordinatensystems an die Streuung der Parameter
  - Rotation + Translation
  - Liefert Orthogonal-Basis des rotierten Raums: die Eigenvektoren
  - Jedes Modell: Linearkombination der Eigenvektoren

Sprecheradaption mit Eigenvoices

- Betrachtet: nur Parameter von 2 HMM-Zuständen

- Realität: Parameter-Raum ist hoch-dimensional
  - Isolet-Task: 26 HMMs, je 6 Zustände = 156 Mittelwerte pro Sprecher
  - Jeder Mittelwert hat hohe Dimension!
  - Grosses Vokabular: Supervektor hat Dimension ~1 Mio!

- Ziel: Dimensions-Reduktion
  - Richtungen mit geringer Streuung weniger interessant
  - Suche Basis eines Raums mit kleinerer Dimension
  - Principal Component Analysis (PCA)



Bestimmung der Eigenvoices: Dimensions-Reduktion

- PCA mit Covarianz-Matrix der R Supervektoren von R Trainings-Sprechern
  - Covarianz-Matrix:

$$\frac{1}{R}\sum_{r=1}^{R}(\hat{\mu}_r - \hat{\mu})(\hat{\mu}_r - \hat{\mu})^T$$

$\hat{\mu}_r$ Super-Vektor von Sprecher r (hohe Dimension D!) ⇒ Matrix D x D

$\hat{\mu}$ Mittelwert (z.B. aus sprecher-unabhängigem Modell)

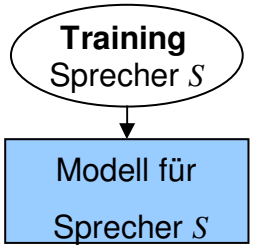
  - Bestimme Eigenwerte e_i, $i=1,\ldots,R$ (nur $R \cdot D$ Werte ⇒ Rang R)
  - e_k ist der zum k-grössten Eigenwert gehörende Eigenvektor

- Ergebnis: R hochdimensionale Eigenvektoren
  - Orthogonal-Basis des rotierten Raumes
  - Vektoren geordnet nach Grösse der Eigenwerte

- Wenige Eigenvektoren decken grossen Teil der Streuung ab!
- Dimensions-Reduktion: Behalte nur die ersten K Eigenvektoren
  - Diese *K* **EIGENVOICES** spannen den **SPEAKER SPACE** auf

Bestimmung der Eigenvoices

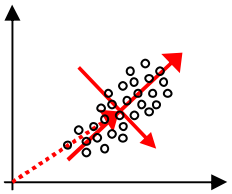
Training von sprecher-abhängigen Modellen mit R Trainings-Sprechern



Konstruktion von R Supervektoren aus den Modellen, jeder von hoher Dim. D

$$\hat{\mu}_1 = \begin{pmatrix} \mu_1^1(1) \\ \mu_1^2(1) \end{pmatrix}, \ldots, \hat{\mu}_{30} = \begin{pmatrix} \mu_1^1(30) \\ \mu_1^2(30) \end{pmatrix}$$

Hauptachsentransformation liefert R Eigenvektoren: $e_1, \ldots, e_R$



Behalte nur die ersten K Eigenvektoren $e_1, \ldots, e_K$

EIGENVOICES

Adaption an den unbekannten Sprecher: Idee

- „Der Speaker Space enthält alle möglichen sprecher-abh. Modelle“

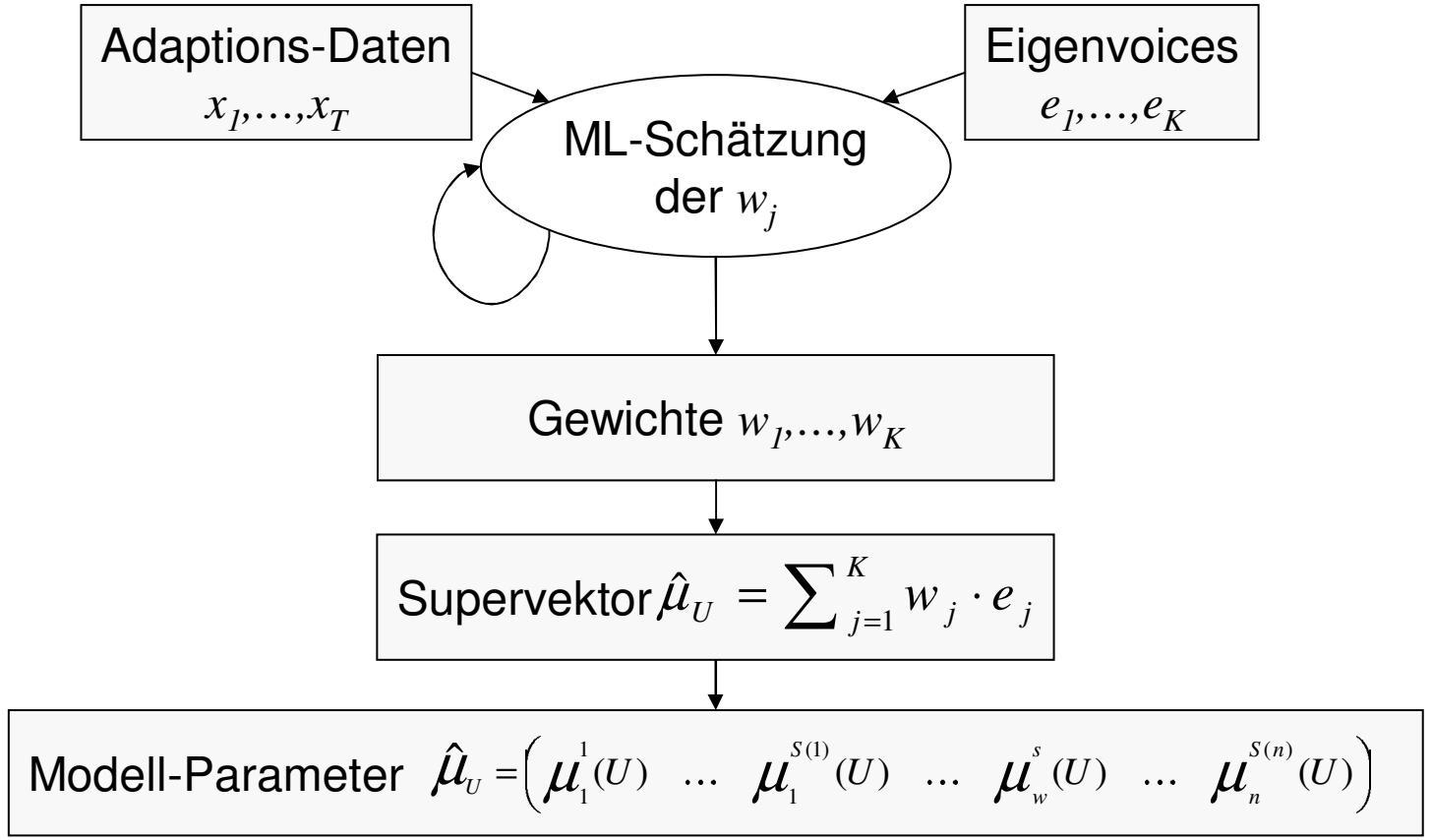
- Erinnerung: Modell = Supervektor
- Jeder Supervektor: Linearkobination der Basisvektoren
- Auch das noch unbekannte Modell!

- Supervektor für unbekannten Sprecher U: $\hat{\mu}_U = \sum_{j=1}^{K} w_j \cdot e_j$

- Ziel:
  - finde Supervektor $\hat{\mu}_U$, der den unbekannten Sprecher repräsentiert
  - **Gegeben:** K Eigenvoices aus PCA, Adaptions-Daten $x_1,\ldots,x_T$
  - **Aufgabe:** Bestimmung der Gewichte w_j

Bestimmen der Gewichte w_j für den unbekannten Sprecher

- Maximum Likelihood Eigen-Decomposition (**MLED**)
  - Maximum Likelihood Schätzung der Gewichte w_j
- Iteratives Bestimmen der Gewichte mit Hilfe der Adaptions-Daten



ML-Schätzung

Schätzen der Gewichte w_j

- Ziel: Bestimme Parameter $\lambda=\{w_1,\ldots,w_K\}$ mit maximalem Likelihood:

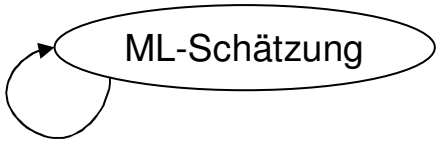
$$\arg\max_{\lambda} \prod_{t=1}^{T} p(x_t \mid \lambda) \quad \text{bzw.} \quad \arg\max_{\lambda} \sum_{t=1}^{T} \log p(x_t \mid \lambda)$$

- Es existiert keine geschlossene Lösung zur Maximierung
  - Zuordnung zu Mischverteilung nicht bekannt ⇒ „versteckte Variable“ y
  - Iterative Lösung: EM-Algorithmus

- Hilfsfunktion Q:

$\tilde{\lambda}$ Aktuelle Zielmenge, λ Menge von Gewichten aus letzter Schätzung

$$Q(\lambda,\tilde{\lambda}) = \sum_{t=1}^{T} \sum_{y} p(y \mid x_t,\lambda) \cdot \underbrace{\log p(x_t, y \mid \tilde{\lambda})}_{\text{Log-Likelihood}}$$



Schätzen der Gewichte w_j

- Q-Funktion für ML-Schätzung:

aus der letzten Iteration

Der gesuchte Wert!

$$Q(\lambda,\tilde{\lambda}) = -\frac{1}{2}\, p(x_1 \ldots x_T \mid \lambda) \cdot \sum_s \sum_l \sum_t \left(\gamma_{l,s}(t) \cdot \left[-n\log(2\pi) - \log\left|\Sigma_{l,s}\right| + (x_t - \tilde{\mu}_{l,s})^T \Sigma_{l,s}^{-1} (x_t - \tilde{\mu}_{l,s})\right]\right)$$

$\gamma_{l,s}(t)$: W'keit, dass Beobachtung x_t Komponente l von Zustand s zugewiesen wird

n: Anzahl der Merkmale

$\Sigma_{l,s}$: Covarianzmatrix für Komponente l, Zustand s

$\tilde{\mu}_{l,s}$: Adaptierter Mittelwert für Komponente l, Zustand s

- Einsetzen: $\tilde{\mu}_{l,s} = \sum_{j=1}^{K} w_j \cdot e_{j,l,s}$

mit $e_{j,l,s}$ Subvektor der j-ten Eigenvoice zu Komp. l, Zustand s

- Maximierung: $\frac{\partial Q}{\partial w(j)} = 0$ für $j = 1,\ldots,K \Leftrightarrow$ K Gleichungen, K Variablen (w_j)

Gliederung

Speaker Adaptation using Eigenvoices

- Einführung und Motivation
- Modell-basierte Sprecheradaption
- Sprecheradaption mit Eigenvoices
- **Ergebnisse**
- Zusammenfassung

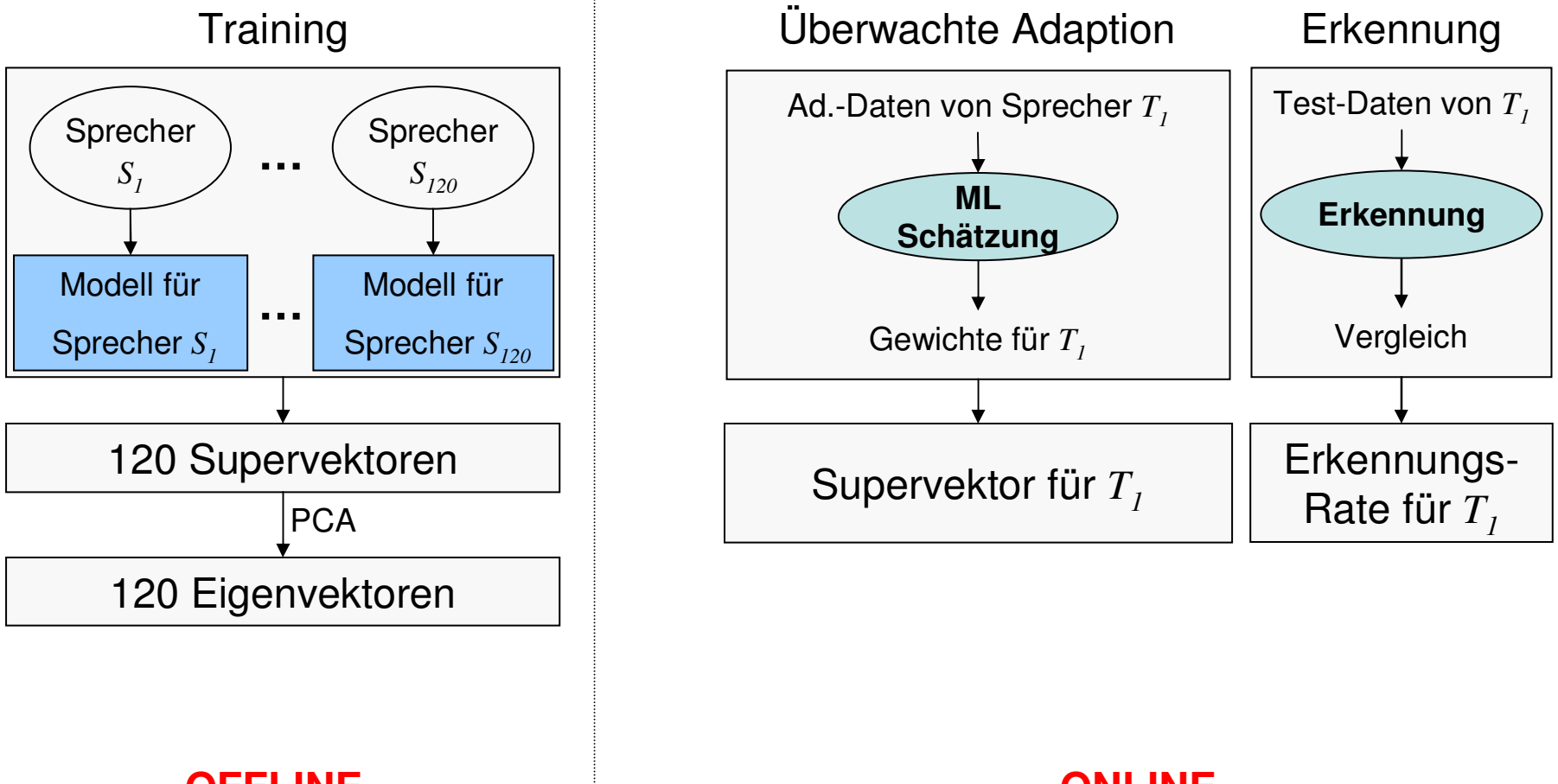
Ergebnisse 1: Kleines Vokabular, isoliert (R.Kuhn, 2000)

- Isolet: Buchstaben des Alphabets, isoliert
- von 150 Sprechern zweimal gesprochen

- Modell für einen Sprecher:
  - 26 HMMs, je 6 Zustände, jeder Zustand mit 18-dim. Mittelwert μ
  - Dimension der Supervektoren: $26 \cdot 6 \cdot 18 = 2808$

- Training mit 120 Sprechern
- Erkennung mit den restlichen 30 Sprechern
  - Variable Anzahl Buchstaben als Adaptions-Daten aus 1. Alphabet
  - Test-Daten aus 2. Alphabet

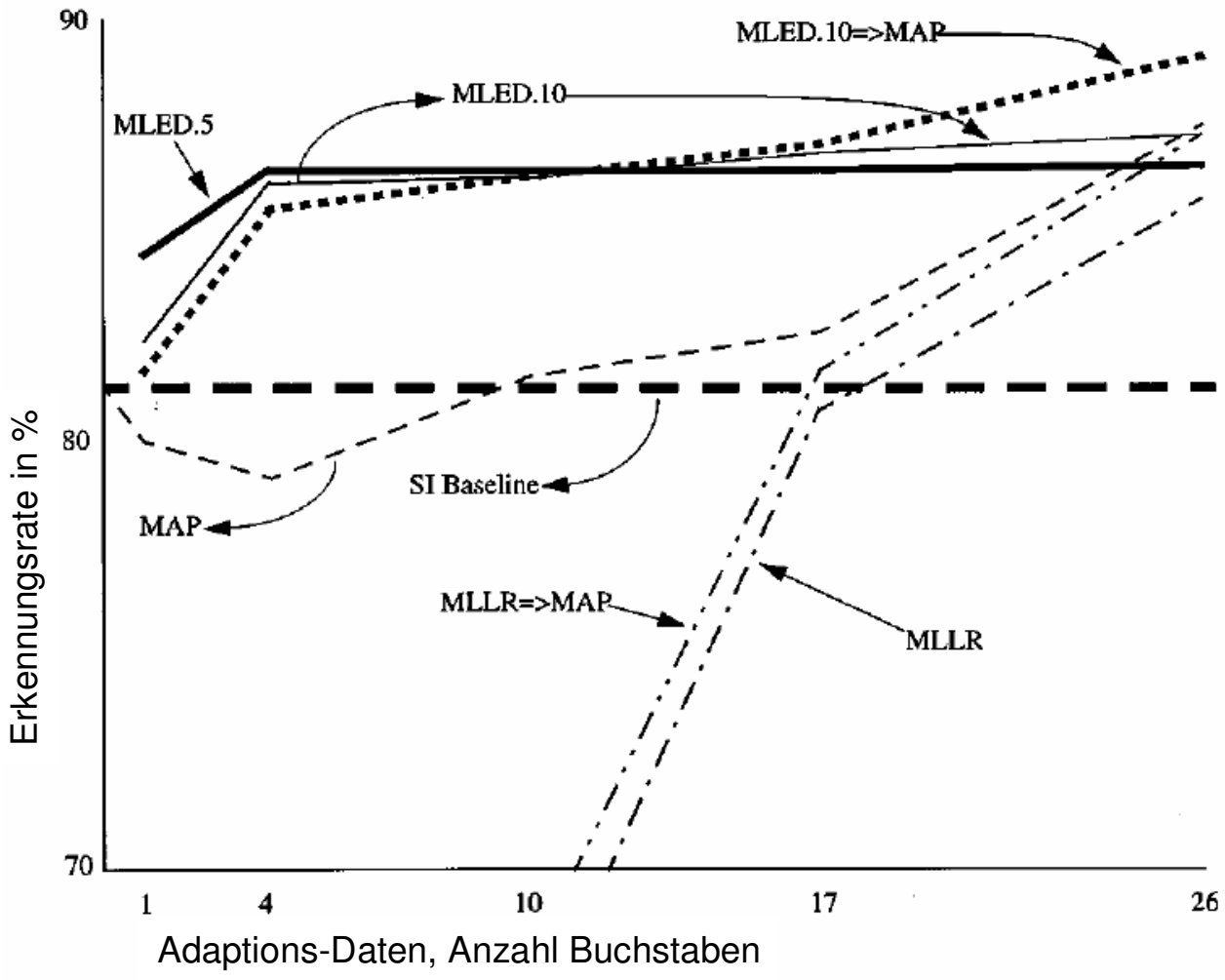
Isolet – Erkennung von isolierten Buchstaben



Isolet – Ergebnisse, überwachte Adaption

Adaptions-Technik	Erkennungsrate in % bei versch. Adaptions-Daten In Klammern: Rel. Änd. in % gg. sprecher-unabh.		
	Ø 1 Buchstabe	{a,b,c,u}	Alphabet
Ohne Adaption, sprecher-unabh.	81,3	81,3	81,3
Eigenvoices, K=5	84,4 (+3,8)	86,3 (+6,2)	86,5 (+6,4)
Eigenvoices, K=10	82,4 (+1,6)	86,4 (+6,3)	87,4 (+7,5)
MAP	80,0 (-1,6)	78,6 (-3,3)	87,4 (+7,5)
MLLR	3,8 (-95,3)	17,0 (-80,1)	85,8 (+5,5)

Ergebnisse



SI Baseline	ohne Adaption
MLED.5	K=5 Eigenvoices
MLED.10	K=10
MAP	Grundlage: Sprecher-unabh. Modell
MLLR	Grundlage: Sprecher-unabh. Modell
MLLR =>MAP	MLLR als Grundlage für MAP
MLED.10 =>MAP	MLED mit K=10 als Grundlage für MLLR

Ergebnisse 2: Grosses Vokabular, kontinuierlich (H. Botterweck, 2000)

- Wall Street Journal (WSJ)
- Problem: Training der R sprecher-abhängigen Modelle
  - Ca. 1 Mio Mittelwerte pro Supervektor
  - Nicht genug Trainingsdaten
- Abhilfe:
  - Zuerst Training eines sprecher-unabhängigen Modells
  - Dann MAP/MLLR-Adaption mit den Trainings-Sprechern -> R Modelle
- PCA / Adaption an unbekannte Sprecher analog

Adaptions-Technik	Erkennungsrate in %		
	3 Sek.	12 Sek.	64 Sek.
Ohne Adaption	79,8	79,8	79,8
MLED, K=100	82,5 (+3,4)	83,8 (+5,0)	84,0 (+5,3)

Adaptions-Technik	Erkennungsrate in % 3 Sek. Adaption
Ohne Adaption	79,8
MLED, K=1	82,3 (+3,1)
MLED, K=10	82,8 (+3,8)
MLED, K=20	82,5 (+3,4)
MLED, K=100	82,5 (+3,4)

Gliederung

Speaker Adaptation using Eigenvoices

- Einführung und Motivation
- Modell-basierte Sprecheradaption
- Sprecheradaption mit Eigenvoices
- Ergebnisse
- **Zusammenfassung**

Zusammenfassung

- Modell-basierte Sprecher-Adaption
  - Adaptionsdaten + trainierte Modelle ⇒ Adaptiertes Modell
- Eigenvoices
  - Streuung der Gauss‘schen Parameter zw. den Sprechern
  - Principal Component Analysis, Dimensionsreduktion
- Unbekannter Sprecher
  - Linearkombination der Eigenvoices
  - Iterative Bestimmung der Gewichte: ML-Schätzung, EM-Algorithmus
- Anwendung:
  - schnelle Adaption
  - Mehr Adaptions-Daten: Kombination mit anderen Methoden