Hochschule Niederrhein

Niederrhein University of Applied Sciences

Elektrotechnik u. Informatik

Automatische Spracherkennung in Theorie und Praxis

Hans-Günter Hirsch, Harald Finster

Hochschule Niederrhein, Krefeld

email: hans-guenter.hirsch@hs-niederrhein.de

http://dnt.kr.hs-niederrhein.de

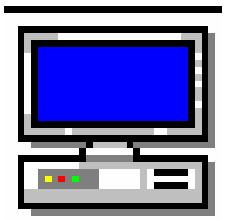
Gliederung

1. **Aufbau eines Spracherkennungssystems**
   - Sprachanalyse, Merkmalsextraktion
   - Training
   - Mustererkennung
2. **Experimente zur phonembasierten Erkennung**
   - Training
   - Erkennungsergebnisse
3. **Sprachdialogsystem**
   - Demonstrationen

Mensch-Maschine Kommunikation

Kommunikation
Mensch – Maschine

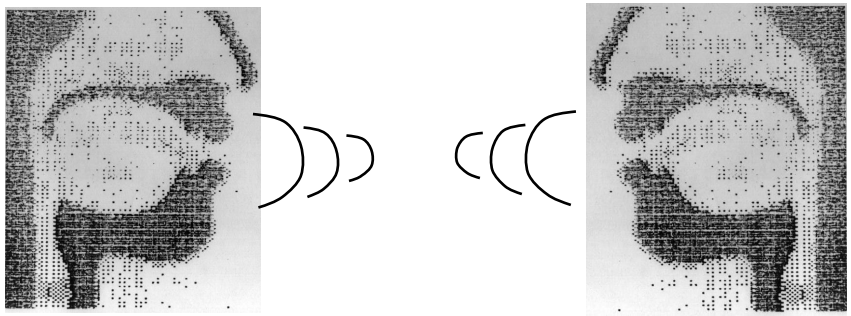
Eingabe:
- Tastatur
- Maus



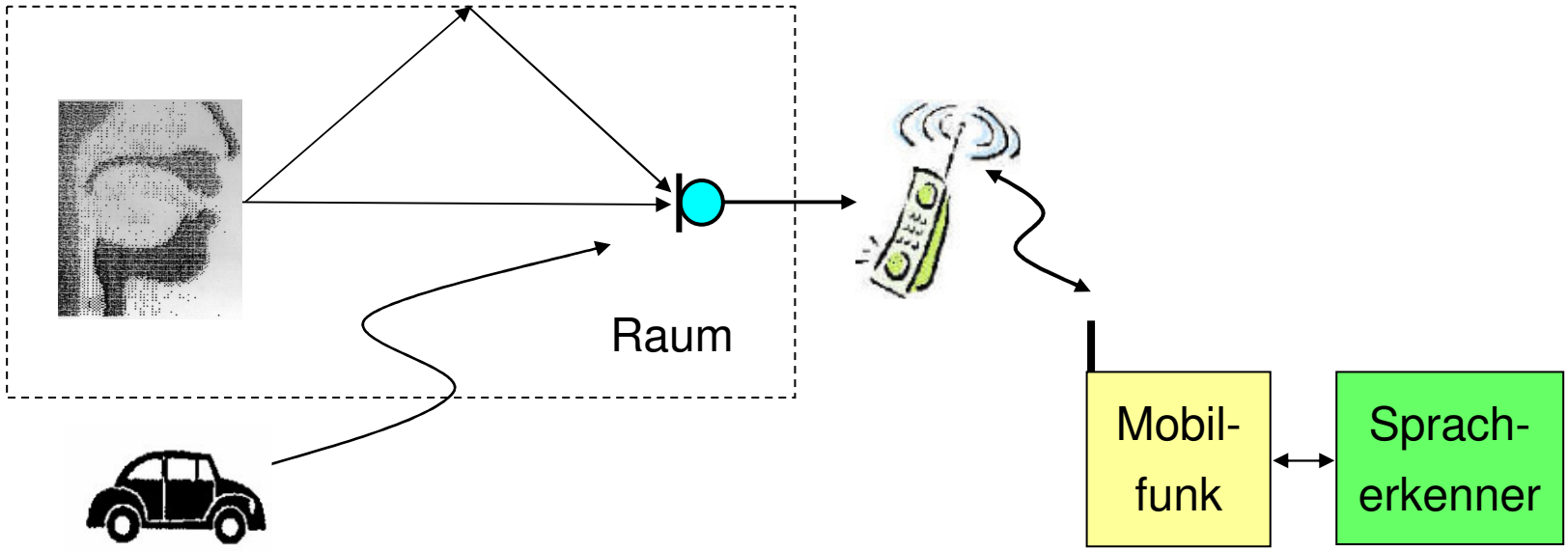
Ausgabe:
- Display (visuell)

Kommunikation
Mensch – Mensch

- Sprache
- Gestik



Störeinflüsse

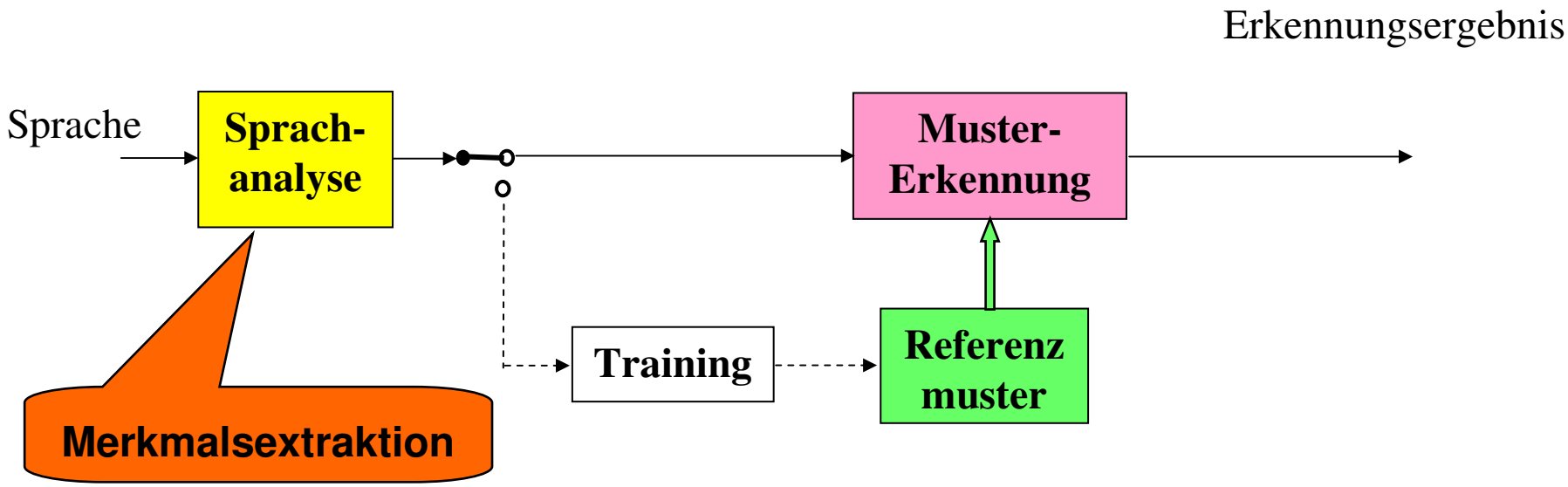


- Additive Hintergrundstörungen
- Unbekannte Frequenzgänge (z.B. Mikrofon, Telefonkanal)
- Hallige räumliche Umgebung
- Mobilfunkkanal (Sprachcodierung, Funkkanal)

Klassifizierung

1. Erkennungsaufgabe
   - Einzelne Wörter
   - Wortketten
   - Ganze Sätze, kontinuierliche Sprache
2. Personenkreis
   - sprecherabhängig (individuelle Trainingsphase)
   - sprecher**un**abhängig
3. Anwendung
   - Steuerung von Geräten
   - Informationssystem
   - Diktiersystem

Aufbau eines Erkennungssystems

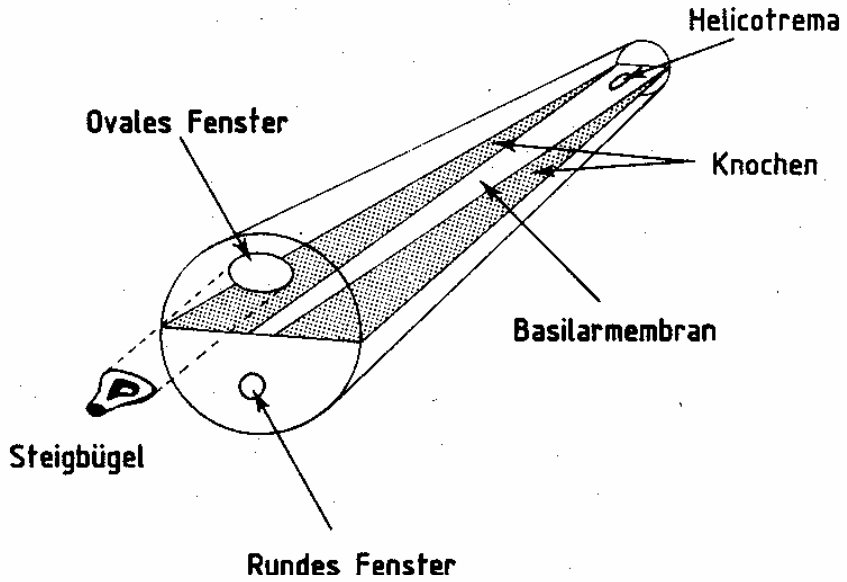


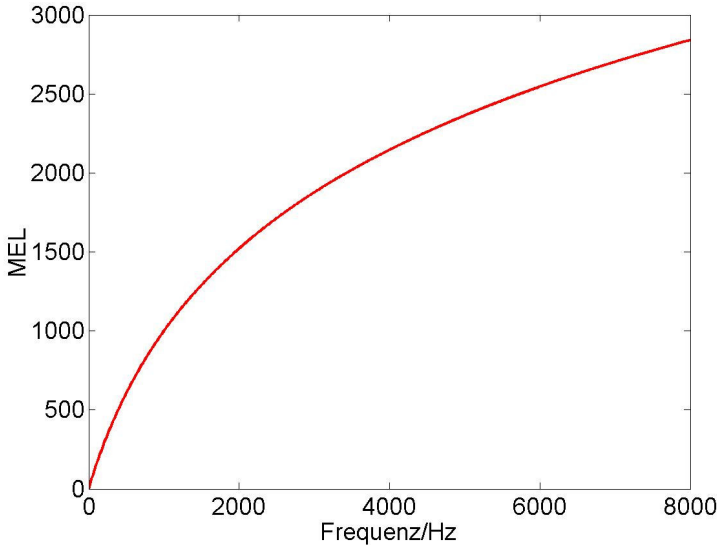
Sprachanalyse

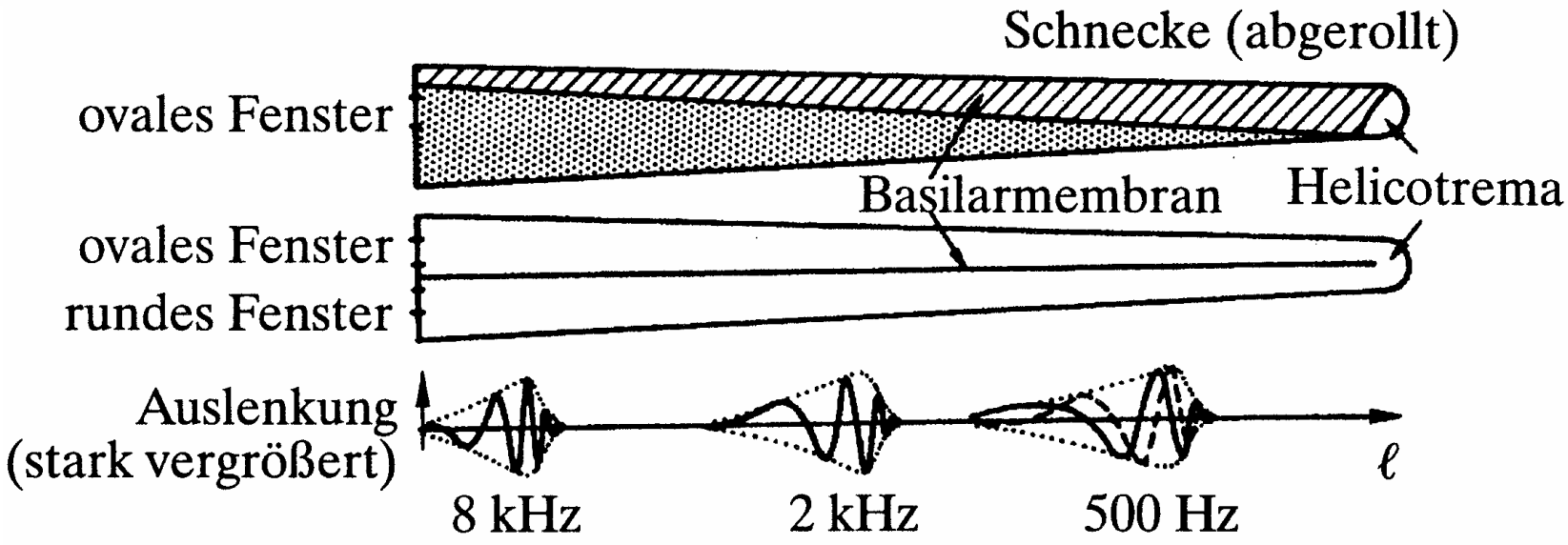
- Der Signalverlauf im Zeitbereich beinhaltet nur sehr begrenzt relevante akustische Merkmale
- Frequenzbereich (ca. bis 6 kHz, Telefon bis 4 kHz)

- Menschliche Gehör:
  - Frequenzanalyse (Frequenz-Orts Transformation im Innenohr) ➔ Filterbank
  - Phase leistet nahezu keinen Beitrag zum Sprachverstehen ➔ Betragsbildung
  - Keine lineare Abhängigkeit zwischen dem Schallpegel und der empfundenen Lautheit ➔ Logarithmischer Zusammenhang

Innenohr mit Basilarmembran

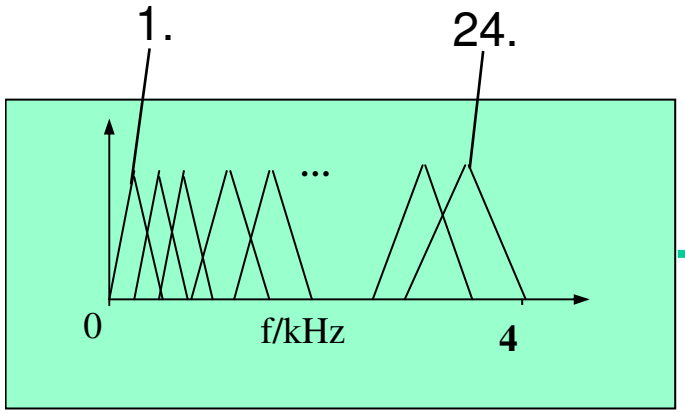




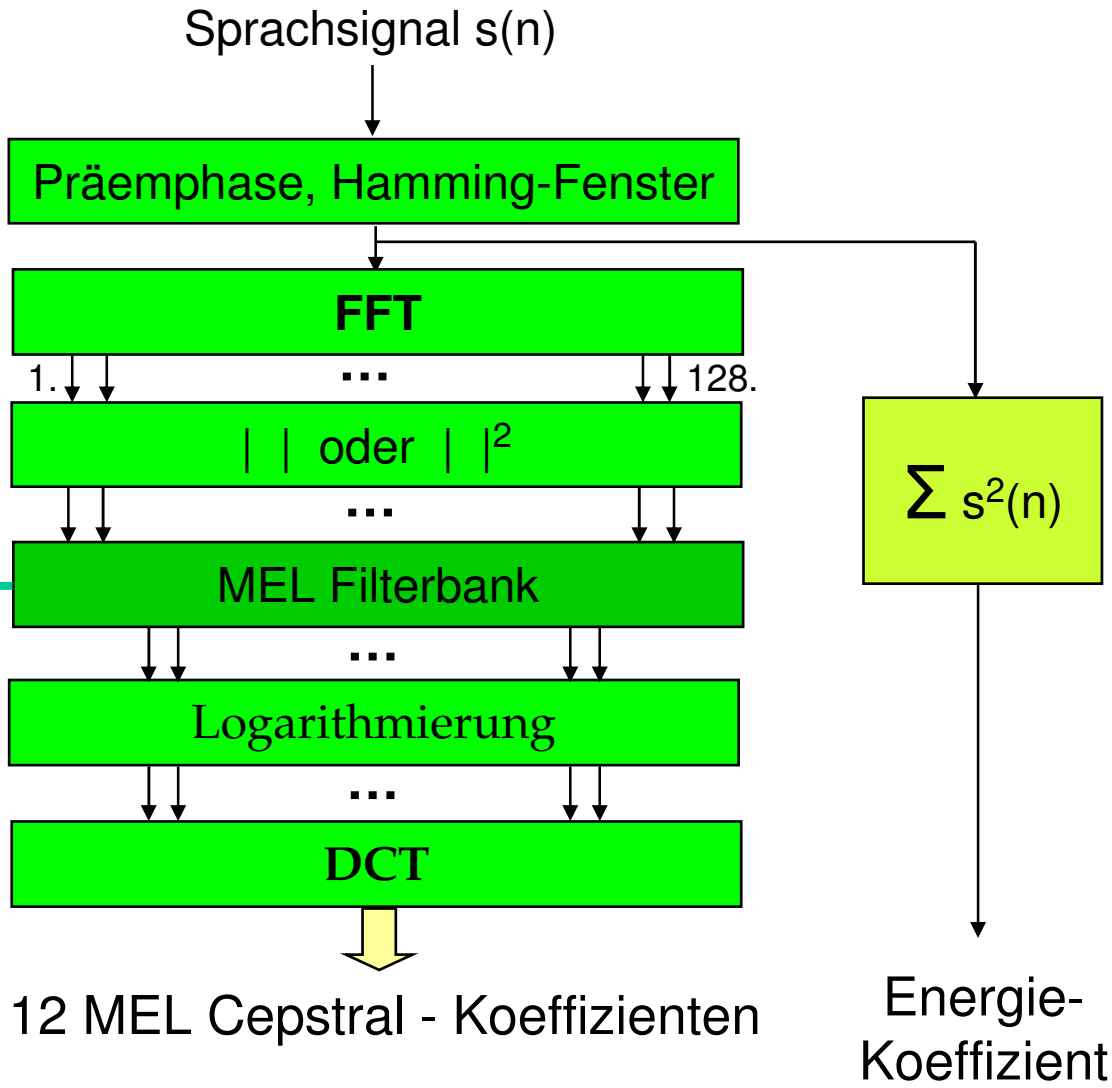


„Cepstral“ Analyse

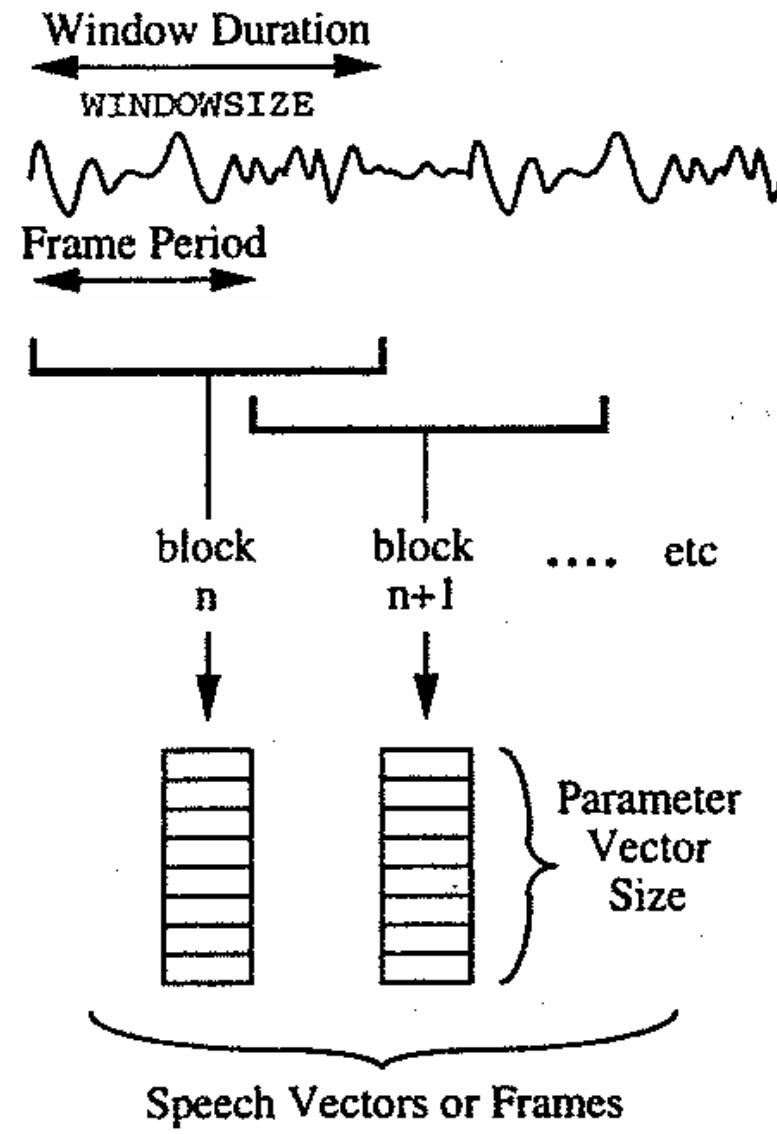
- Kurzzeit-Spektralanalyse von ca. 20 -30 ms langen Signalabschnitten



DCT = Diskrete Cosinus Transformation



Merkmalsextraktion

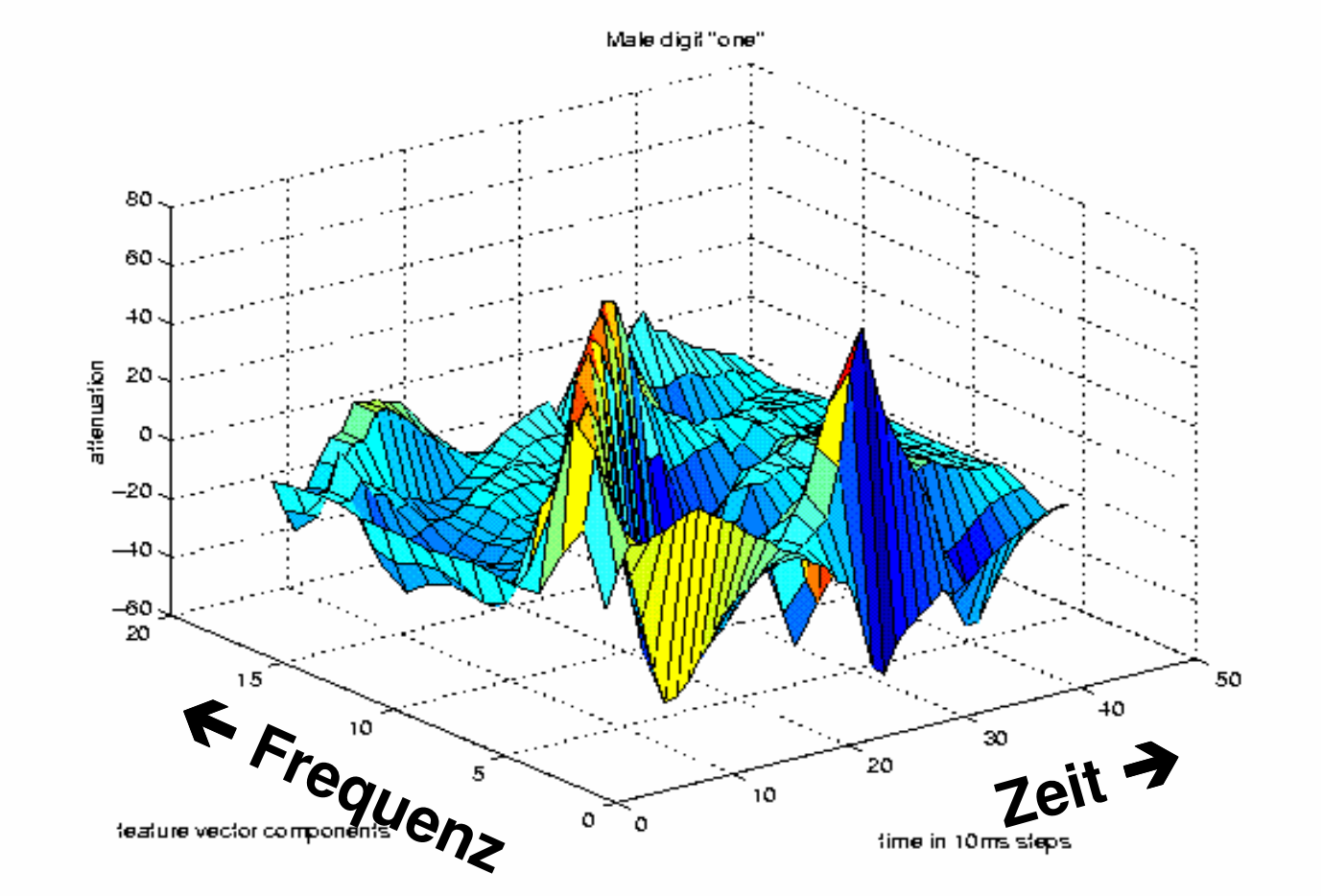


- Fensterbreite: ~ 20 … 30 ms
- Verschiebung um ~ 10 ms
  ➔ 100 Vektoren/Sekunde

- Jeder Vektor besteht aus ~13 akustischen Parameter (12 Cepstral- und 1 Energieparameter)
- Neben den statischen Parametern werden die zeitlichen Ableitungen jedes Parameters über der Zeit verwendet➔ Delta-Parameter
- 2. Ableitung: Delta-Delta Parameter

➔ 100 * (13+13+13) = 3900 Parameter/Sek.

Spektrales Muster der Ziffer ‘one’

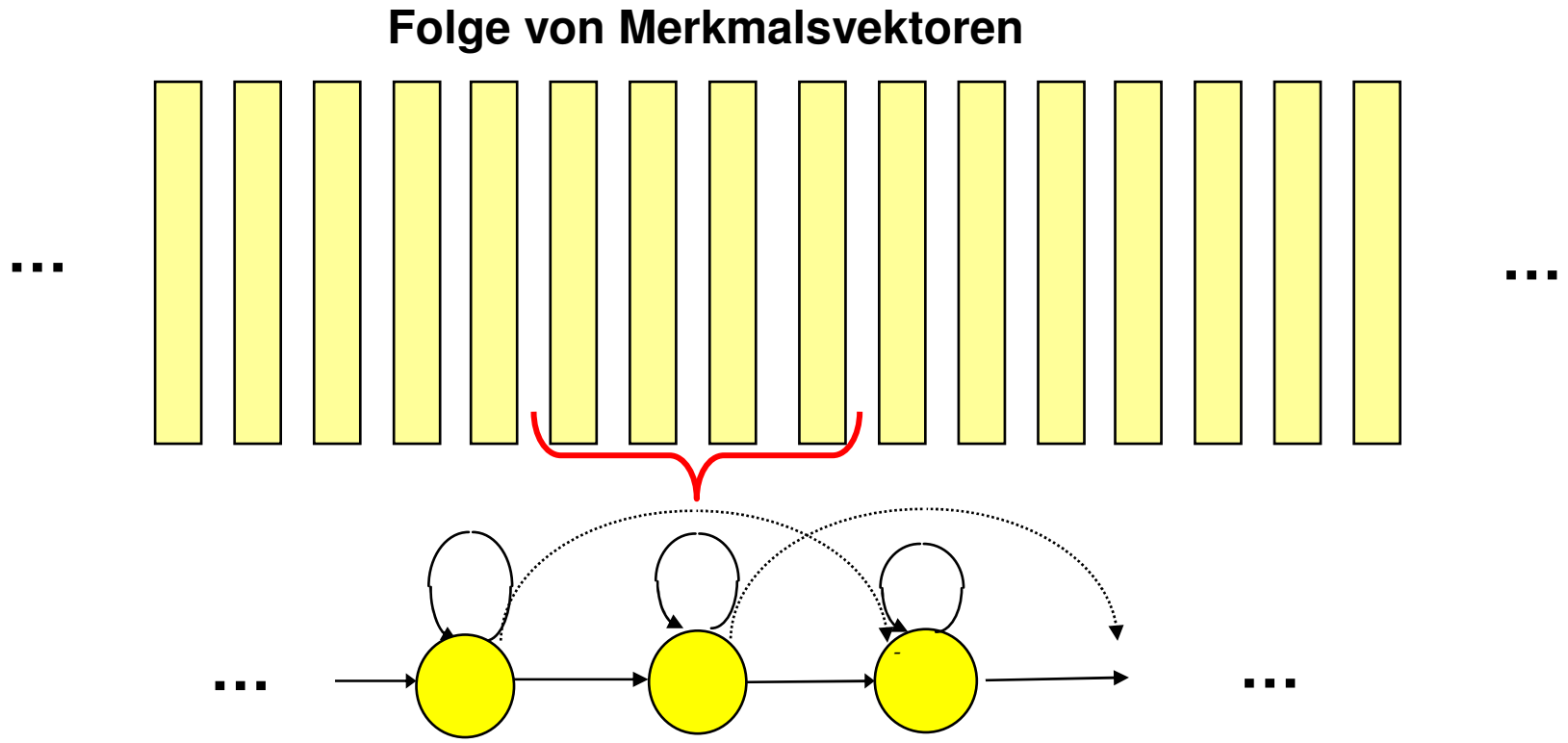


Training

- Aus der Merkmalsextraktion resultiert ein Datenstrom mit ~3900 Parameter je Sekunde
- Die Parameter des zeitlichen Abschnitts eines Wort (Lauts) können als Referenzmuster verwendet werden
- Zur Datenreduktion Auswahl oder Mittelung von Vektoren ➔ Anwendung zur sprecherabhängigen Erkennung

- Zur sprecher**un**abhängigen Erkennung wird eine statistische Modellierung mit sogenannten Hidden-Markov Modellen (HMMs) eingesetzt.

- Mathematisches, statistisches Modell zur Beschreibung von Sprachabschnitten



- Jeder Zustand wird beschrieben durch Gauß-Verteilungen jedes Parameters eines Merkmalsvektors (z.B. 39 Mittelwerte & 39 Varianzen)

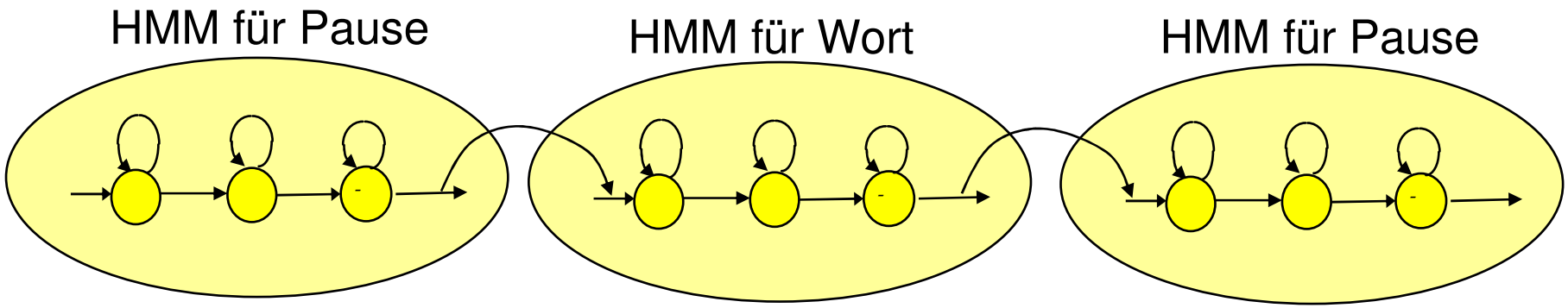
- Initialmodell durch lineare Segmentierung eines Sprachsignals (z.B. Worts) gemäß der Anzahl von Zuständen des HMMs
- „Erzwungene“ Erkennung mit möglichst vielen, bekannten Trainingsäußerungen (grosse Sprachdatenbasis)
- Aus der erzwungenen Erkennung resultiert eine zeitliche Zuordnung der Vektoren zu Zuständen
- Bestimmung der Gauß-Verteilungen aus allen Vektoren, die einem Zustand zugeordnet werden
- Iterative Wiederholung

- Kurze Wörter (~500 … 1000 ms) werden mit etwa 16 Zuständen und Laute (phoneme) mit etwa 3 bis 6 Zuständen als HMM modelliert.

- Bestimmung der Wahrscheinlichkeit, dass eine Folge von Merkmalsvektoren durch ein HMM modelliert werden kann
- Suche nach dem Maximum der bedingten Wahrscheinlichkeiten für das HMM w_i , gegeben eine Folge von Merkmalsvektoren [X1, X2, …, Xn]

  MAX { P(w_i | [X1, X2, …, Xn] }

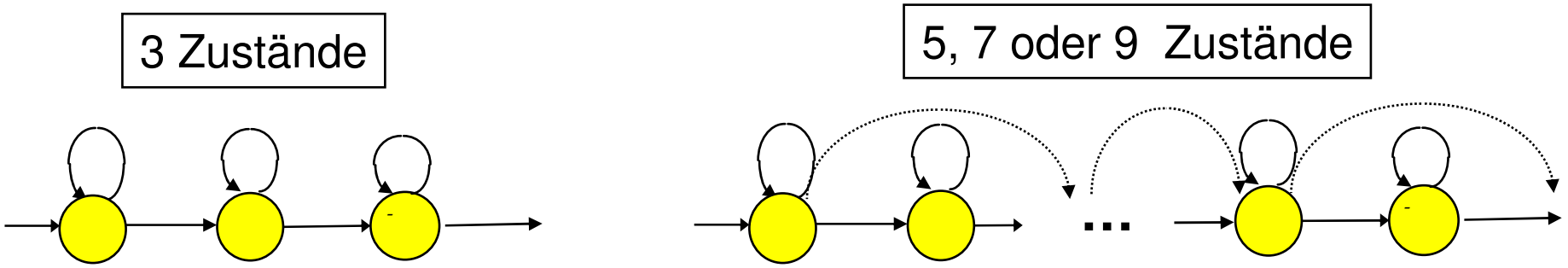
- Wortkettenerkennung:

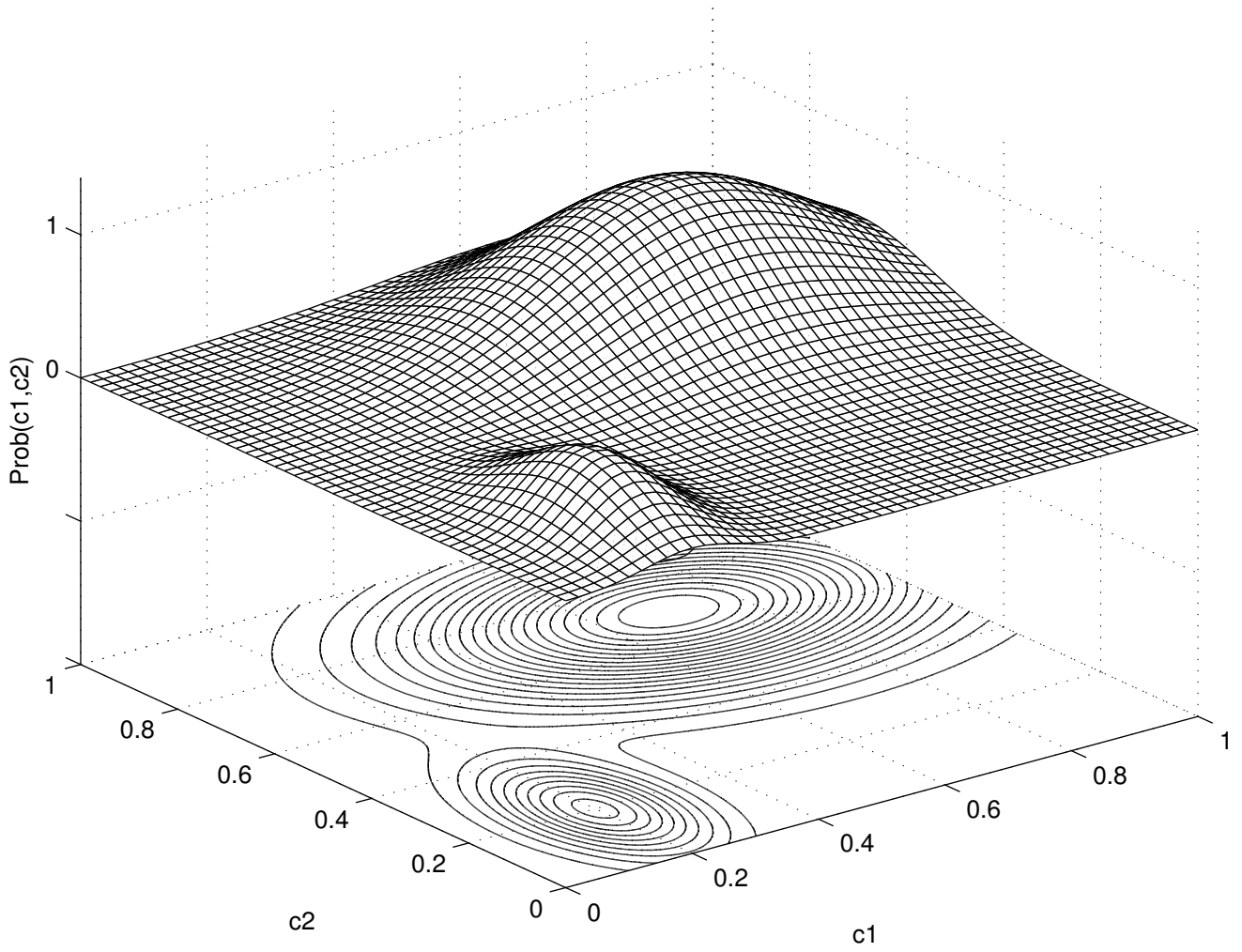


- Modellierung von Wörtern als Folge von Lauten (Phonemen)
  ➔ Aneinanderhängen der Phonem-HMMs zu einem Wort-HMM
- etwa 50 Phoneme im Deutschen
- Training der Phoneme nötig
- Benötigt wird eine Sprachdatenbasis:
  - Sollte alle Phoneme in ausreichender Anzahl beinhalten
  - Neben der Sprache wird eine Lautbeschreibung jeder Äußerung mit einer zeitlichen Zuordnung benötigt

- „SpeechDat“ – Sammlung über Telefon von ca. 1000 Sprechern:
  - ~32000 Äußerungen (= ~48 h) zum Training verwendet (Wörter, Ziffern, Zahlwörter, Sätze, …)

- Betrachtung der Laute ohne Berücksichtigung des vorhergehenden und des nachfolgenden Lauts ➔ „Monophon“
- Modellierung von 53 unterschiedlichen Monophonen durch HMMs mit definierter Anzahl von Zuständen und definierter Anzahl von Gauß-Verteilungen je Zustand
- Untersucht wurden:
  - HMMs mit 3, 5, 7 oder 9 Zuständen
  - 1, 2, 4, 8 oder 16 Gauß-Verteilungen je Merkmal und je Zustand




Prob(c1,c2)
1
0
1
0.8
0.6
0.4
0.2
0
c2
0
0.2
0.4
0.6
0.8
1
c1

- Betrachtung der Laute **mit** Berücksichtigung des vorhergehenden und des nachfolgenden Lauts ➔ „Triphon“
- Relativ große Anzahl: ~9000
- Problem: genügend viele Beispiele jedes Triphons in den Trainingsdaten?
- Kompromiss: Modellierung in Abhängigkeit der **Lautklasse** des vorhergehenden und des nachfolgenden Lauts ➔ ~750 Triphone

Kürzel	Klasse	Elemente
V	Vokal	a: e: i: o: u: a E I O U y: oe Oe ae ar an E: @
F	Frikativ	h f v z x s S C
N	Nasal	m n l r j Y N Z On
P	Plosiv	t g p d k b
D	Diphon	al OY aU

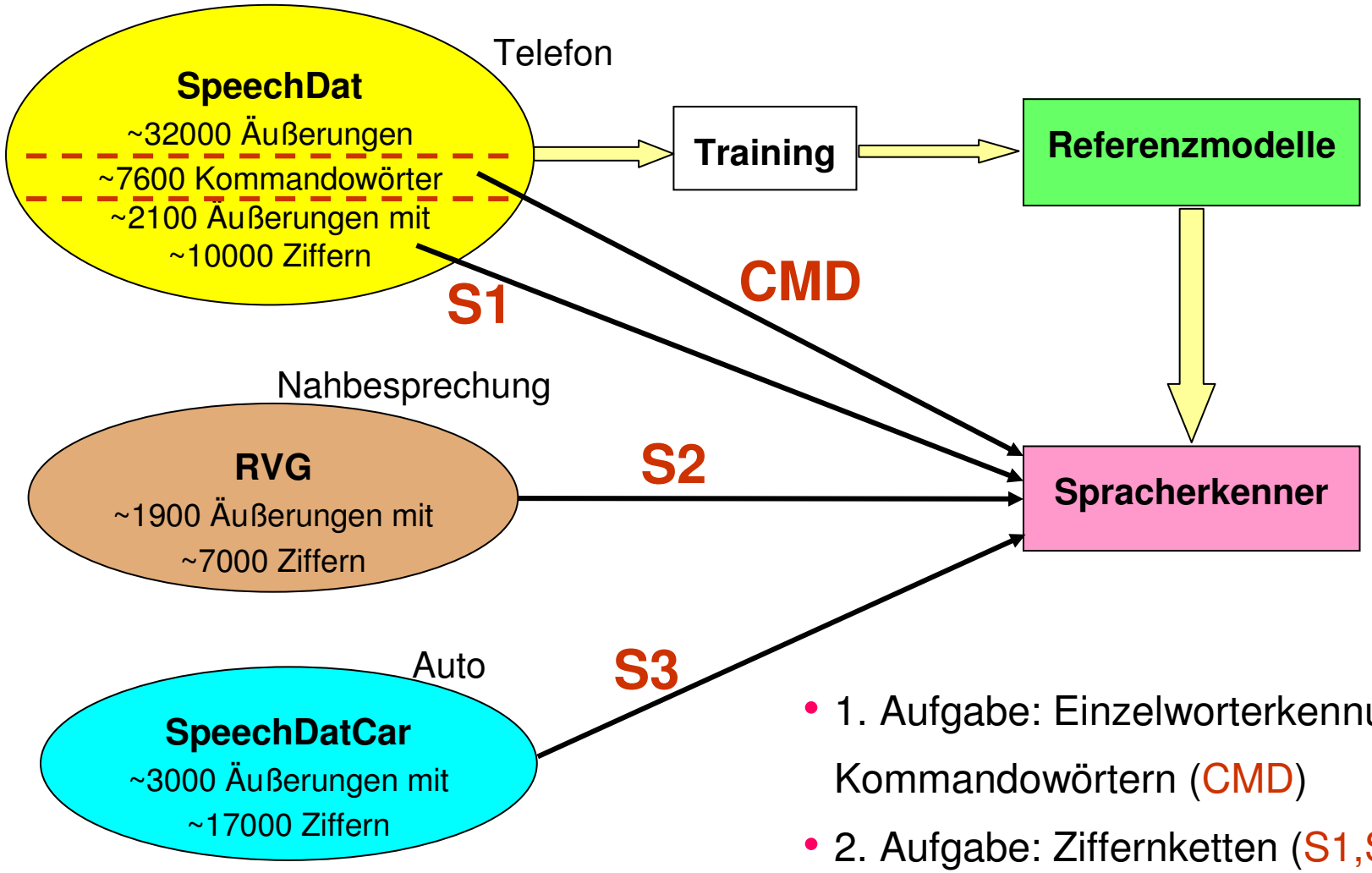
- Untersucht wurden wieder:
  - HMMs mit 3, 5, 7 oder 9 Zuständen
  - 1, 2, 4, 8 oder 16 Gauß-Verteilungen je Merkmal und je Zustand

- Sprachanalyse mit einer von ETSI standardisierten „robusten“ Cepstralanalyse (Module zur Störunterdrückung und zur Frequenzgangkompensation)

- Training und Erkennung mit frei verfügbarer Software zum Training und zur Erkennung mit HMMs (HTK – Hidden Markov Model Toolkit, Cambridge, UK)

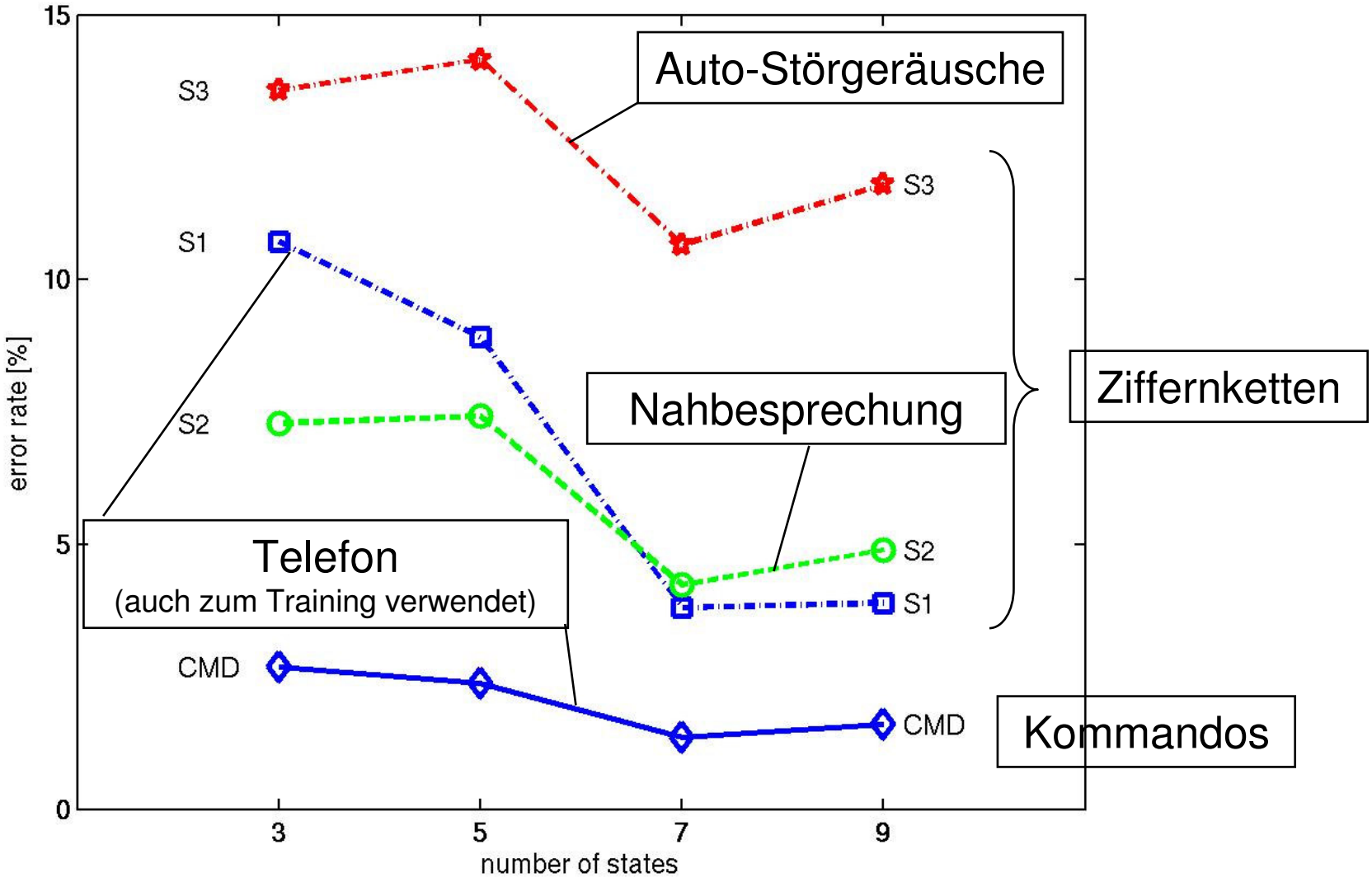
- Erkenntnisse dieser Experimente können übertragen und genutzt werden für eigenen Erkenner

Sprachdaten zum Training und zur Erkennung

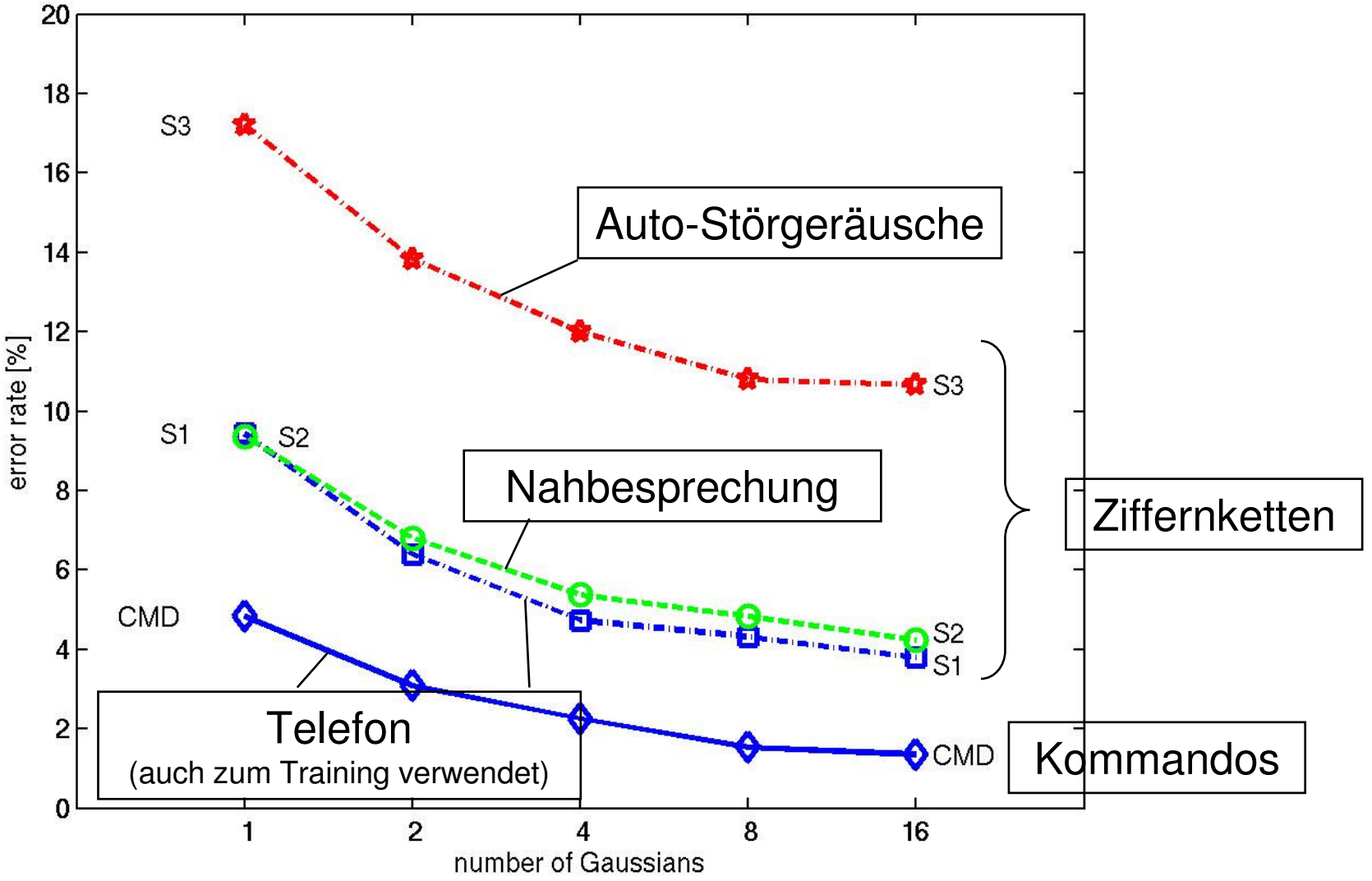


- 1. Aufgabe: Einzelworterkennung von Kommandowörtern (CMD)
- 2. Aufgabe: Ziffernketten (S1,S2,S3)

Wortfehlerraten (16 Gauß-Verteilungen pro Zustand)

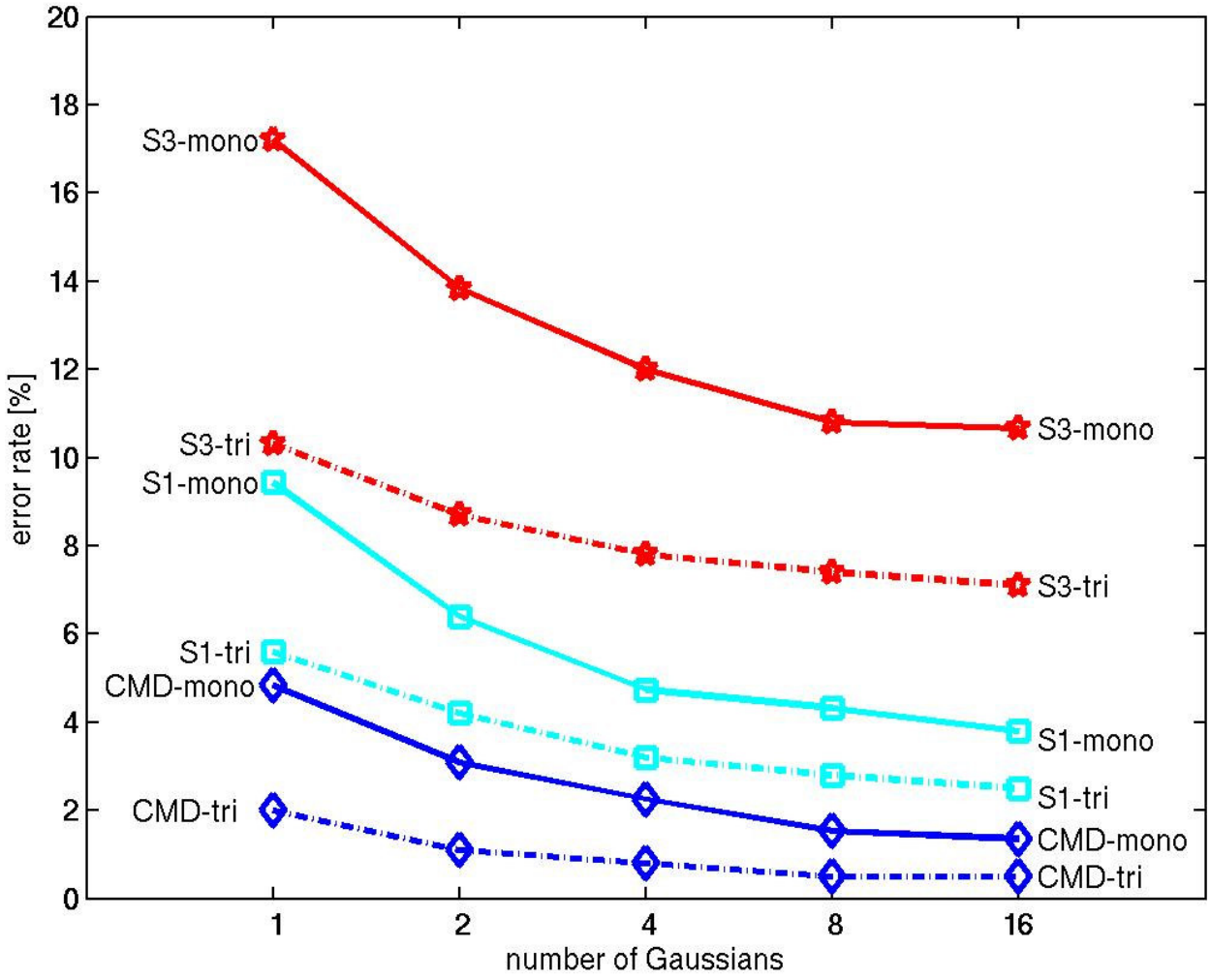


Wortfehlerraten (7 Zustände je HMM)



Wortfehlerraten (7 Zustände je HMM)

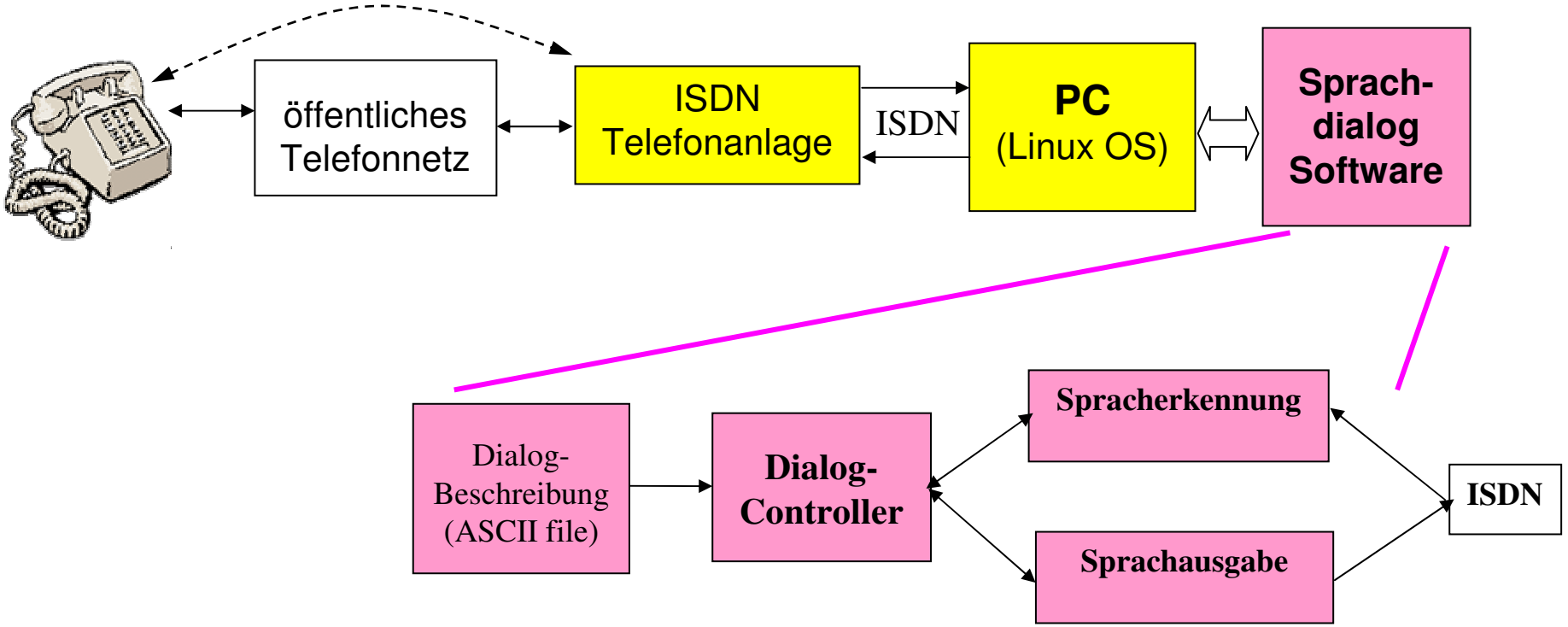
Vergleich von Monophon- und Triphon Modellen



Simulationsergebnisse

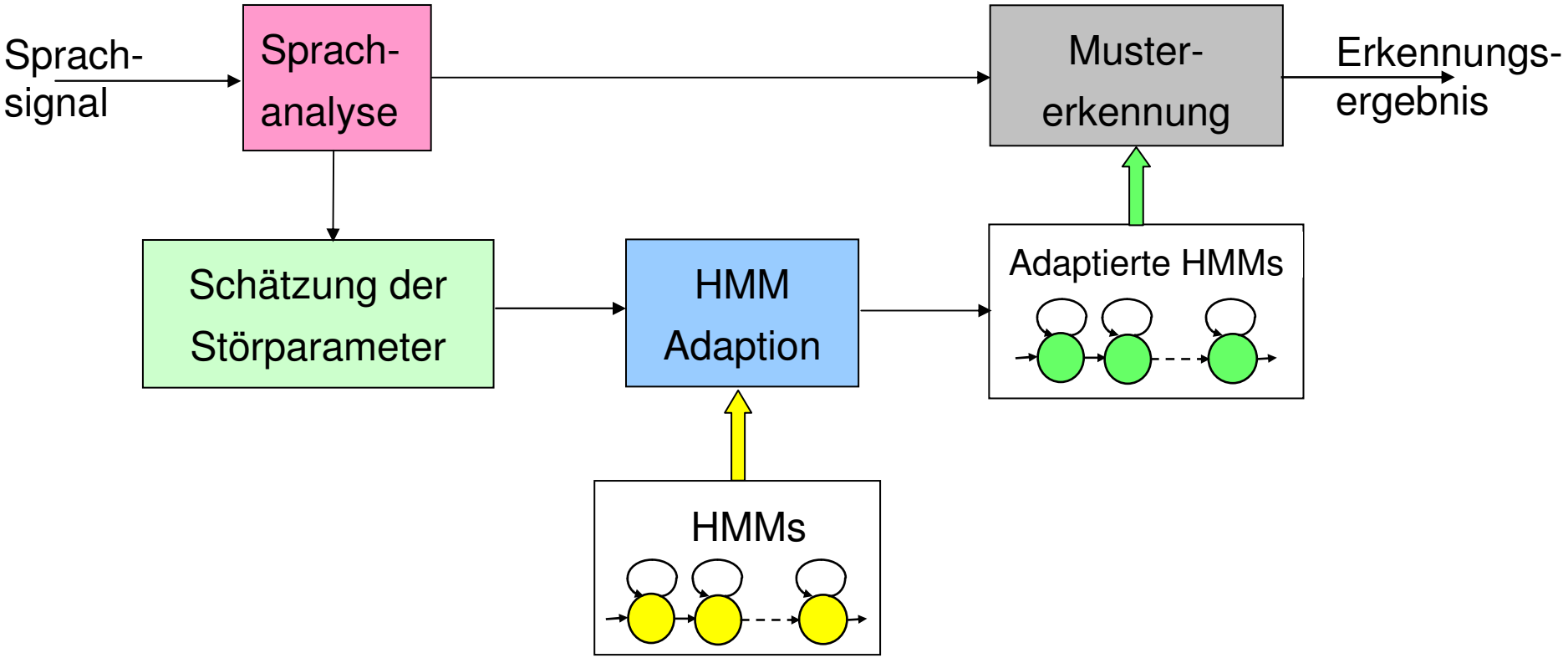
- Geringste Fehlerraten für HMMs mit 7 Zuständen
- Keine wesentliche Verbesserung mehr für mehr als 8 Gauß-Verteilungen je Zustand
- Triphon-Modelle besser als Monophon-Modelle

Sprachdialogsystem

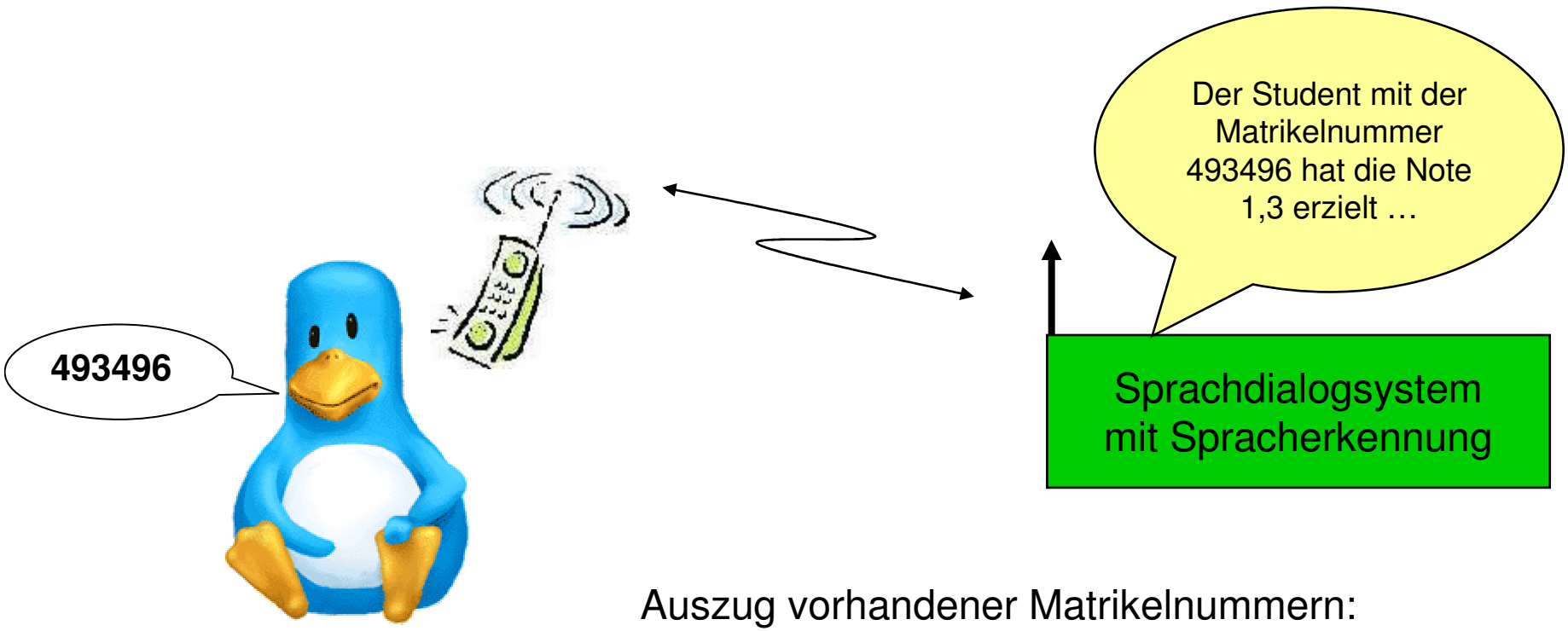


- Spracherkenner zeichnet sich durch relativ hohe Robustheit gegenüber stationären Hintergrundstörungen aus
- wird erzielt durch eine Adaption der Referenzmuster

Adaption der Referenzmuster



Automatische Notenauskunft

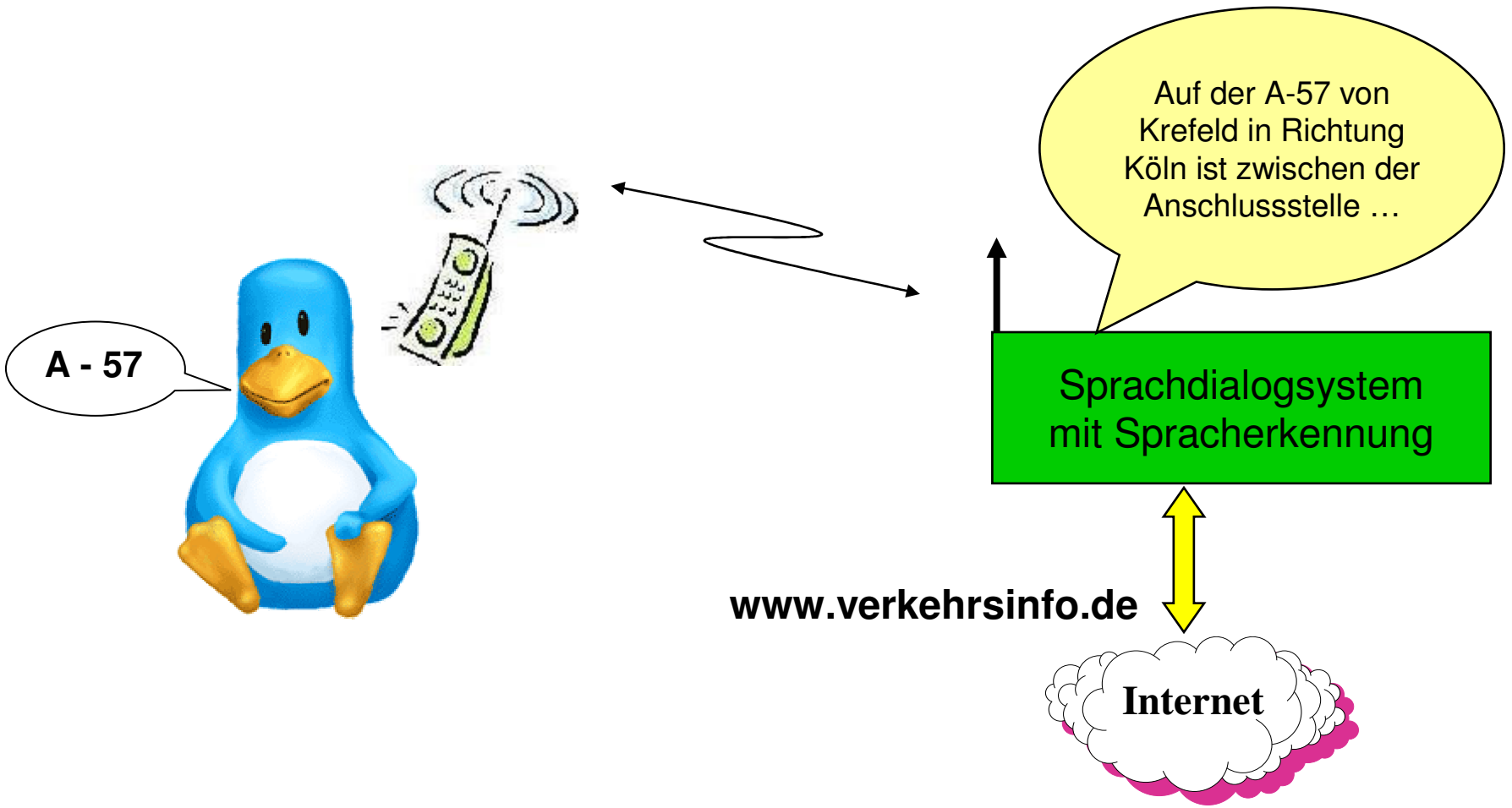


Auszug vorhandener Matrikelnummern:

356738	402828	448333
449698	470353	470239
470079	478062	507379
529704	…	

Tel.-Nr. 02151 643894

Verkehrsinfo



Tel.-Nr. 02151 643893

Demo zur phonembasierten Erkennung

Index	Name
0	Rainer Wallnig
1	Hermann Ostendorf
2	Gabriele Behler
3	Peer Steinbrück
4	Edelgard Buhlman
5	Gerhard Schröder