Verfahren zur robusten automatischen Spracherkennung

Hans-Günter Hirsch

Hochschule Niederrhein
Fachbereich Elektrotechnik und Informatik
Digitale Nachrichtentechnik

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

- Ansätze zur robusten Spracherkennung
- Robuste Merkmalsextraktion (ETSI Standard)
- Aurora Experimente
- Adaption der Referenzmuster

Berufliche Stationen

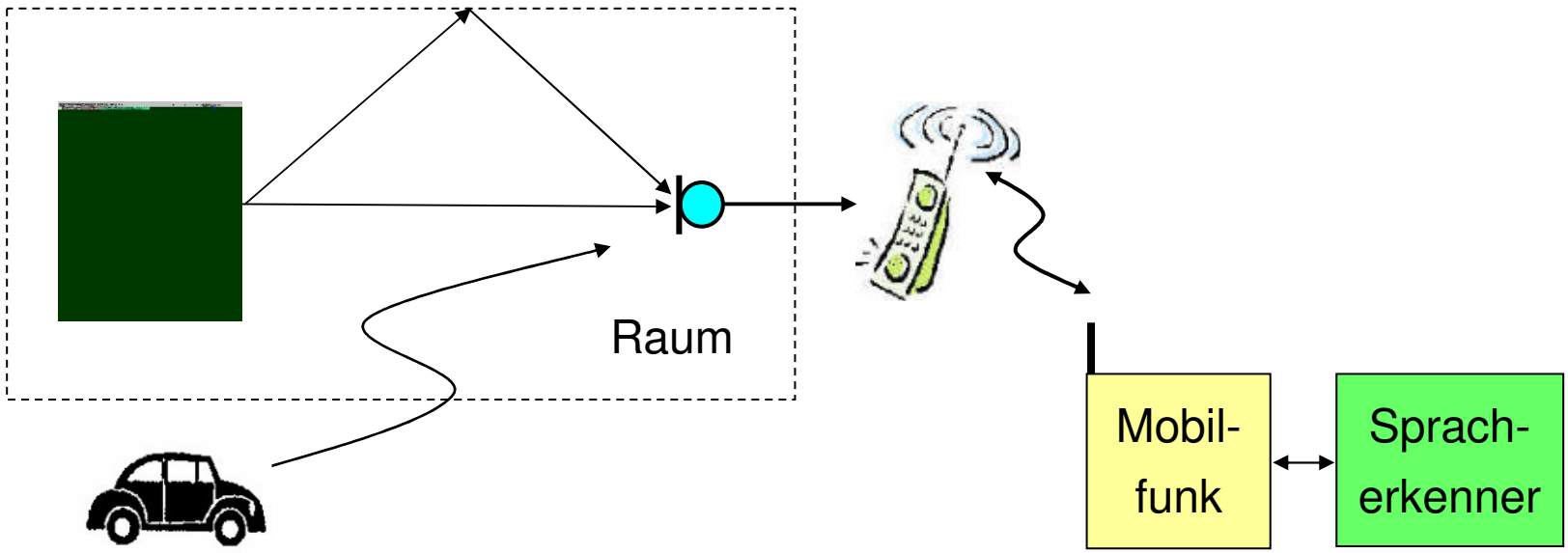
- 82 - 94: Forschung & Lehre am Institut für Nachrichtengeräte und Datenverarbeitung, RWTH Aachen
- 94 - 95: Schnurlostelefon-Entwicklung bei Ascom (CH)
- 95 - 01: Forschung & Entwicklung beim Ericsson Eurolab in Nürnberg
- seit Feb. 01: HS Niederrhein

- Forschungsaufenthalte:
  - 87-88: 6 Monate bei Philips Kommunikations Industrie (Nürnberg)
  - 92 und 93: 4 bzw. 3 Monate am Int. Computer Science Institute (Berkeley)
  - 97: 3 Monate an der KTH (Stockholm)
  - Juli 01 & Sep 02: International Computer Science Institute (Berkeley)

Aktivitäten in Lehre und Forschung

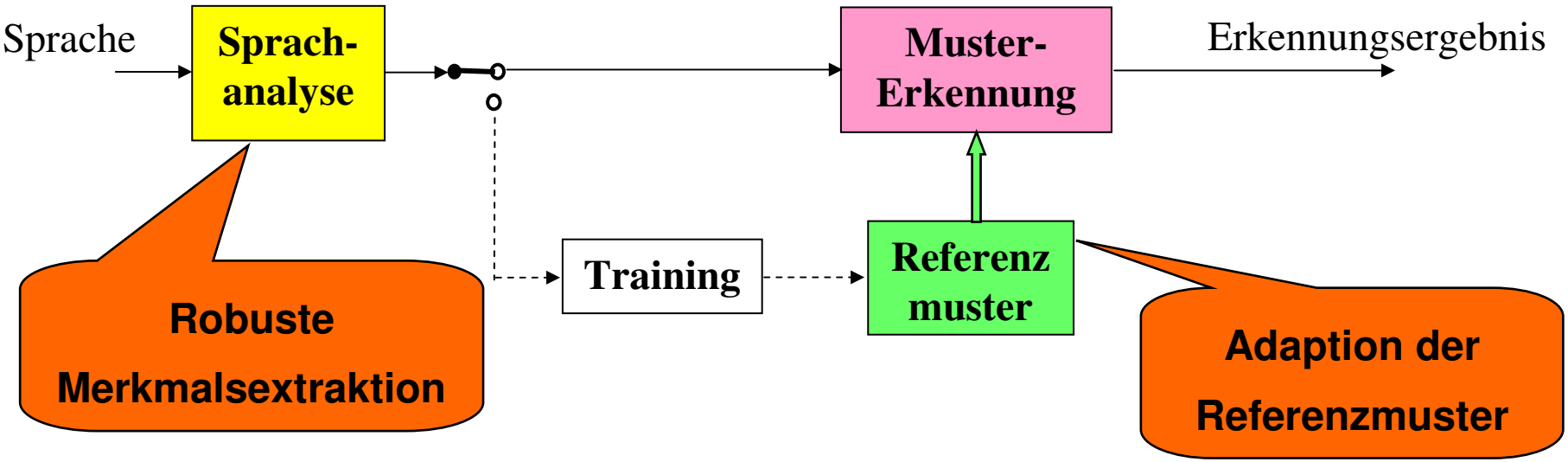
- **Lehre:**
  - "Digitale Verfahren in der Nachrichtentechnik" (Pflicht-Vorlesung für Nachrichtentechniker im 6. & 7. Semester)
  - "Digitale Sprach- und Audiosignalverarbeitung" (Wahlvorlesung für alle Studienrichtungen im 6. & 7. Semester)
  - Projektfächer zur „Audiosignalverarbeitung mit Matlab" im 2. und 3. Studienjahr
- **Forschung und Entwicklung:**
  - Sprachcodierung
  - Automatische Spracherkennung (speziell: Robustheit)
  - Sprachdialogsysteme
  - Aktuelle Projekte:
    - Lautbasierte Erkennung zum Erkennen von Eigennamen
    - Automatische Notenauskunft per Telefon
    - Verkehrsinformationen per Telefon
    - Sprachgesteuerte Nebenstellenanlage

Störeinflüsse



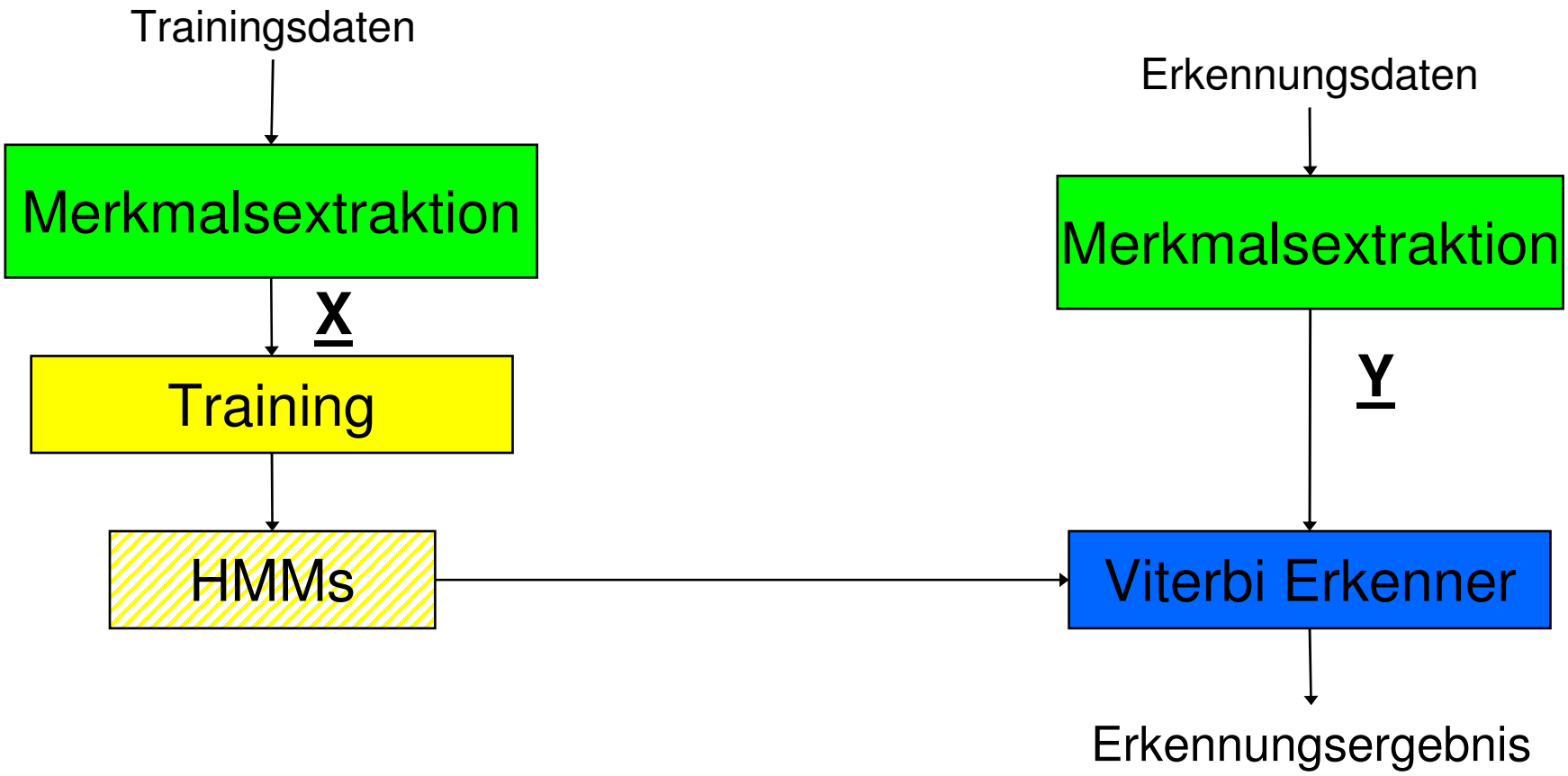
- Additive Hintergrundstörungen
- Unbekannte Frequenzgänge (z.B. Mikrofon, Telefonkanal)
- Hallige räumliche Umgebung
- Mobilfunkkanal (Sprachcodierung, Funkkanal)

Ansätze zur robusten Erkennung



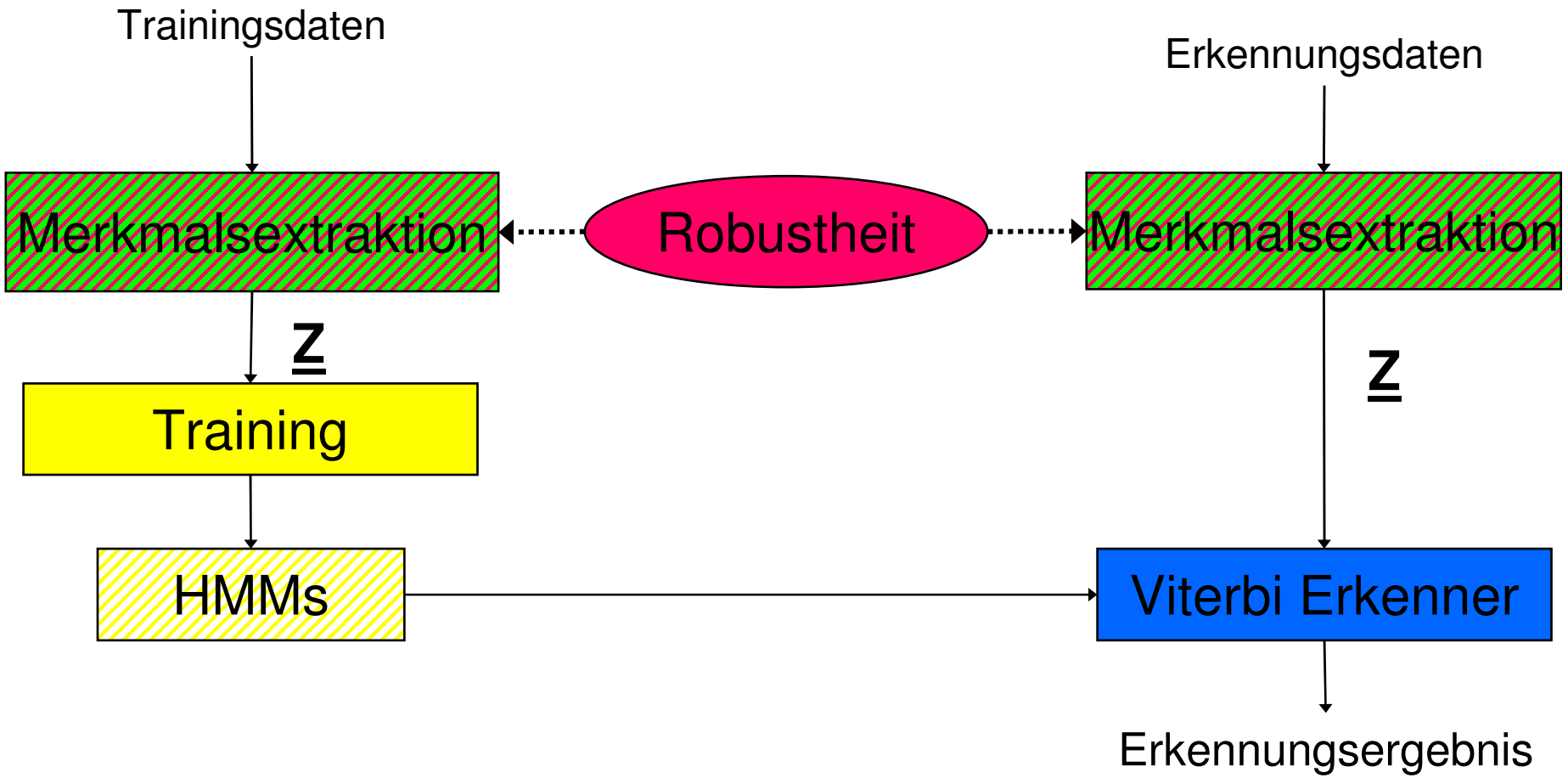
1. Extraktion robuster akustischer Merkmale bei der Sprachanalyse
2. Adaption der Referenzmuster auf die aktuelle Störumgebung

Problem

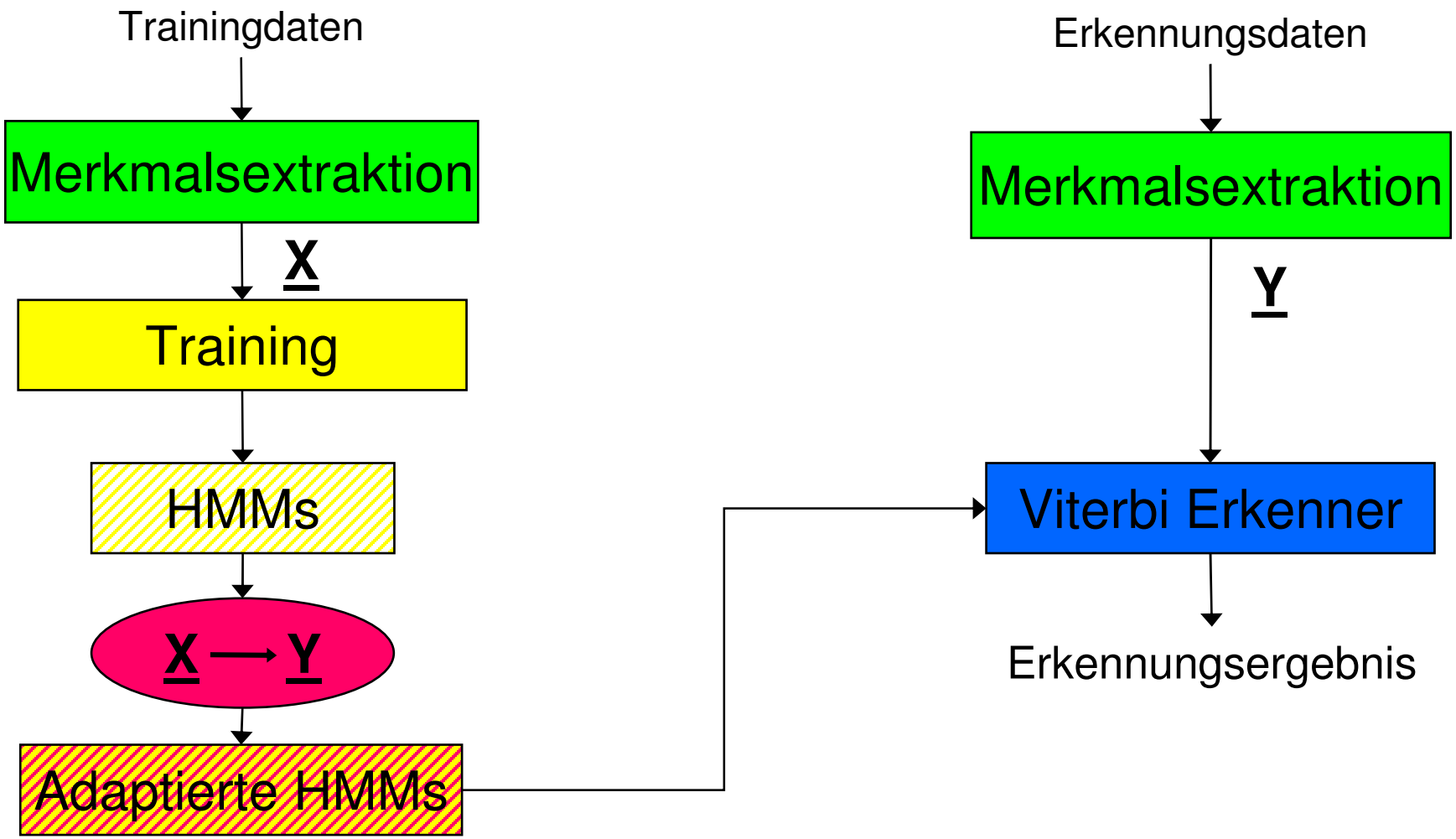


X und **Y** repräsentieren u.U. verschiedene Bereiche im Merkmalsraum

Robuste Merkmalsextraktion



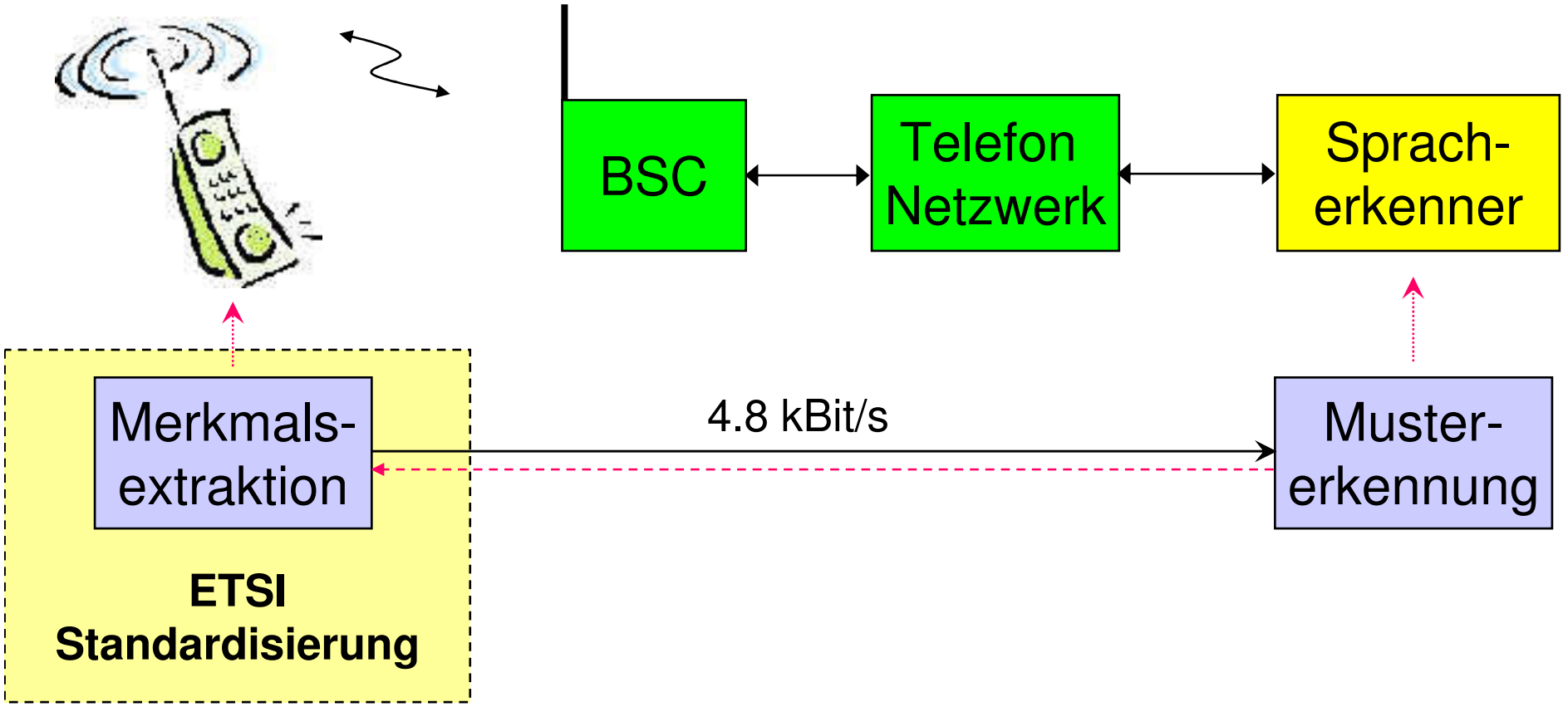
HMM Adaption (X → Y)



Aurora

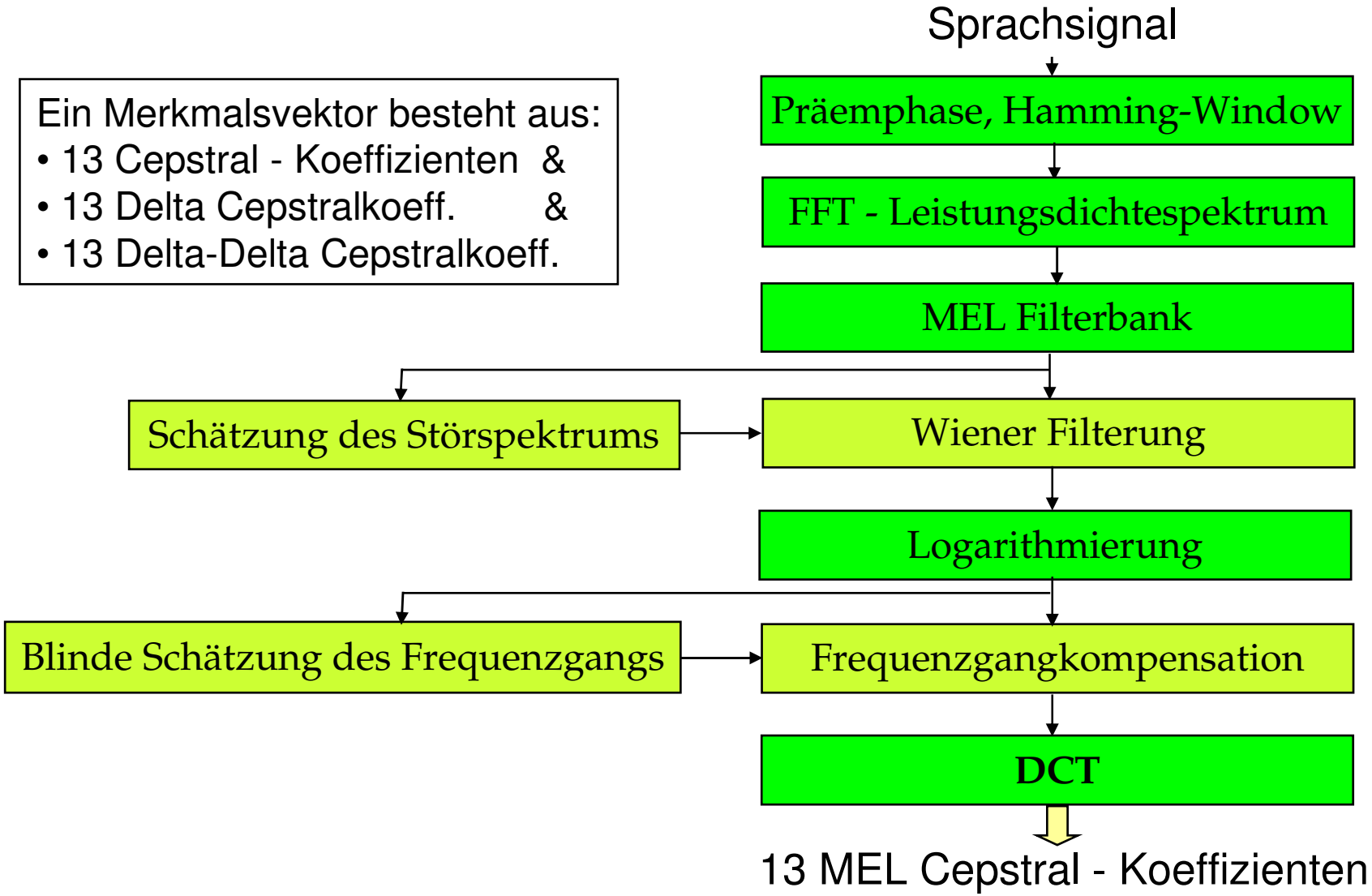
- Ziel: Standardisierung von Verfahren zur Merkmalsextraktion
  ➔Implementierung in Telekommunikationsendgeräten zur Realisierung einer verteilten Spracherkennung
- 1. Standard: herkömmliche MFCC-Analyse (2000)
- 2. Standard: Robuste MFCC Analyse (2002)
- Definition einiger Erkennungsexperimente

Verteilte Spracherkennung



Standardisierte (ETSI), robuste Merkmalsextraktion

Ein Merkmalsvektor besteht aus:
- 13 Cepstral - Koeffizienten &
- 13 Delta Cepstralkoeff. &
- 13 Delta-Delta Cepstralkoeff.



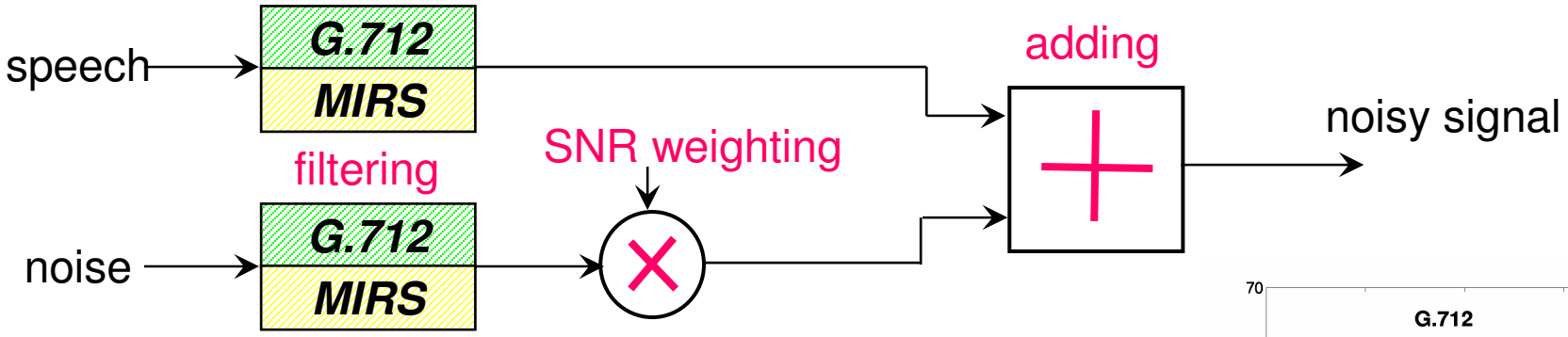
Erkennungsexperimente

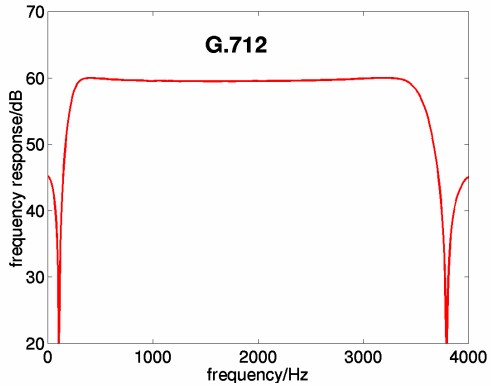
- **wortbasierte Erkennung**
  - TIDigits plus künstliche Störung
  - Extrakte der SpeechDatCar Datensammlungen

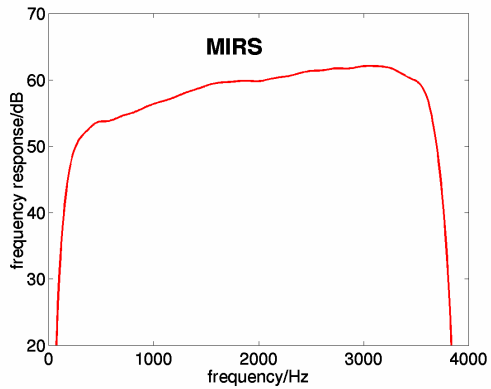
    (Ziffernketten in 4 bzw. 5 verschiedenen Sprachen)

- **phonembasierte Erkennung**
  - WSJ (wall street journal) Datensammlung plus künstliche Störung

Erzeugung gestörter Daten







- SNR definiert für Sprache und Sörung nach Filterung mit der G.712 Charakteristik
- Sprachpegelbestimmung und Filterung gemäß ITU (STL2000 Softwarepacket)

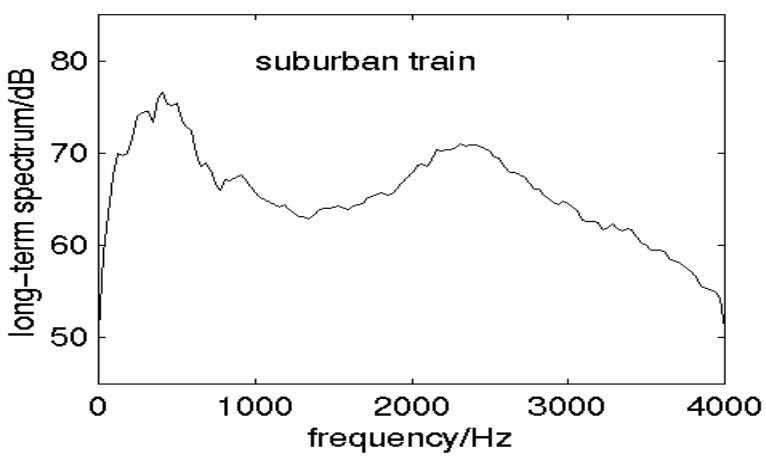
Sprache and Störgeräusche

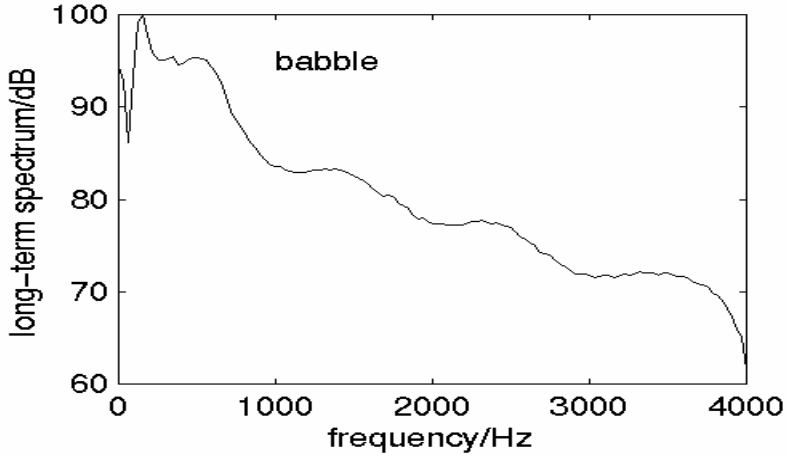
- TIDigits: Ziffernketten (US Englisch)

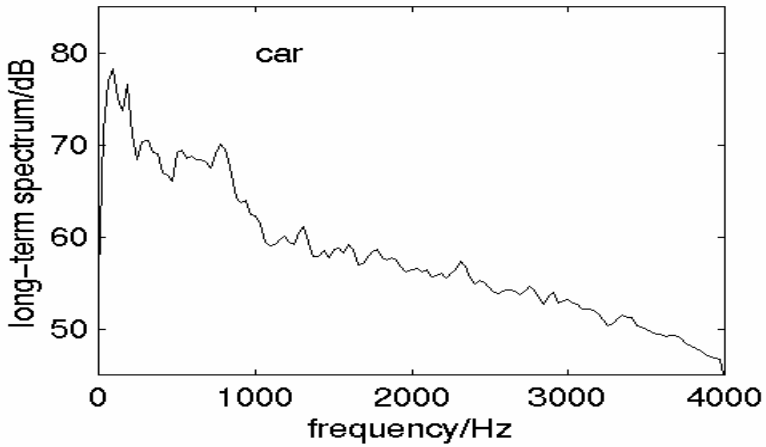
  (jeweils ca. 8700 Ketten = 28000 Ziffern für Training bzw.Test)

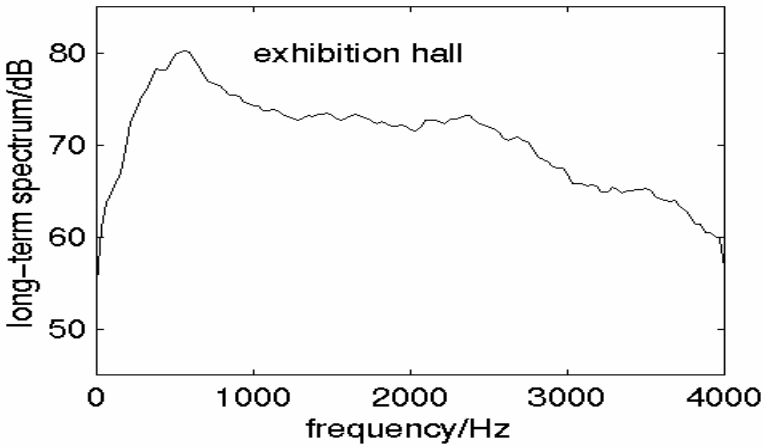
- Repräsentative Störgeräusche für Anwendungsumgebungen von Mobiltelefonen:
  - car
  - babble
  - subway
  - exhibition hall
  - restaurant
  - street
  - airport
  - train station
- SNRs: 20, 15, 10, 5, 0, -5 dB

Langzeitspektren









HTK Erkenner und Ergebnisse

- Ganzwort HMMs für alle Ziffern mit
  - 16 Zuständen pro Wort und
  - mixture of 3 Gaussians pro Zustand
- Wortfehlerraten (%):

ETSI -1

training mode	test set A	test set B	test set C
multi-condition	**11.93**	**12.78**	**15.44**
clean	**41.26**	**46.60**	**34.00**

ETSI -2

training mode	test set A	test set B	test set C
multi-condition	**7.88**	**8.04**	**9.43**
clean	**12.56**	**13.00**	**14.45**

SpeechDatCar

- Ziffernketten, aufgenommen im KFZ unter verschiedenen Störbedingungen
- Wortfehlerraten (%):

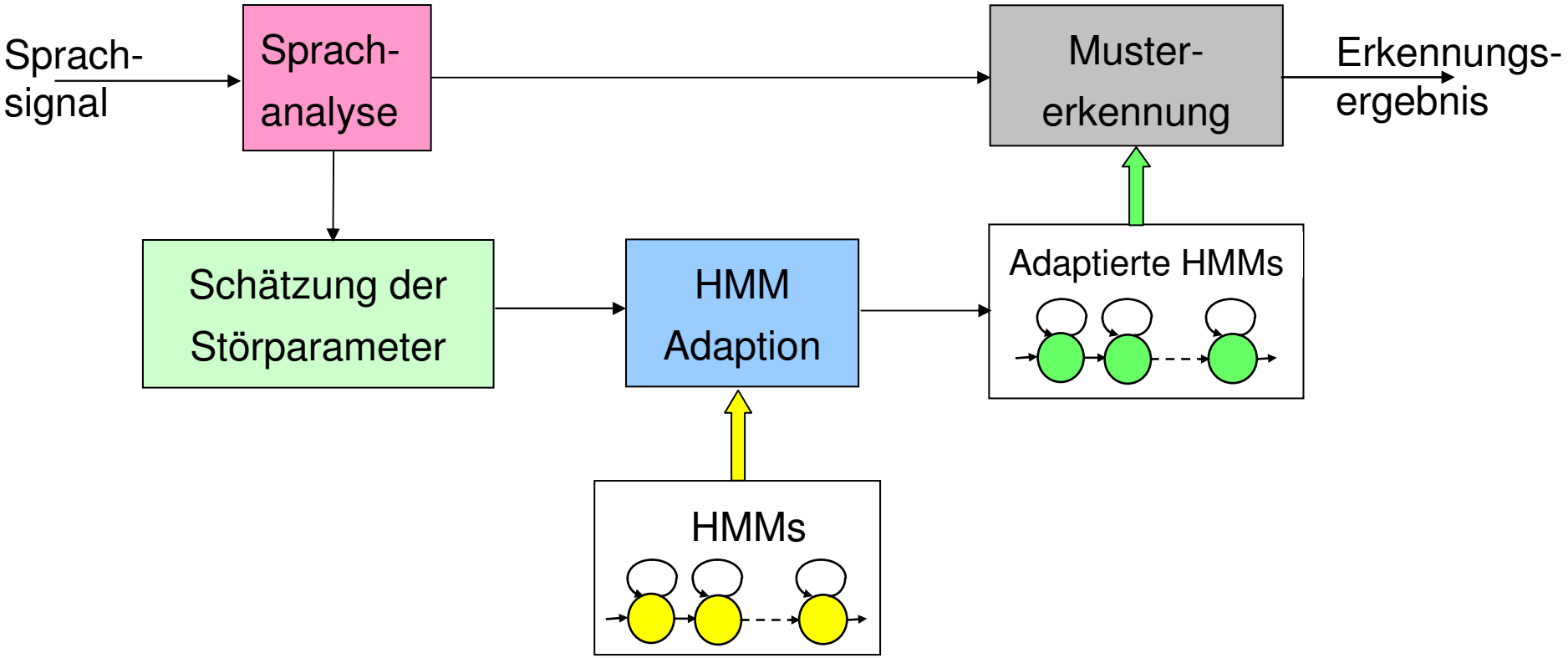
ETSI -2

mode	Finnisch	Spanisch	Deutsch	Dänisch
well matched	**7.26**	**7.06**	**8.80**	**12.72**
medium mismatch	**19.49**	**16.69**	**18.96**	**32.68**
high mismatch	**59.47**	**48.45**	**26.83**	**60.63**

Wall Street Journal

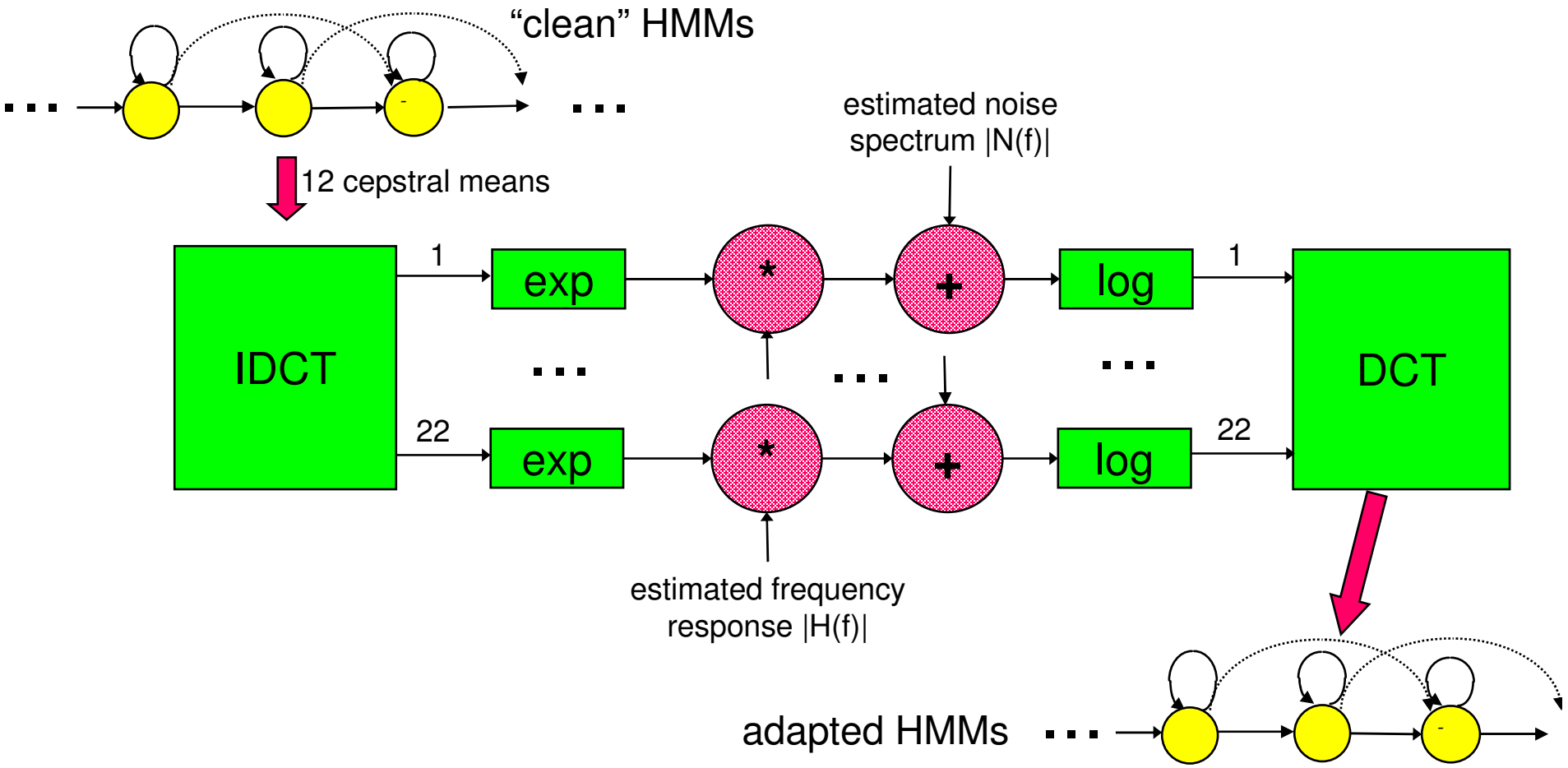
- Ganze Sätze (US Englisch)
- Definiertes Experiment vorhanden (DARPA Untersuchungen)
- erweitert um:
  - Untersuchung von 16 kHz und 8 kHz
  - additive Überlagerung von Störgeräuschen
  - clean und multi-condition Training Sets
- Erkennungssystem der Mississippi State University (Joe Picone)
- Mittlere Wortfehlerrate für ETSI-2: 34.1 %

Adaption der Referenzmuster



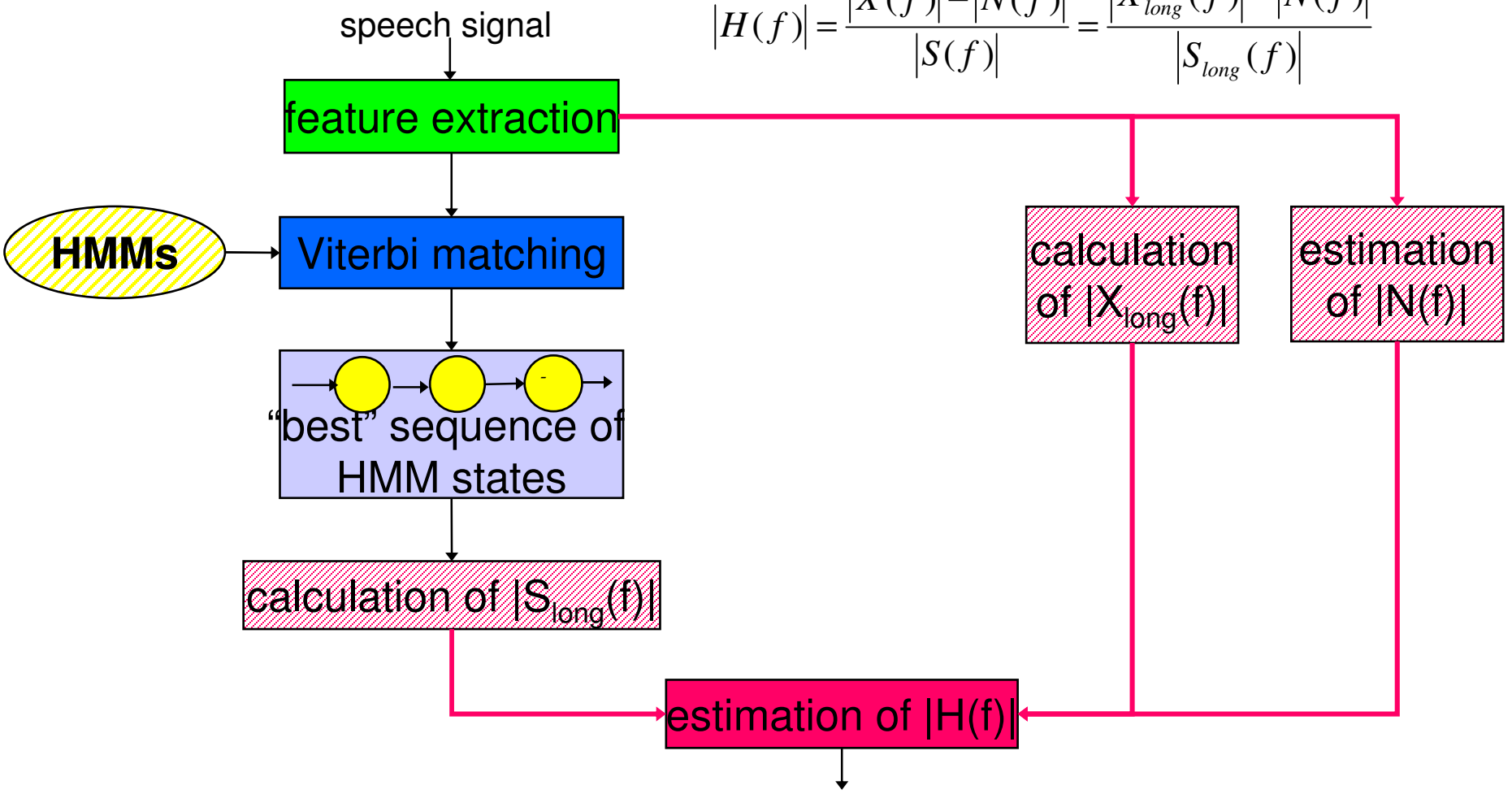
Referenzmuster: statistische Modellierung mit Hidden Markov Modellen (HMM)

Adaptation Scheme (Young&Gales, Cambridge University)



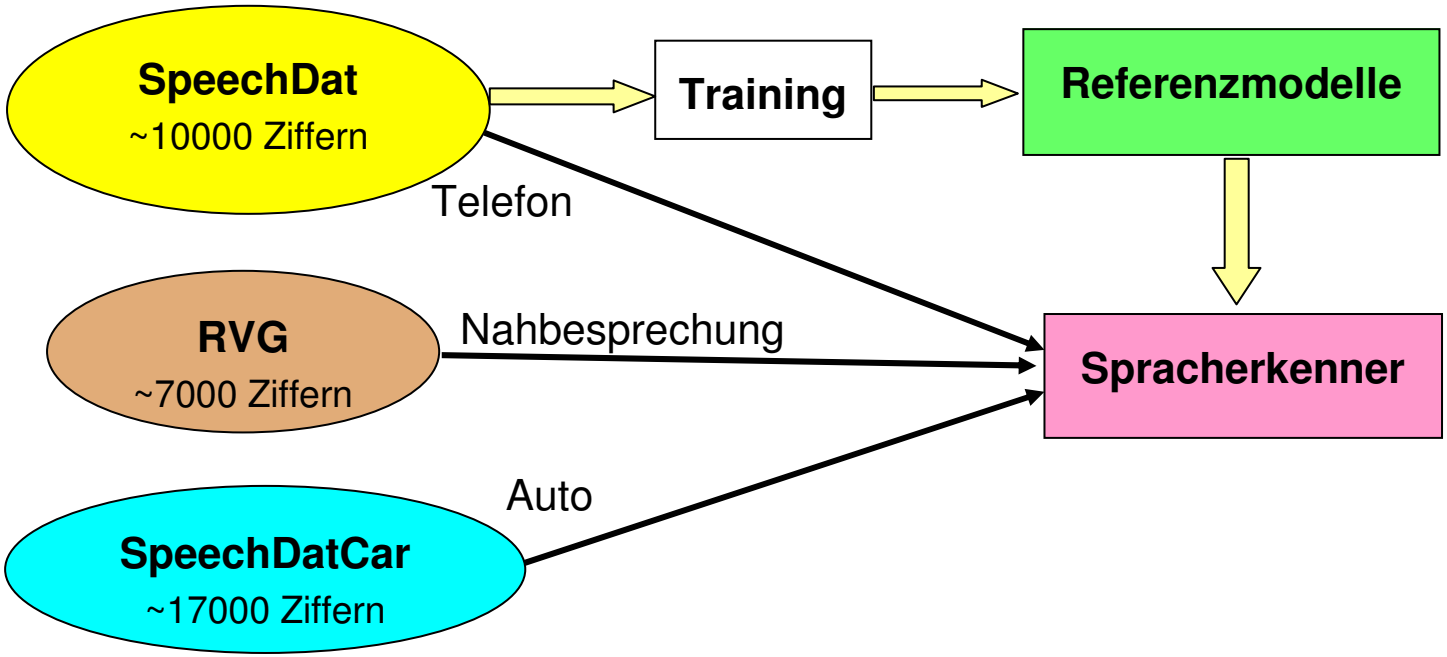
Estimation of frequency response

$$|H(f)| = \frac{|X(f)| - |N(f)|}{|S(f)|} = \frac{|X_{long}(f)| - |N(f)|}{|S_{long}(f)|}$$



Erkennungsexperimente

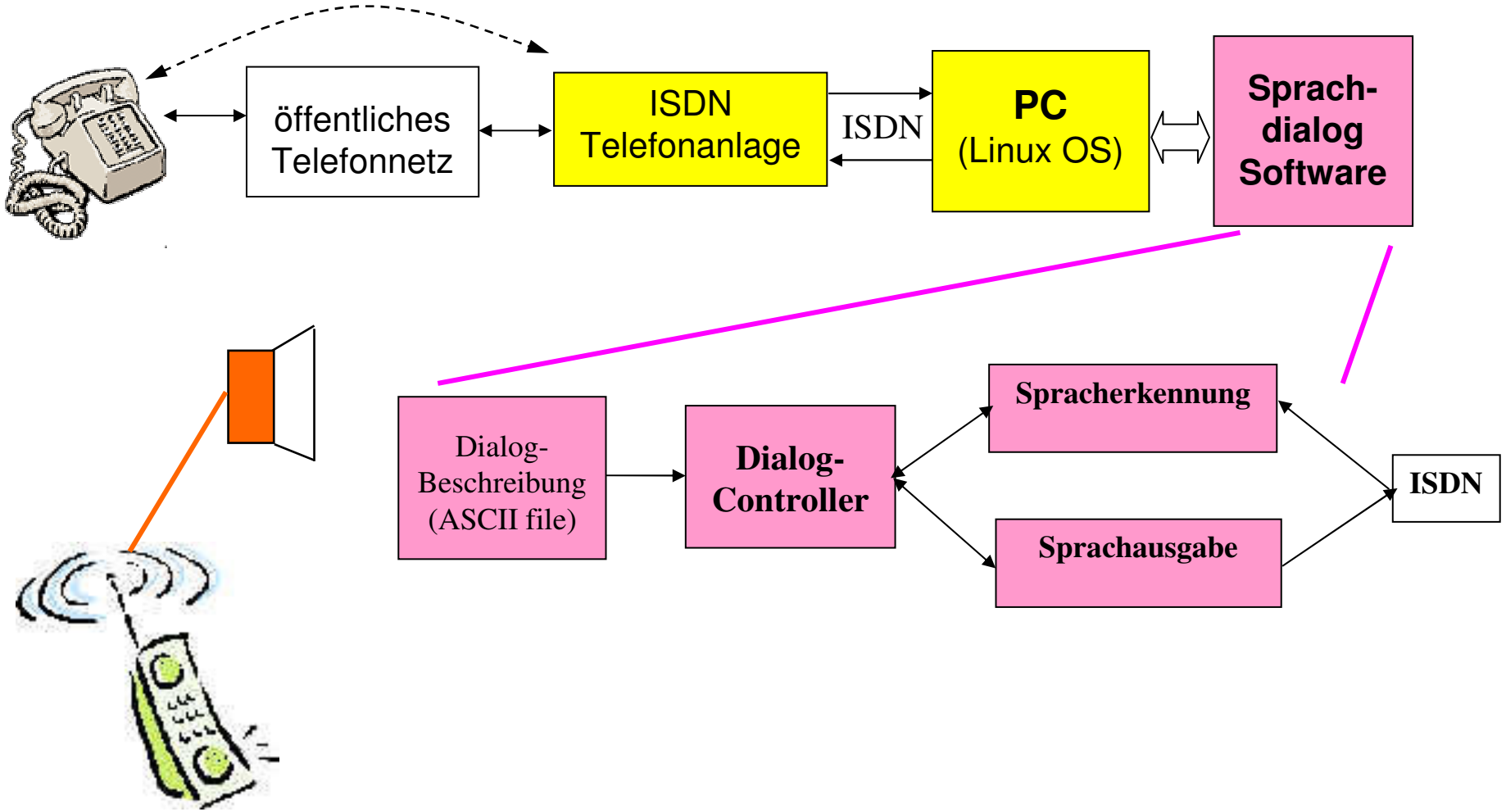
- Aufgabe: Erkennung deutscher Ziffern (ketten)



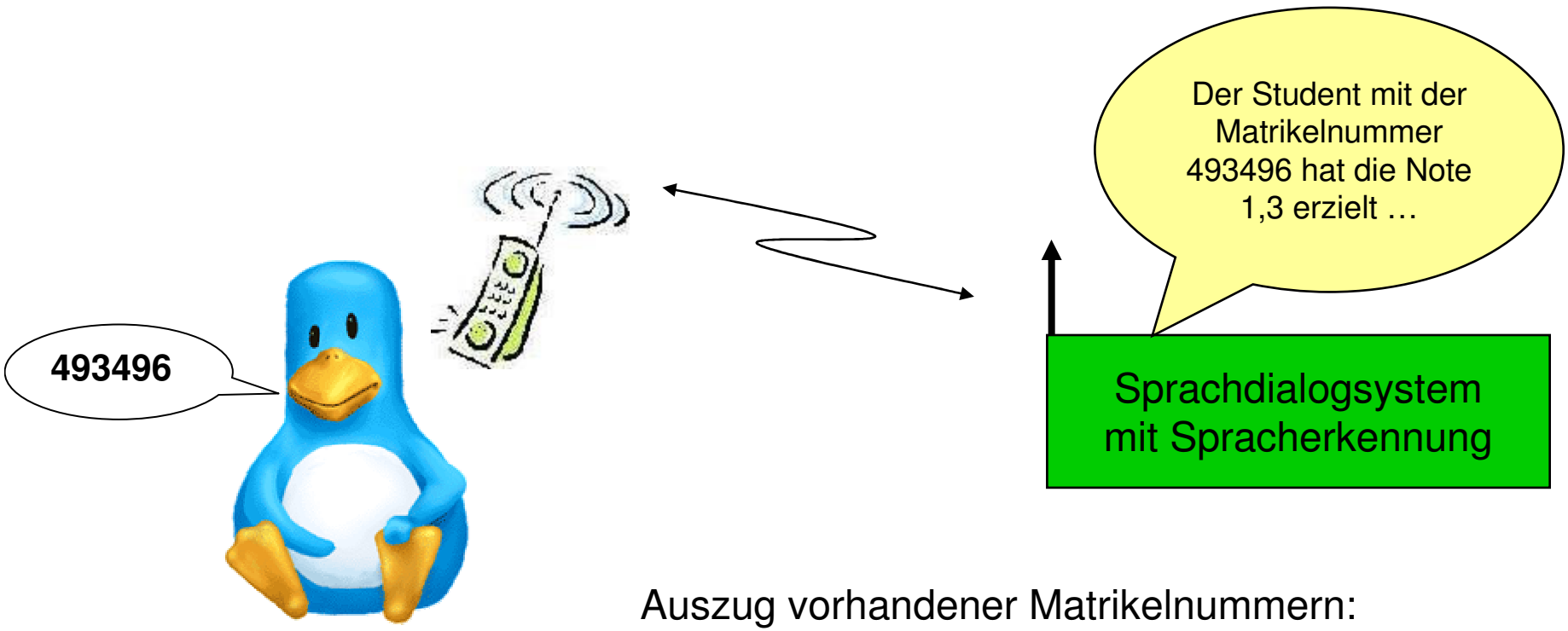
Worterkennungsraten

	SpeechDat	RVG	SpeechDatCar
39 Koeffizienten/Vektor → Standardisierte Merkmalsextraktion (ETSI-2)	96.11%	93.47 %	90.02 %
24 Koeffizienten/Vektor → Eigene Adaptionstechnik	95.88 %	93.88 %	87.21 %

Sprachdialogsystem



Automatische Notenauskunft

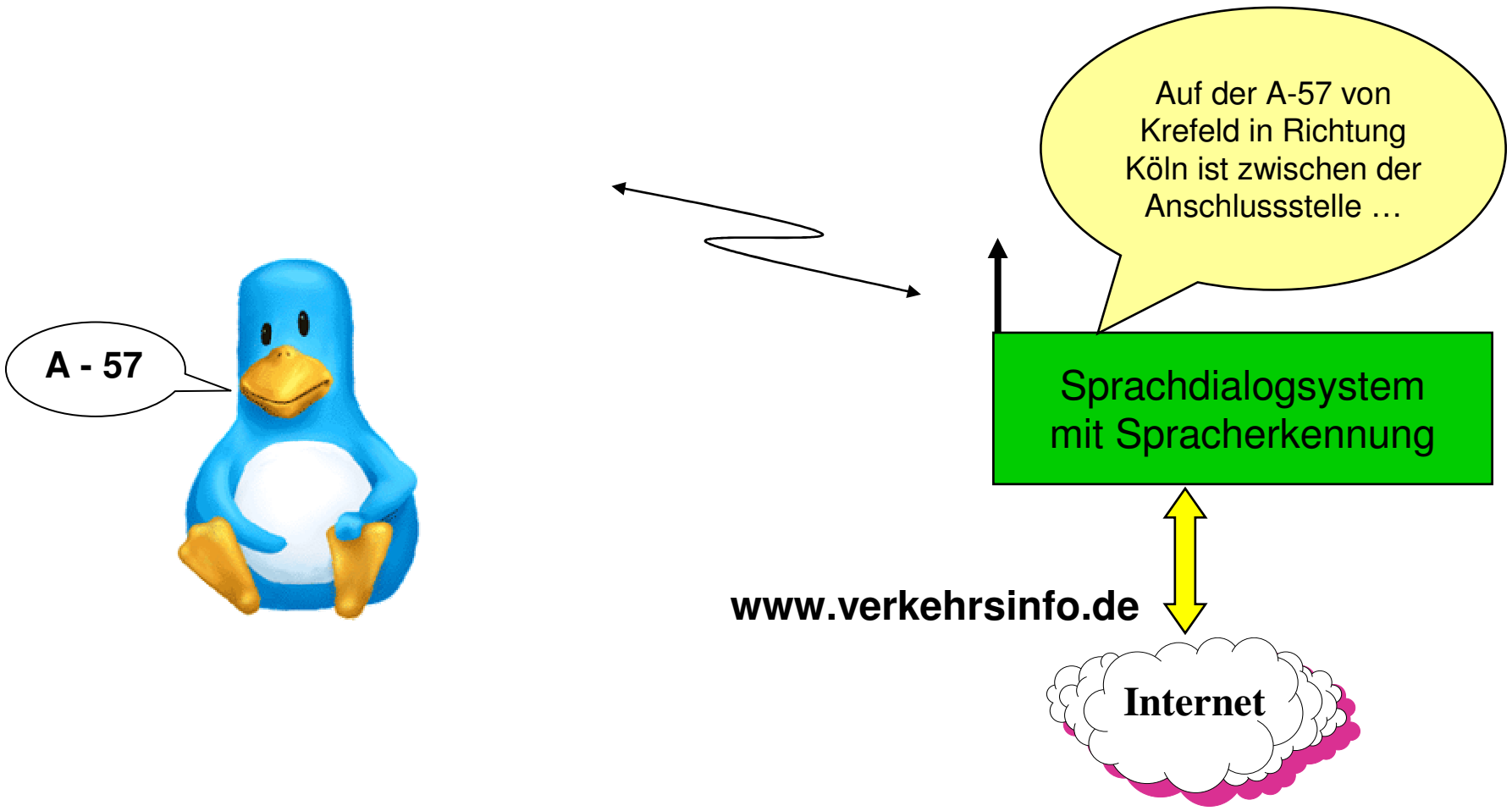


Auszug vorhandener Matrikelnummern:

3567380	4028280	4483330
4496980	4703530	4702390
4700790	4780620	5073790
5297040	…	

Tel.-Nr. 02151 643894

Verkehrsinfo



Tel.-Nr. 02151 643893

Sprachgesteuerte Nebenstellenanlage