Speech Communication II

Summer term 2004

Erhard Rank/Franz Pernkopf

Speech Communication and Signal Processing Laboratory
Graz University of Technology
Inffeldgasse 16c , 8010 Graz, Austria
Tel: 0 316 873 4436
E-Mail: rank@inw.tugraz.at

Organizational

- Lecture: 2 h/week
- Wednesday, 11.00 - 12.30
- Lecture room: i11
- Office hour: 9 to 5, phone 873 - 4436
- Exam: oral?
- Speech Communication 2 Laboratory

  First meeting: Fri., 5.3.04, 9.00 HS i11

  Speech signal analysis and synthesis, coding (DSP), recognition (HTK)
- Projects %

Projects

- Informationstechnik Projekt: 441.115 (6 PR)
- Telecom & Mobile Computing Project 441.117 (6 PR)

Speech synthesis and recognition algorithms

(MATLAB, C/C++, or on DSP)

- Harmonic-plus-noise model
- Glottis closure instant detection
- Real-time pitch modification (Dipl.th.?)
- Oscillator model
- Feature extraction

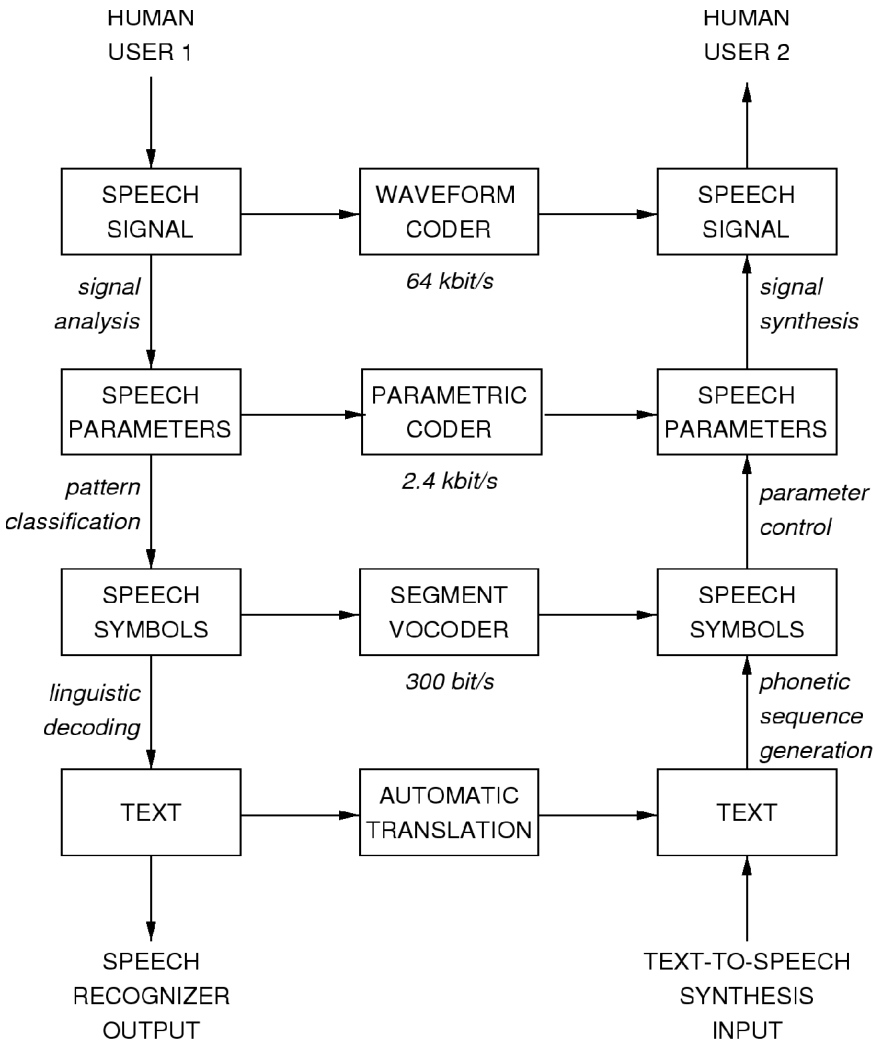
References

- L. Rabiner, B. H. Juang: "Fundamentals of Speech Recognition" Prentice Hall, Englewood Cliffs, NJ, 1993.
- E.G. Schukat-Talamazzini:"Automatische Spracherkennung", Vieweg Verlag, Braunschweig, 1995.
- R.A. Cole et al.: Survey of the State of the Art in Human Language Technology. WWW publication at www.cse.ogi.edu/CSLU/HLTsurvey, 1996 (accessed March 11, 2001).
- F. Jelinek: Statistical Methods for Speech Recognition (Language, Speech, and Communication). MIT Press 1999.
- D. Jurafsky et al: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice-Hall 2000.
- X. Huang, A. Acero, H.-W. Hon: Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR 2001.

Contents

- Automatic speech recognition (ASR)
  - Introduction
  - Feature extraction
  - Classification
  - Markov models
  - Hidden Markov models (HMMs)
  - Phonetic elements
  - Grammar models
  - Decoding (Viterbi decoder)
- Speech synthesis
  - Harmonic-plus-noise model
  - Oscillator-plus-noise model

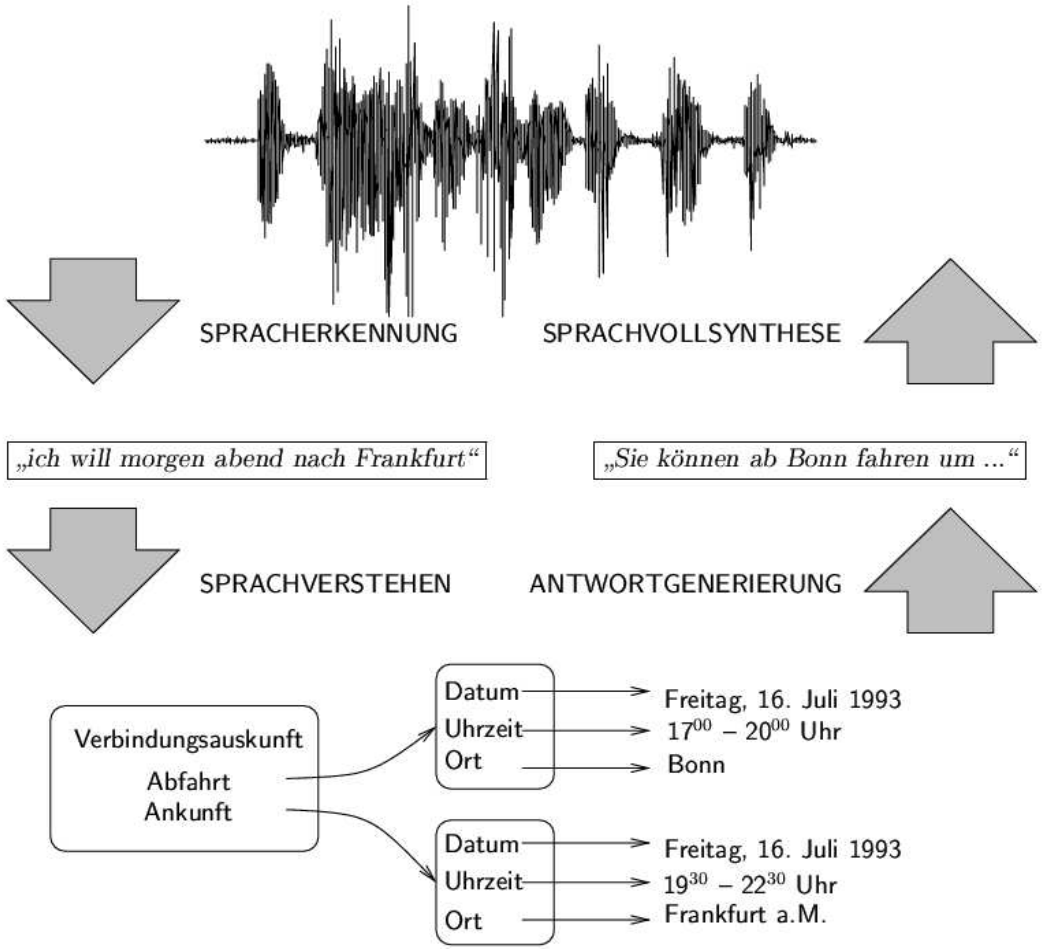
Speech coding levels:



Dialogue system:

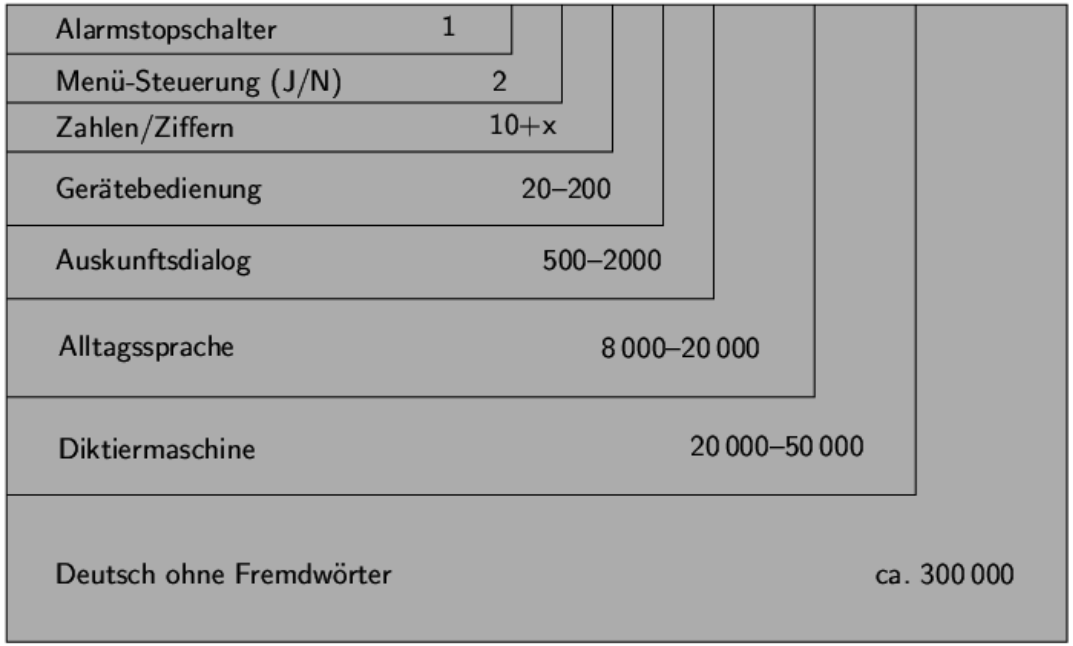
Bahnauskunftsdialog:

S: Hier ist die automatische InterCity-Auskunft. Was kann ich für Sie tun?
B: *Ich will morgen abend nach Frankfurt.*
S: Sie können ab Bonn fahren um [...]
B: *Gibt es auch noch einen früheren Zug?*
S: Bis wann möchten Sie spätestens in Frankfurt ankommen?
B: *Bis einundzwanzig Uhr.*
S: Sie können ab Bonn fahren um [...]
B: *Vielen Dank. Auf Wiedersehen.*

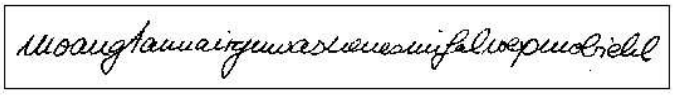
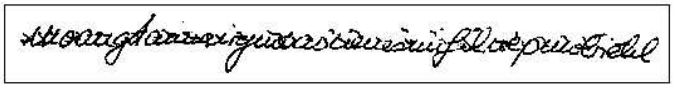
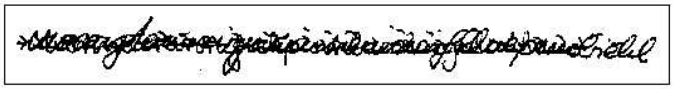


Introduction

- Entwurfsparameter spracherkennender Systeme:
  - Darbietungsform: isolierte Einzelwörter, kontinuierliche Sätze und Passagen
  - Kommunikationsmodus: Kommandos, Dialog, Übersetzung, ...
  - Sprecherabhängigkeit: ein Sprecher, Sprechergruppe, beliebige Sprecher, adaptiv
  - Sprachsignalqualität: Sprachaufnahmequalität, Bandbreite, Störgeräusche, Raum
  - Wortschatz: Umfang, Satzgrammatik, ...

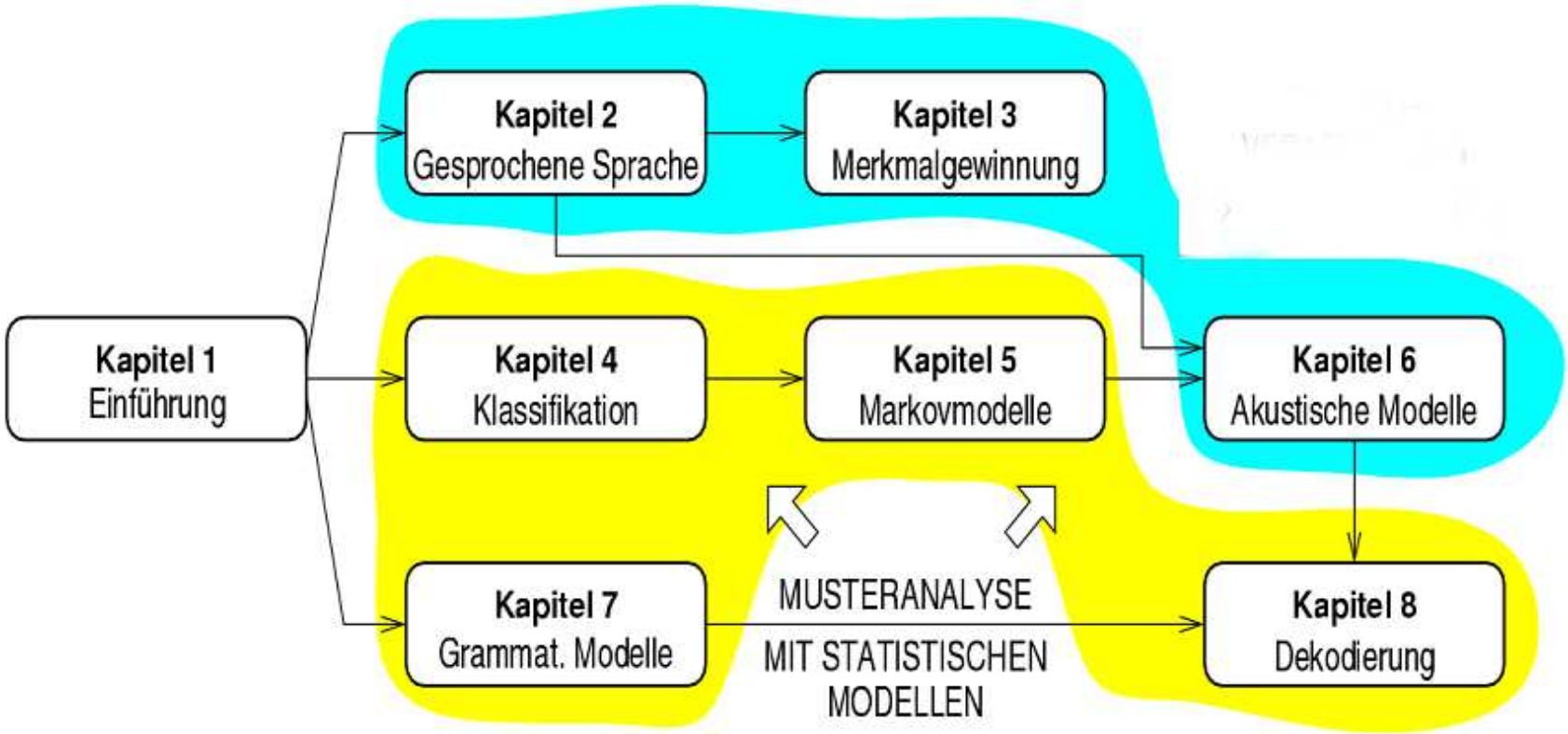


Gesprochene & handgeschriebene Sprache ...

Guten Morgen, Herr Hauptkomissar Thanner. *Gibt es irgendetwas Neues im Fall "Verbmobil"?*	der Text in "Schönschrift"
Morgen, Thanner. *Irgendwas Neues im Fall "Verbmobil"?*	spontan gesprochene Sprache
morgen thanner irgendwas neues im fall verbmobil	Großschreibung? Satzzeichen?
morgenthannerirgendwasneuesimfallverbmobil	kontinuierliche Sprache
moangtannairgnwasneuesimfalwerpmobiehl	Aussprachevarianten
	artikulatorische Verschleifung
	Störungen & Verzerrungen des akustischen Kanals
	Überlagerung d. Fremdstimmen „Cocktailparty-Effekt“

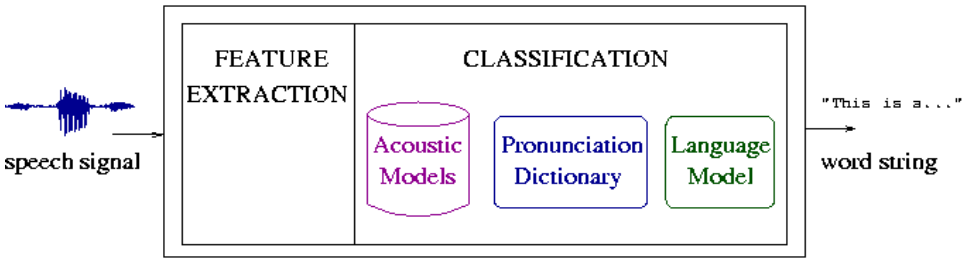
Introduction

Speech recognition



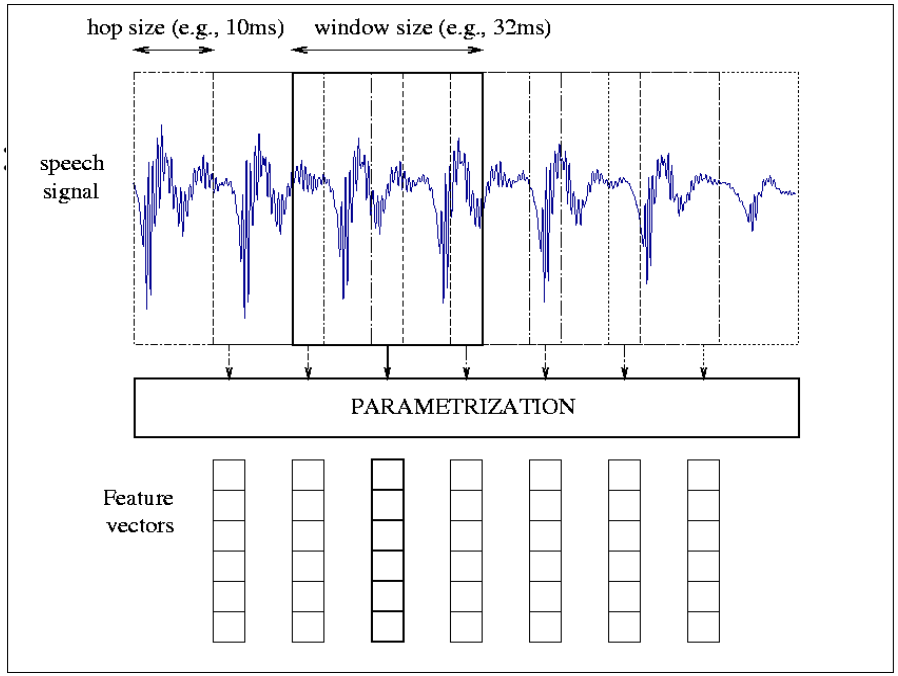
Speech recognition

System overview:



Parametrization

"Feature extraction":



Feature extraction

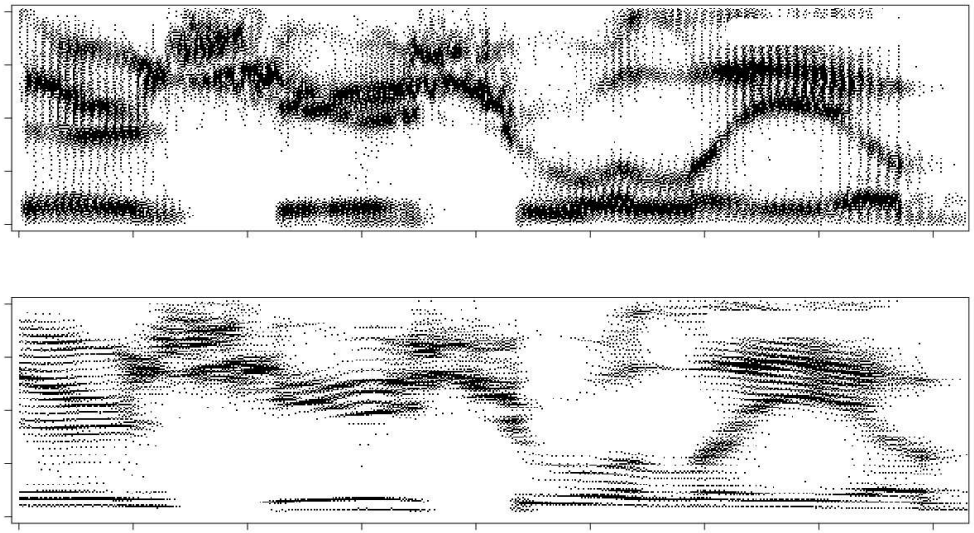
Speech is sampled at a rate between 6.6 kHz and 20 kHz

Every 10-20 ms a feature vector is computed (e.g., 39 parameters):

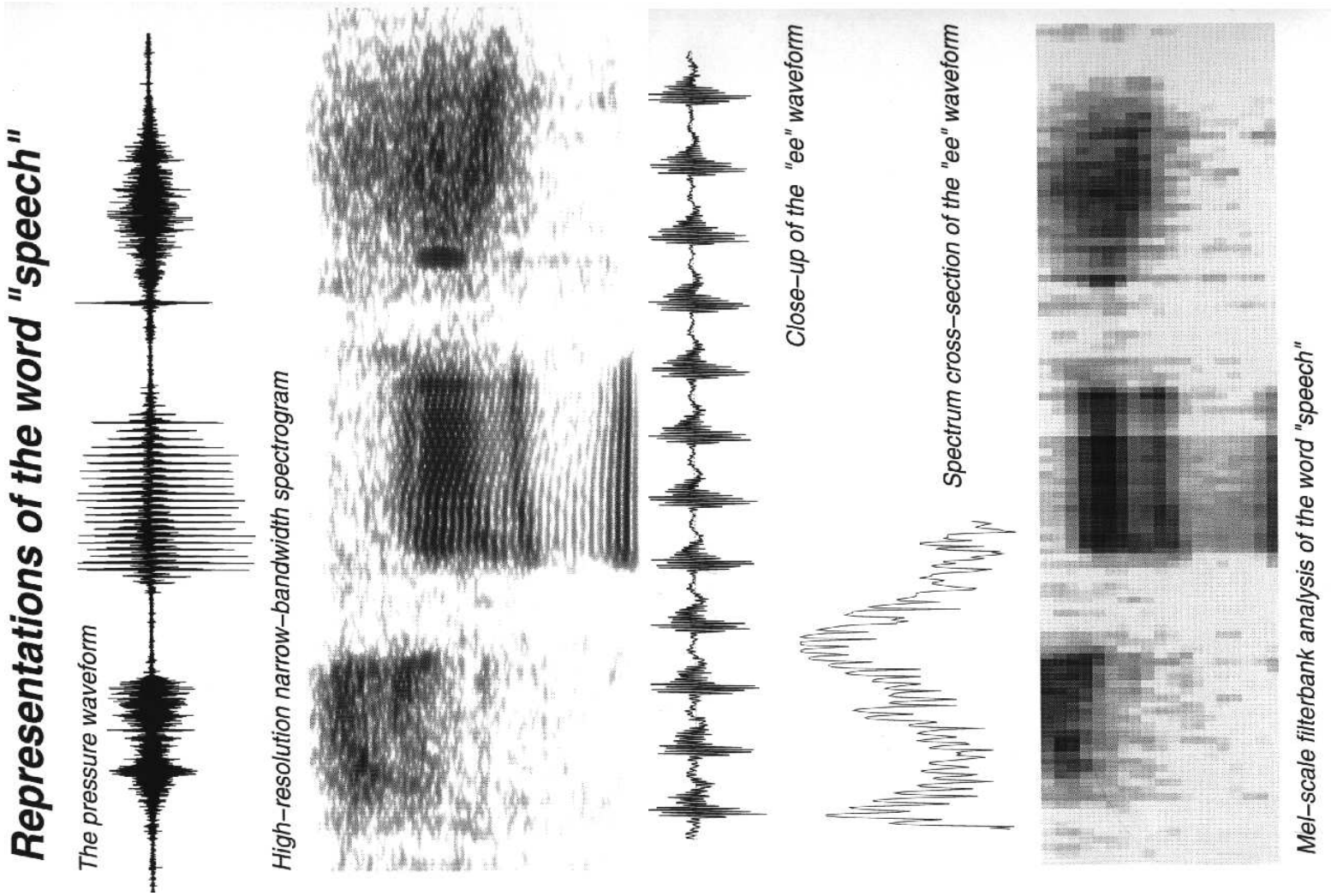
- 1. Parameter is the energy
- 12 Parameters are (often) Mel Frequency Cepstral Coefficients (MFCCs), computed from FFT or LP spectrum.
- 14.-26. Parameter: time derivative of each parameter (delta features)
- 37.-39. Parameter: time-acceleration of each parameter (delta-delta features)

Spektrogrammdarstellung

Folge breitbandiger/schmalbandiger Kurzzeitspektren

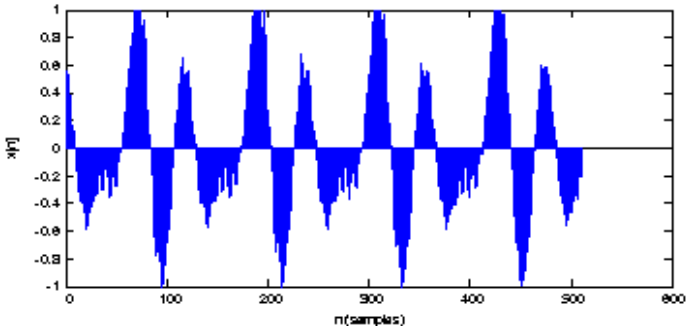


Feature extraction



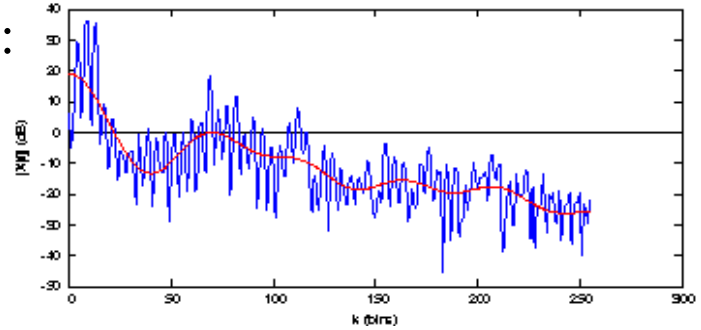
Feature extraction

Time domain:



x[n]

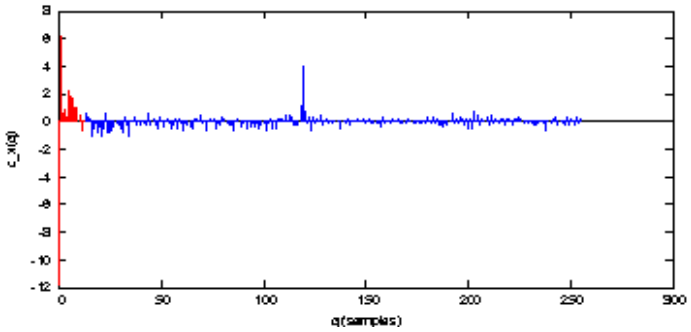
Spectral domain:



X[k] = FFT(x[n])

plotted: 20 log(|X[k]|)

Cepstral domain:



Xc[q] = IFFT(20 log(|X[k]|))

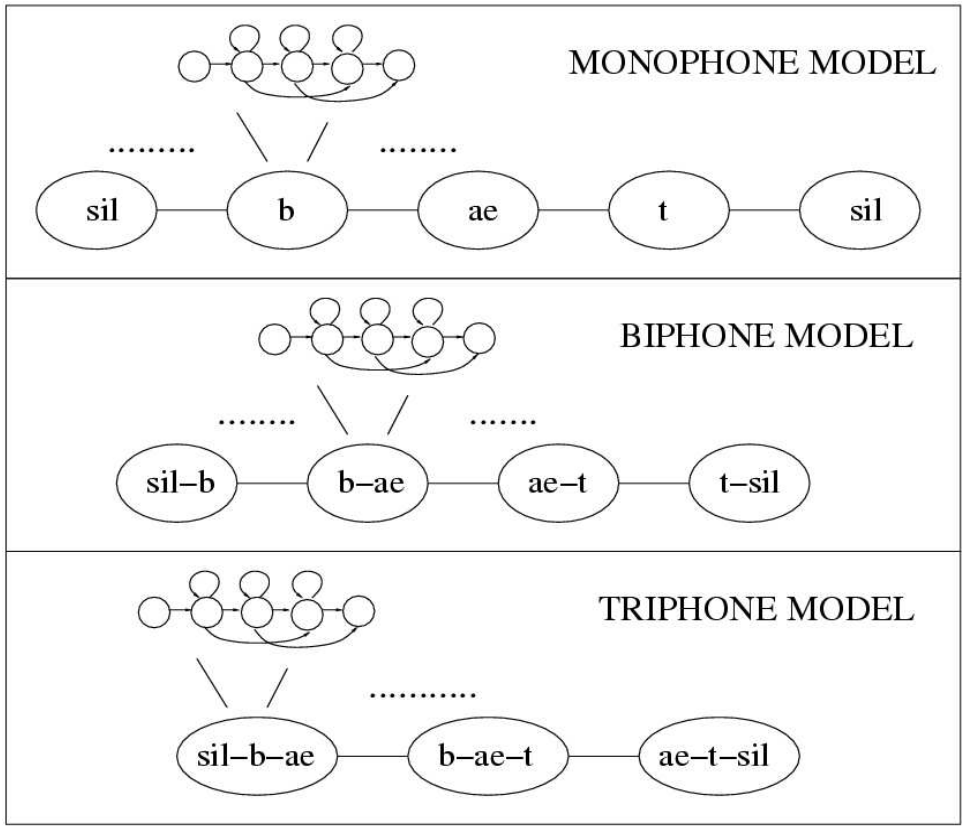
Feature extraction

- Small number of features
  - Relevant acoustic information
  - Robust to acoustic variation (fundamental frequency, pronuciation variants, speaker identity)
  - Robust against noise, etc.
  - Sensitive to linguistic content

- Cepstral features capture the spectral envelope
- Nonlinear “warping” of frequency axis ⇔ human auditory processing
- Delta features ⇔ variation in natural speech: “co-articulation”

Acoustic modeling

Phoneme models:

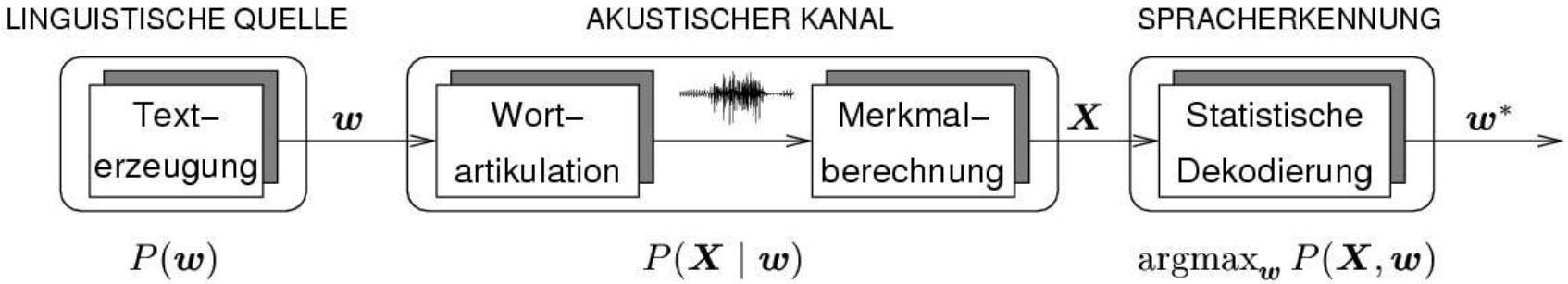


Pronunciation Dictionary:

Table 1: Extract from a dictionary

word	*pronunciation*
INCREASE	ih n k r iy s
INCREASED	ih n k r iy s t
INCREASES	ih n k r iy s ah z
INCREASING	ih n k r iy s ih ng
INCREASINGLY	ih n k r iy s ih ng l iy
INCREDIBLE	ih n k r eh d ah b ah l

Probabilistic model



Suche diejenige Wortfolge $\boldsymbol{w}$ mit maximaler *a posteriori* Wahrscheinlichkeit

$$P(\boldsymbol{w}|\boldsymbol{X}) \stackrel{\text{def}}{=} \frac{P(\boldsymbol{w}) \cdot P(\boldsymbol{X}|\boldsymbol{w})}{P(\boldsymbol{X})}$$

Dabei bezeichnen

- $\boldsymbol{w}$ eine Folge $w_1 ... w_m$ von Wörtern
- $\boldsymbol{X}$ eine geeignete Parametrisierung des Eingabesprachsignals
- $P(\boldsymbol{w})$ die Wahrscheinlichkeit des Auftretens einer Wortfolge
- $P(\boldsymbol{X}|\boldsymbol{w})$ die Wahrscheinlichkeit dafür, daß $\boldsymbol{w}$ durch ein Schallsignal $\boldsymbol{X}$ *realisiert* wird