Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Sprachverarbeitung I / 11 HS 2016

Spracherkennung: Merkmalsextraktion

Buch: Kapitel 11.7 und 11.8

Beat Pfister

Programm heute:

Vorlesung:
- Aufnehmen einer Äusserung
- Sprachmerkmale für die Spracherkennung
- Mel-Cepstrum als Sprachmerkmal

Übung:
- ⋆ Sprachmerkmale zur Lautunterscheidung

Spracherkennung

Aufgabe: Ermitteln der Aussage aus einem Sprachsignal

Annahme: Sprachsignal enthält eine (ganze) Äusserung!

Frage: Wie kommt ein Sprachsignal mit einer ganzen Äusserung in den Computer bzw. in den Spracherkenner?

$\longrightarrow$ Äusserung muss aus dem vom Benutzer aufgenommenen Signal detektiert werden!

Detektieren einer Äusserung

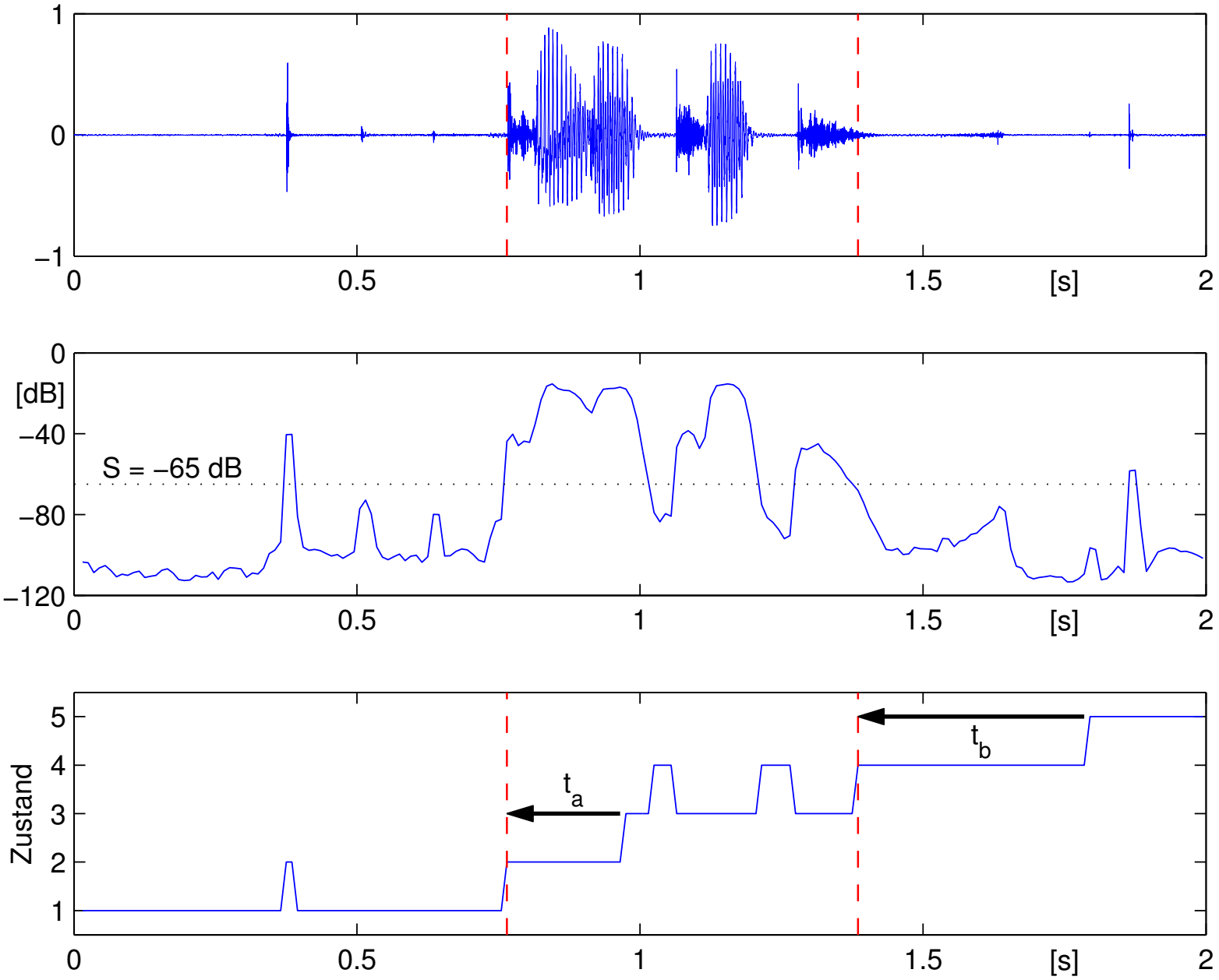
(Buch: Kapitel 11.8)

manuell:	Sprechtaste (push-to-talk)	– robust bzgl. Umgebungslärm – eine Hand muss frei sein – Äusserung wird oft abgeschnitten
halbautomatisch:	Starttaste (tap-and-talk)	– ziemlich robust – eine Hand muss frei sein
automatisch:	Anfangs- und Endpunktdetektion	

Automatische Anfangs- und Endpunktdetektion

Anforderungen bei Dialogsystemen:

- Flexible Antwortzeit des Benutzers
  $\longrightarrow$ Zeitfenster für Eingabe genügend lang
- Kurze Reaktionszeit des Systems
  $\longrightarrow$ Aufnahmedauer darf nicht fix sein
- Robust für leise oder kurze Störungen
  $\longrightarrow$ Geräuschpegel viel tiefer als Leistung des Sprachsignals
  $\longrightarrow$ starke Störungen kürzer als jede gültige Äusserung


1
0
−1
0
0.5
1
1.5
[s]
2
0
[dB]
−40
−80
−120
S = −65 dB
0
0.5
1
1.5
[s]
2
5
4
3
2
1
Zustand
t_a
t_b
0
0.5
1
1.5
[s]
2

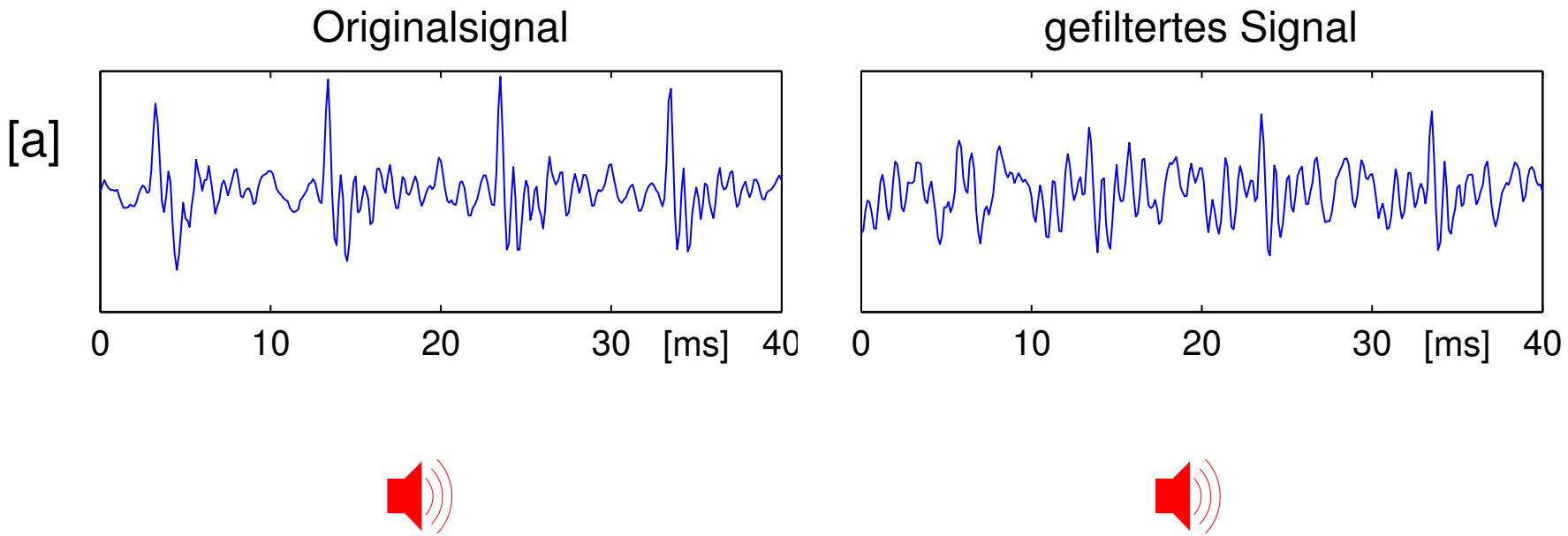
Spracherkennung

Resultat der Anfangs- und Endpunktdetektion:

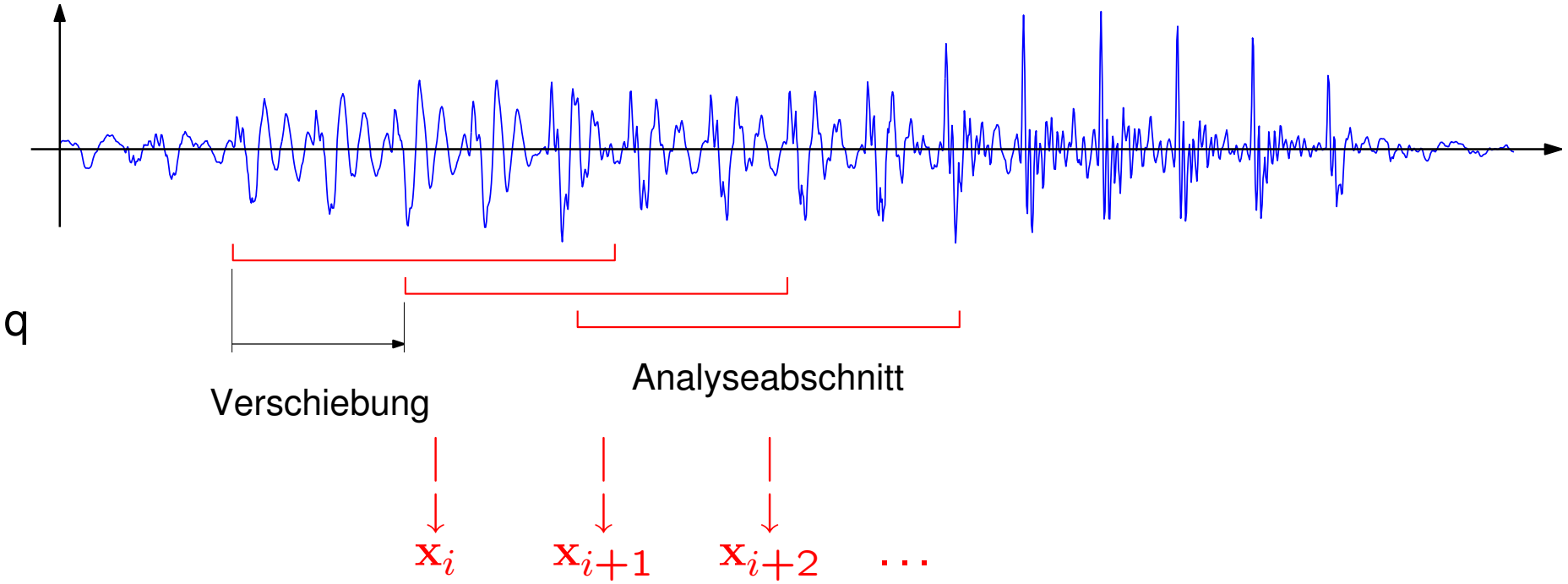
$\longrightarrow$ Sprachsignal mit einer Äusserung
Input des Spracherkenners

Sprachsignale im Zeitbereich

Signal kann nicht direkt für die Erkennung verwendet werden!



Merkmalsextraktion Kurzzeitanalyse des Sprachsignals



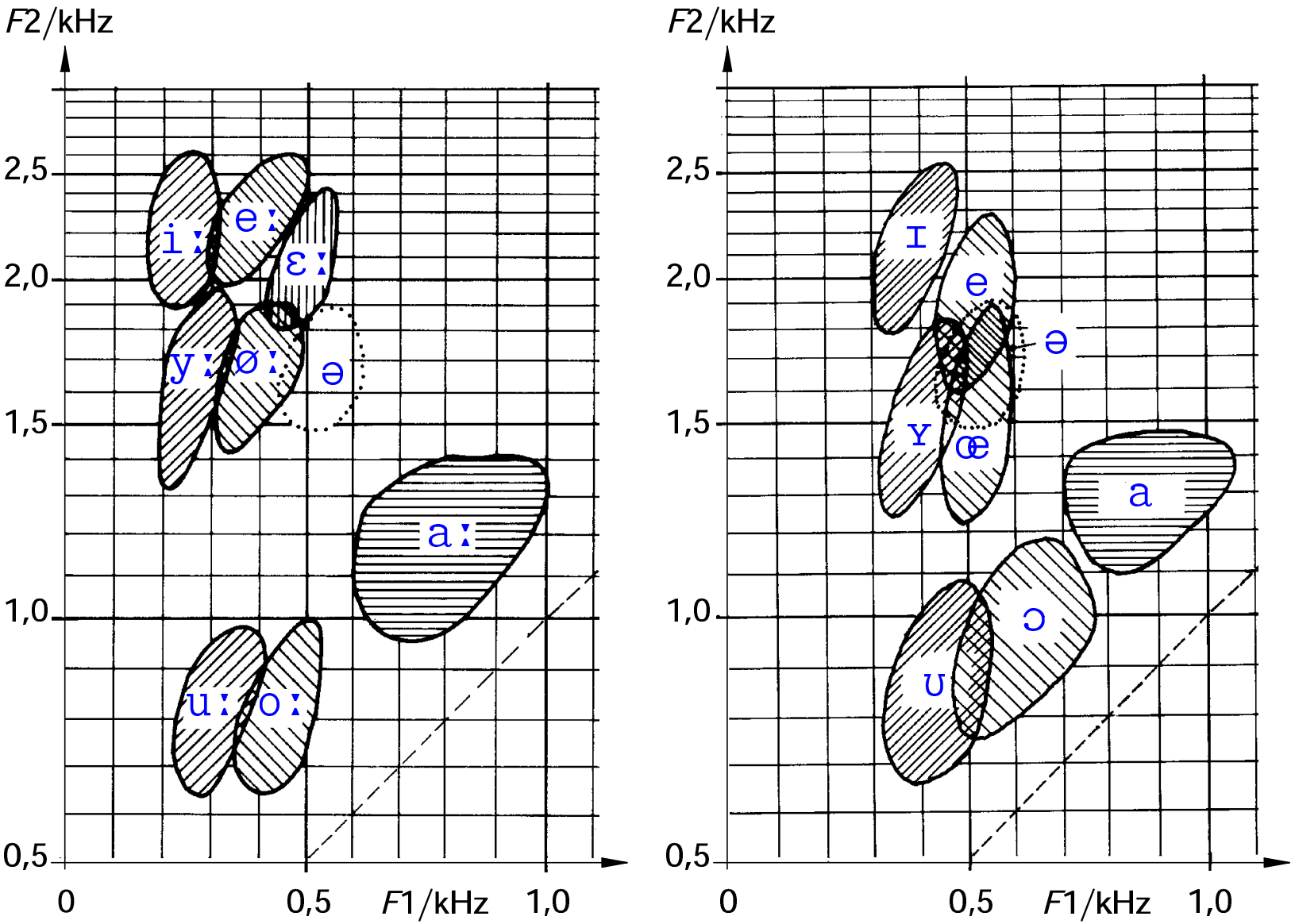
Merkmalssequenz: $\mathbf{X} = \mathbf{x}_1\, \mathbf{x}_2\, \mathbf{x}_3 \ldots \mathbf{x}_T$

Fragen

Welche Merkmale sind für die Spracherkennung geeignet?

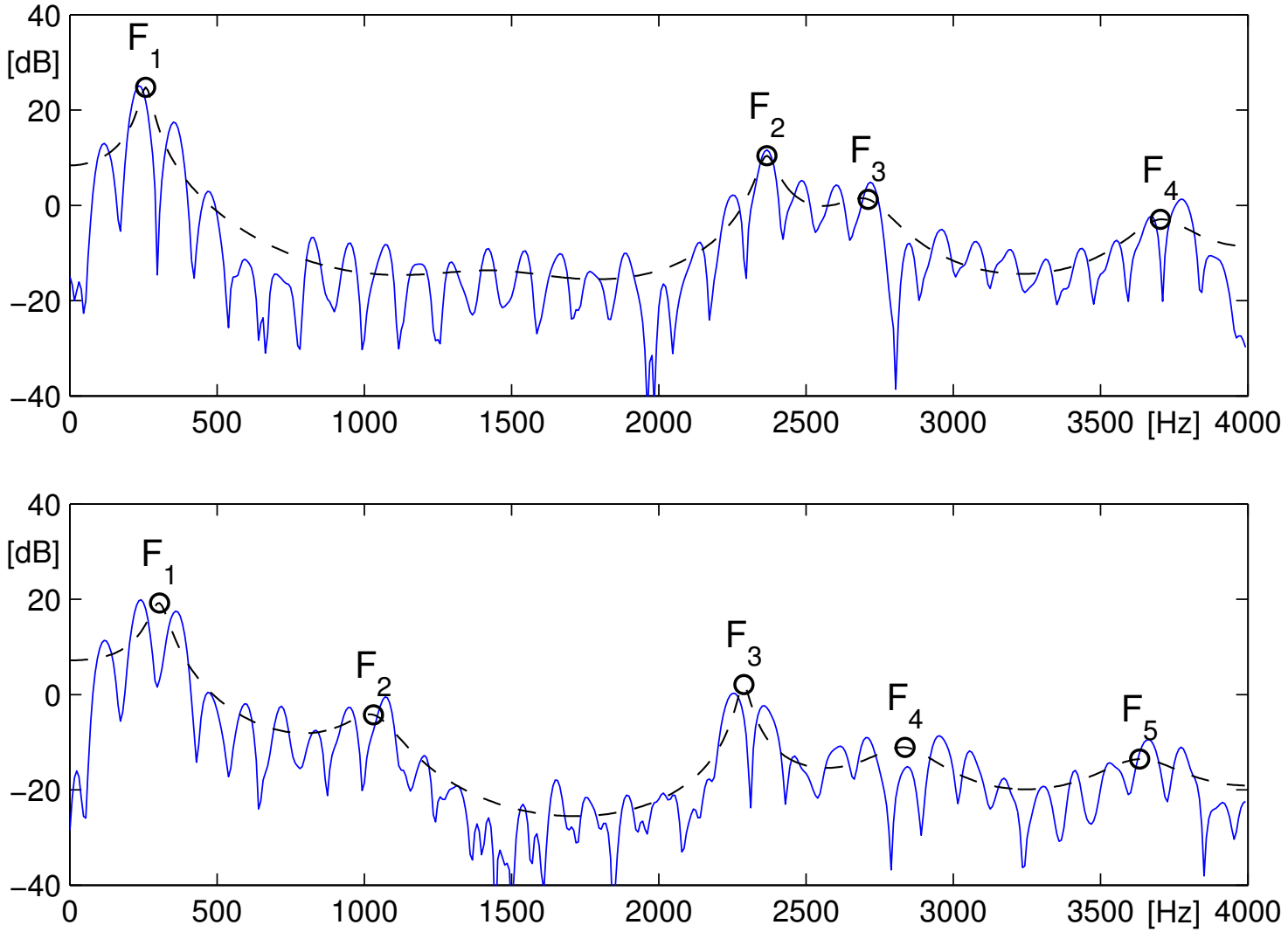
Gibt uns die Phonetik nützliche Hinweise?

W. Hess. *Grundlagen der Phonetik, 2. Deskriptive Phonetik*. Lehrskript, Institut für Kommunikationswissenschaften, Universität Bonn.



Formanten

Laut: [iː]



Anforderungen an Sprachmerkmale

Grundsatz: Ein Merkmal $\mathbf{x}$ soll Laute $L_1, L_2, L_3 \ldots$ unterscheiden

Problem: Sprachsignale eines Lautes L_i stark verschieden

Frage: Wie kann man ein Merkmal $\mathbf{x}$ beurteilen?

Kriterium: Eigendistanzen d_E vs. Kreuzdistanzen d_K

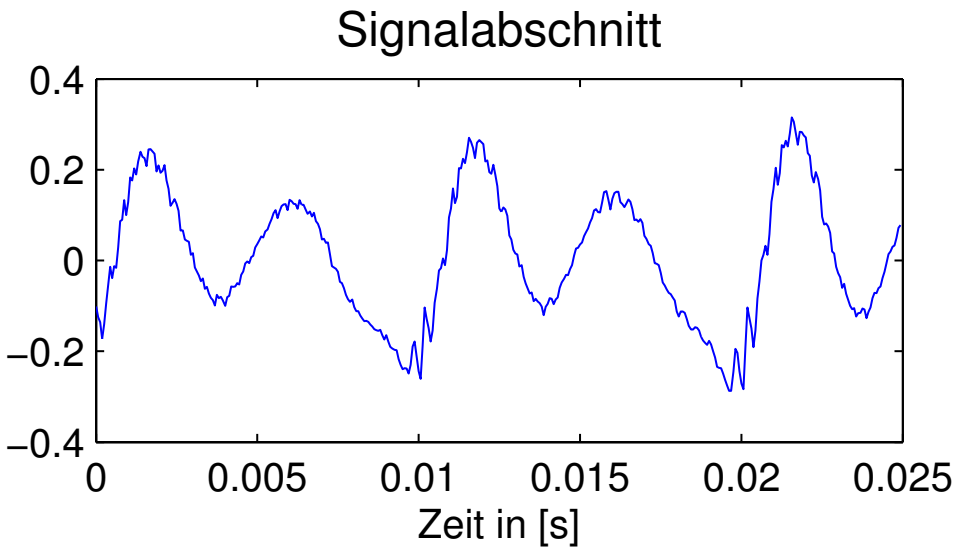
- $d_E = d(\mathbf{x}\{L_i\}, \mathbf{x}\{L_j\})$ mit $L_i = L_j$ $\longrightarrow$ klein
- $d_K = d(\mathbf{x}\{L_i\}, \mathbf{x}\{L_j\})$ mit $L_i \neq L_j$ $\longrightarrow$ gross

Merkmal für die Spracherkennung: MFCC

(MFCCs: mel frequency cepstral coefficients)

Ermittlung des Mel-Cepstrums:

Analyseabschnitt

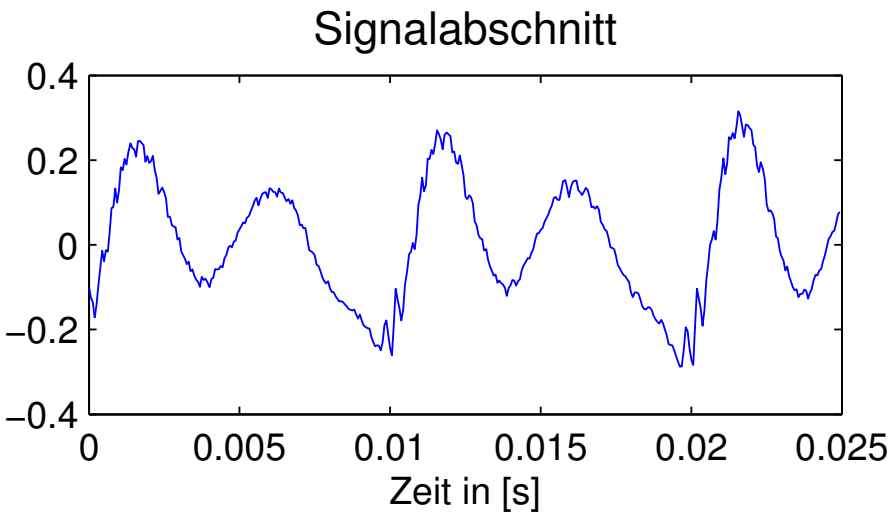


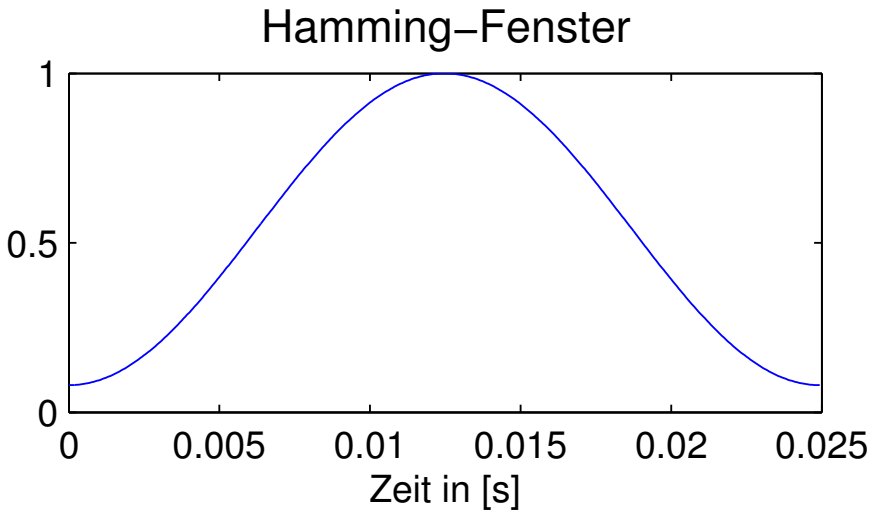
Merkmal für die Spracherkennung: MFCC

(MFCCs: mel frequency cepstral coefficients)

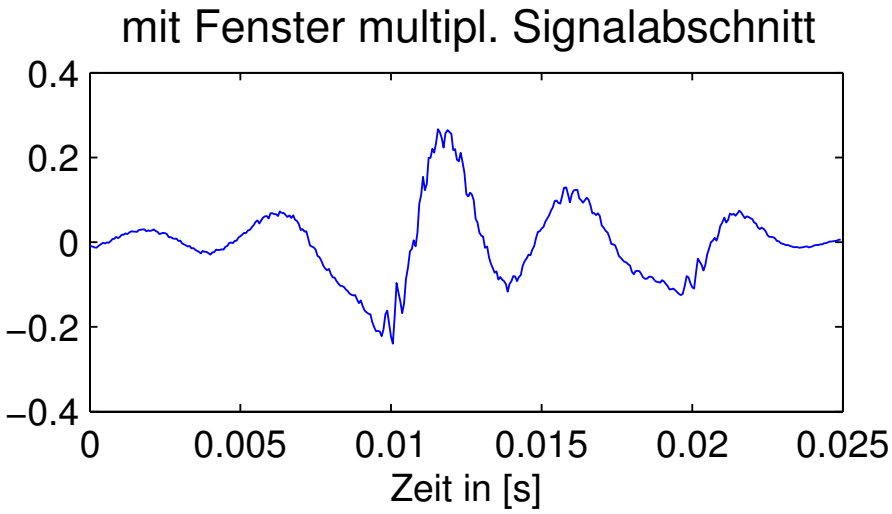
Ermittlung des Mel-Cepstrums:

Multiplikation mit Fensterfunktion





Signalabschnitt · Fensterfunktion



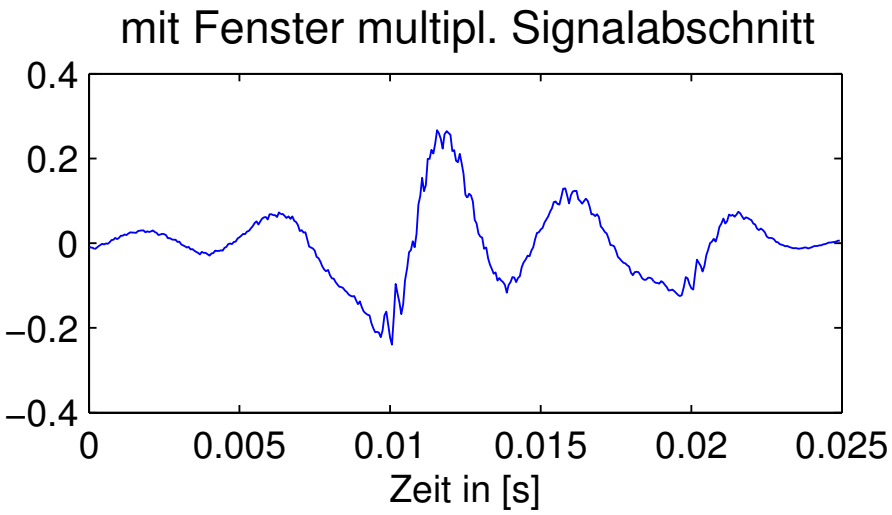
Merkmal für die Spracherkennung: MFCC

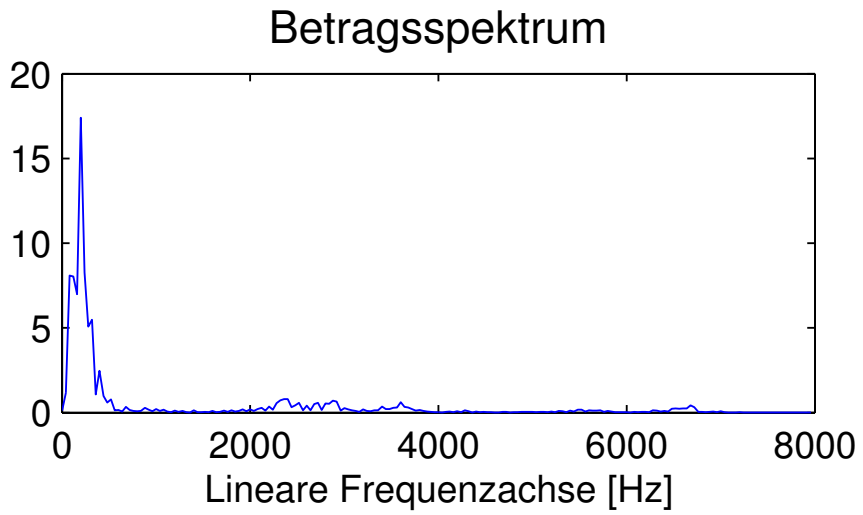
(MFCCs: mel frequency cepstral coefficients)

Ermittlung des Mel-Cepstrums:

Betragsspektrum

DFT $\longrightarrow$ Betrag



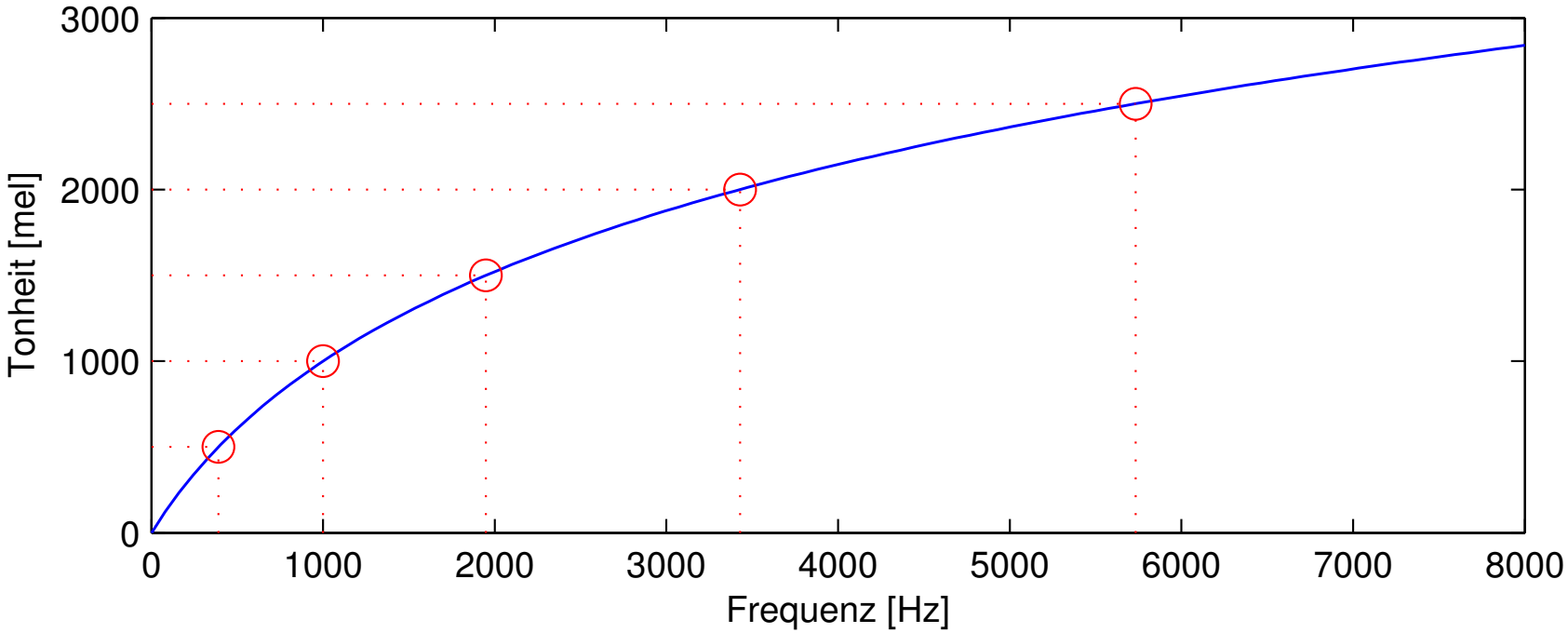


Merkmal für die Spracherkennung: MFCC

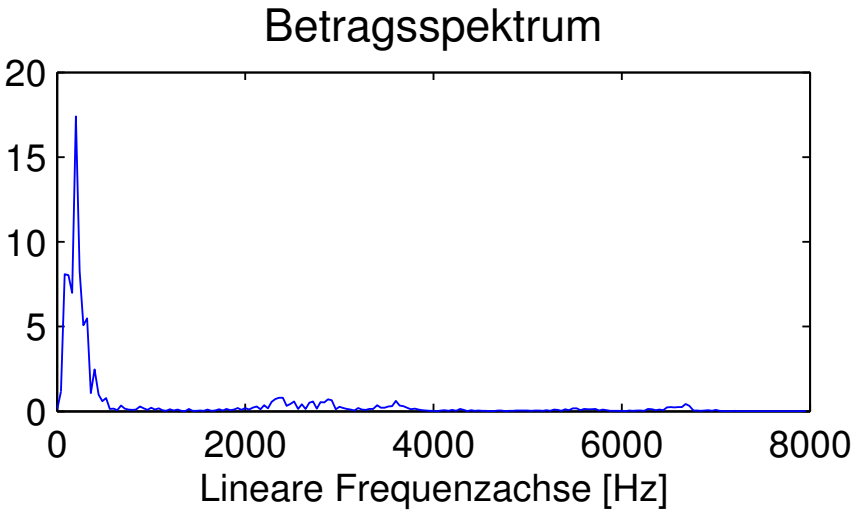
(MFCCs: mel frequency cepstral coefficients)

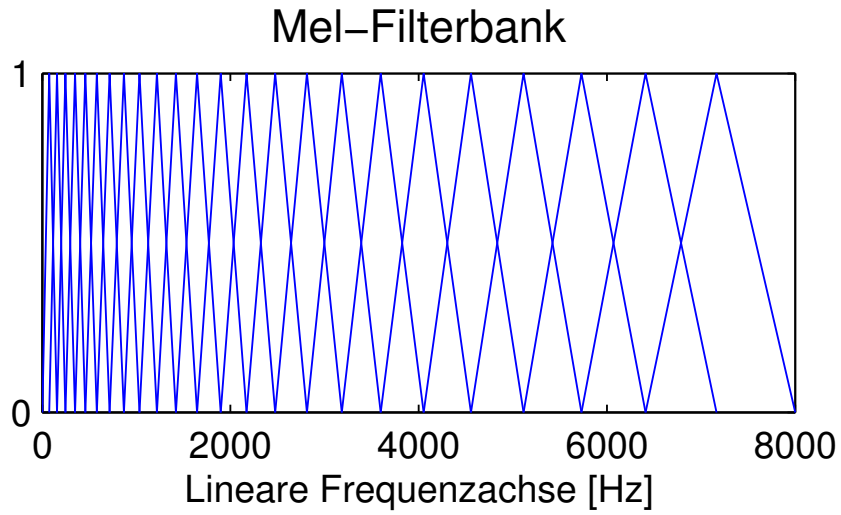
Ermittlung des Mel-Cepstrums:

Mel-Skala

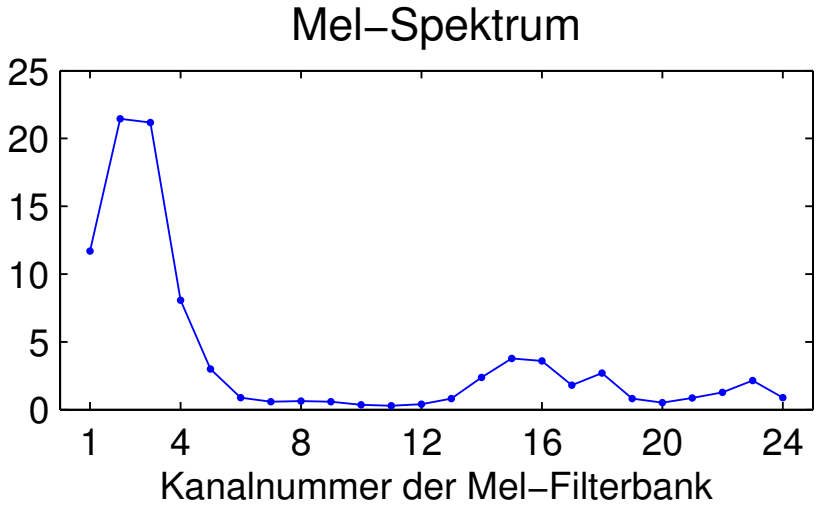

3000
2000
1000
0
Tonheit [mel]
0
1000
2000
3000
4000
5000
6000
7000
8000
Frequenz [Hz]

Mel-Spektrum





Betragsspektrum · Mel-Filterbank



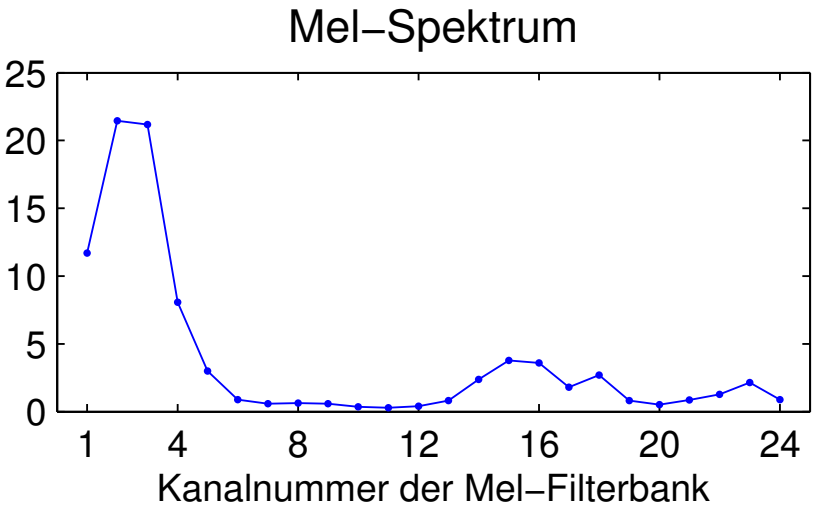
Merkmal für die Spracherkennung: MFCC

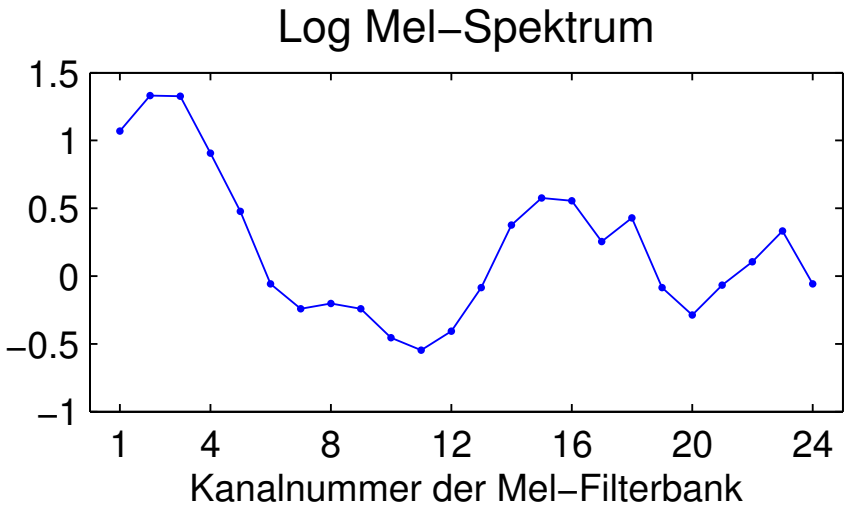
(MFCCs: mel frequency cepstral coefficients)

Ermittlung des Mel-Cepstrums:

Log-Mel-Spektrum

Mel-Spektrum $\rightarrow$ Log





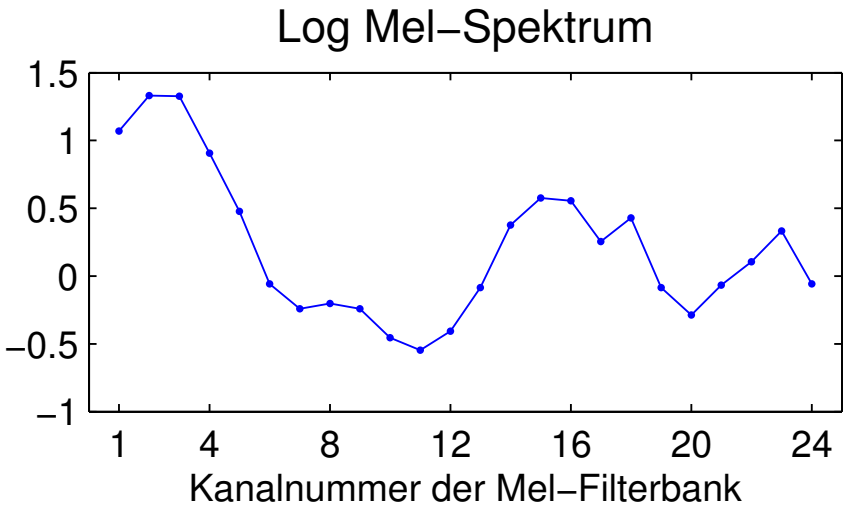
Merkmal für die Spracherkennung: MFCC

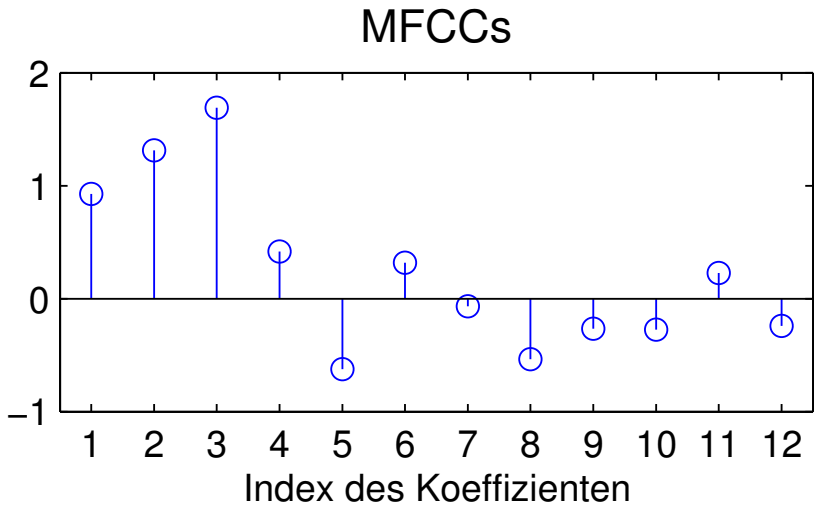
(MFCCs: mel frequency cepstral coefficients)

Ermittlung des Mel-Cepstrums:

Mel-Cepstrum

Log-Mel-Spektrum $\longrightarrow$ DCT





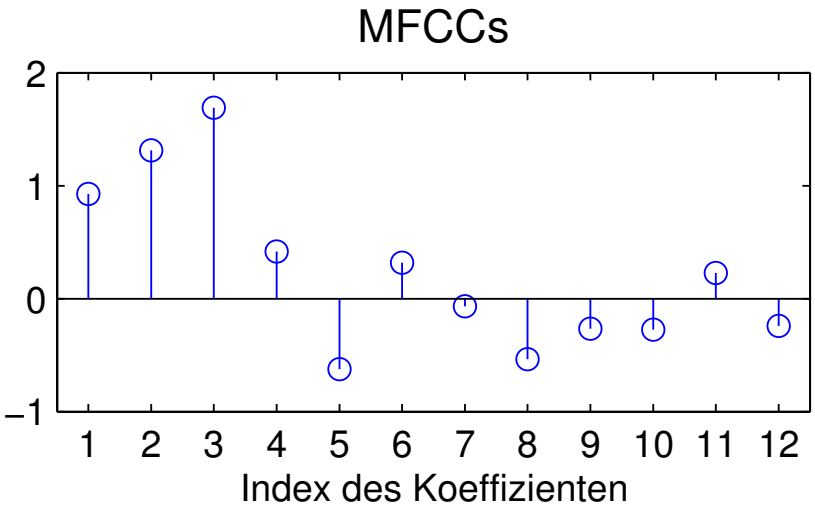
Merkmal für die Spracherkennung: MFCC

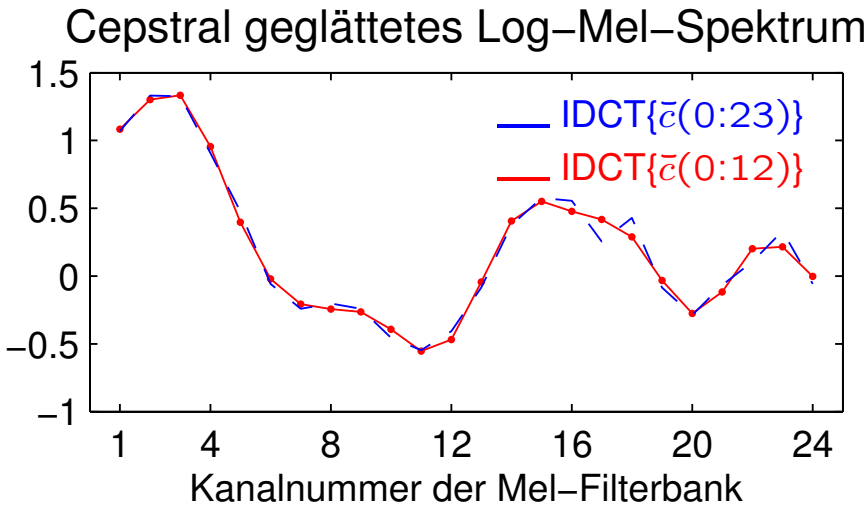
(MFCCs: mel frequency cepstral coefficients)

Ermittlung des Mel-Cepstrums:

Was beschreiben die MFCCs?

IDCT(MFCCs) >>>



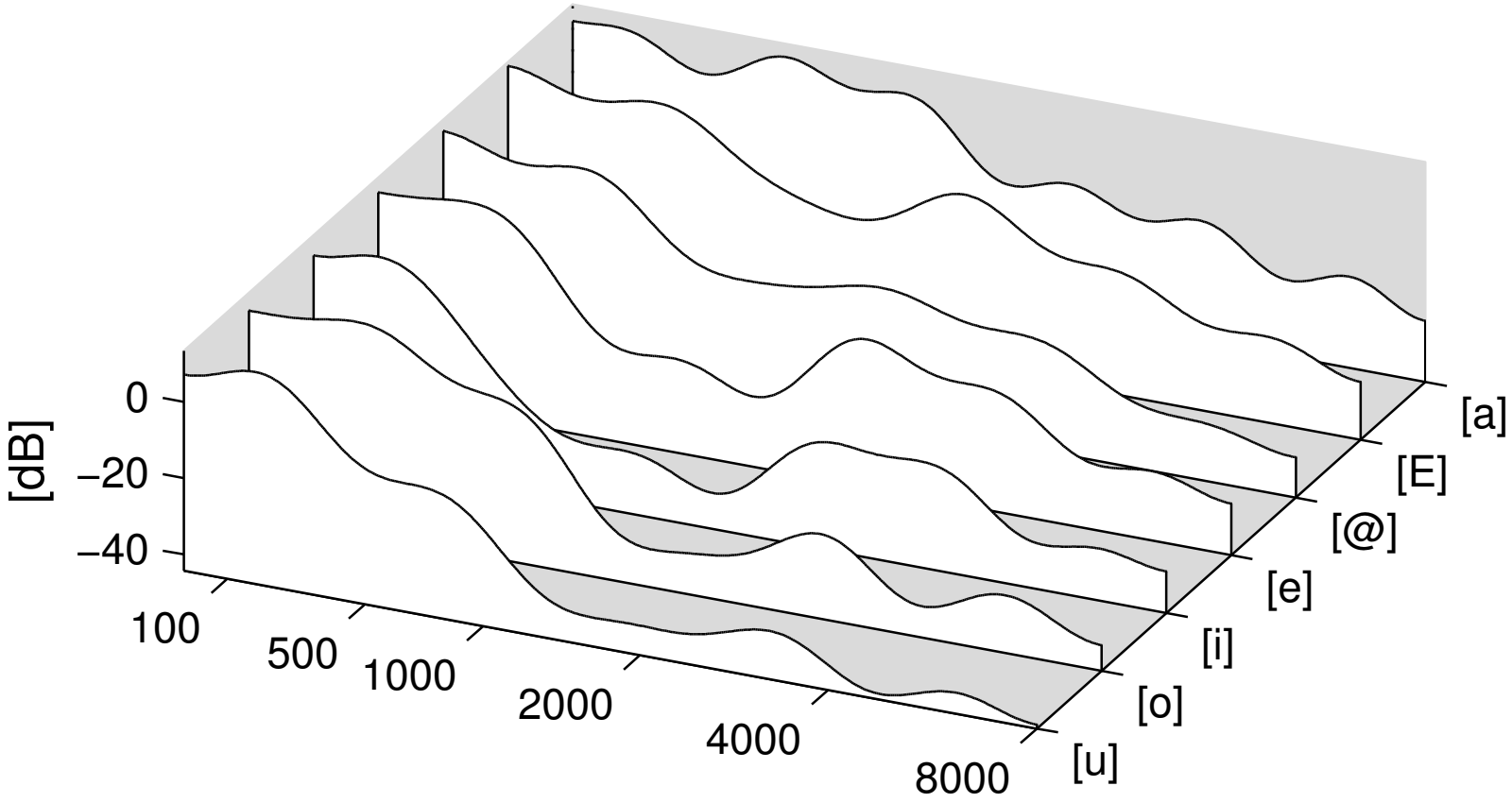


Der Merkmalsvektor repräsentiert das cepstral geglättete Log-Mel-Spektrum

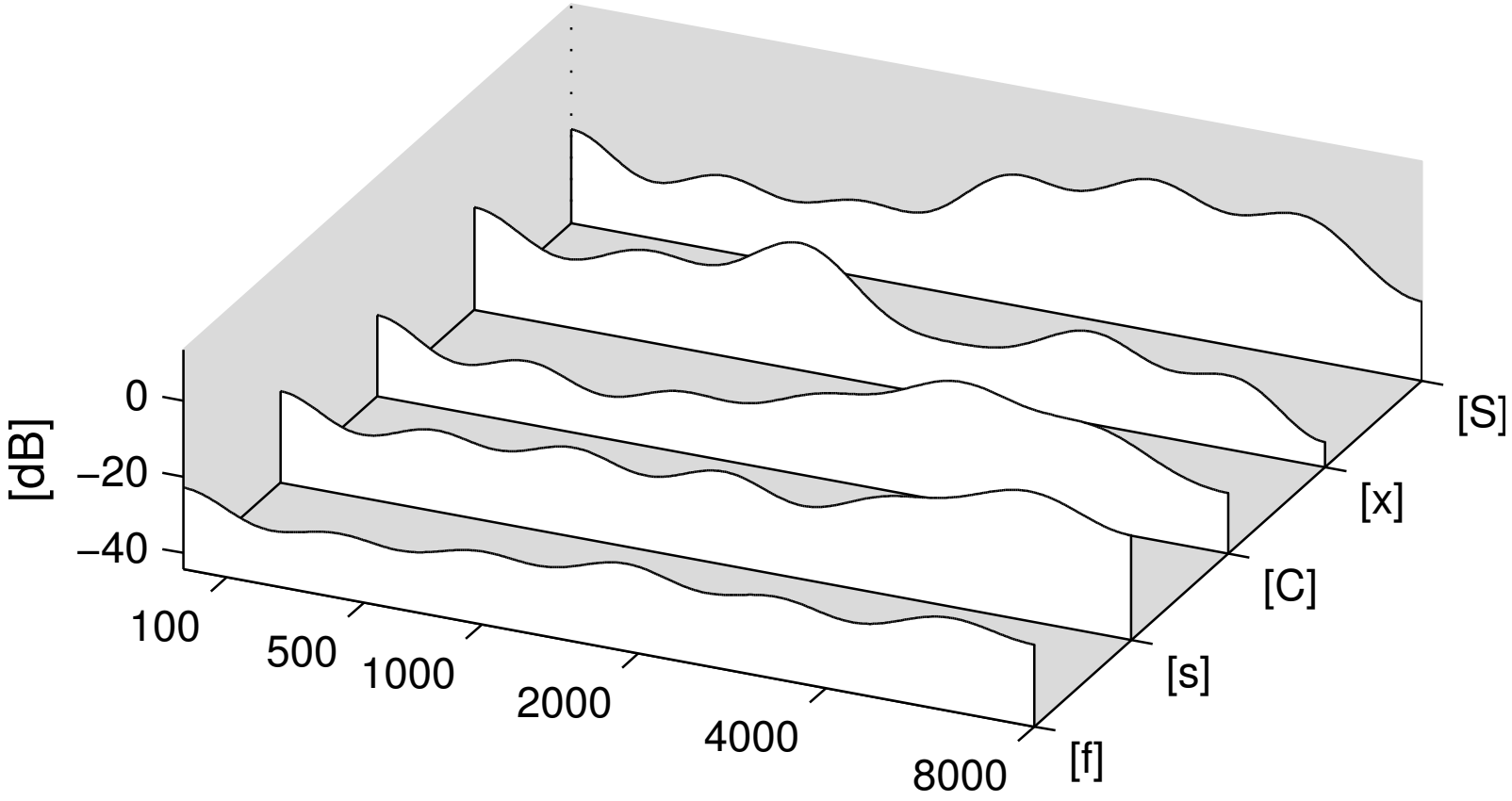
Wahl der Parameter

Abtastfrequenz	8/16 kHz
Länge des Analysefensters	25 ms
Verschiebung des Analysefensters	10 ms
Anzahl Dreiecksfilter	24
Cepstrale Koeffizienten	$\bar{c}(0) - \bar{c}(12)$

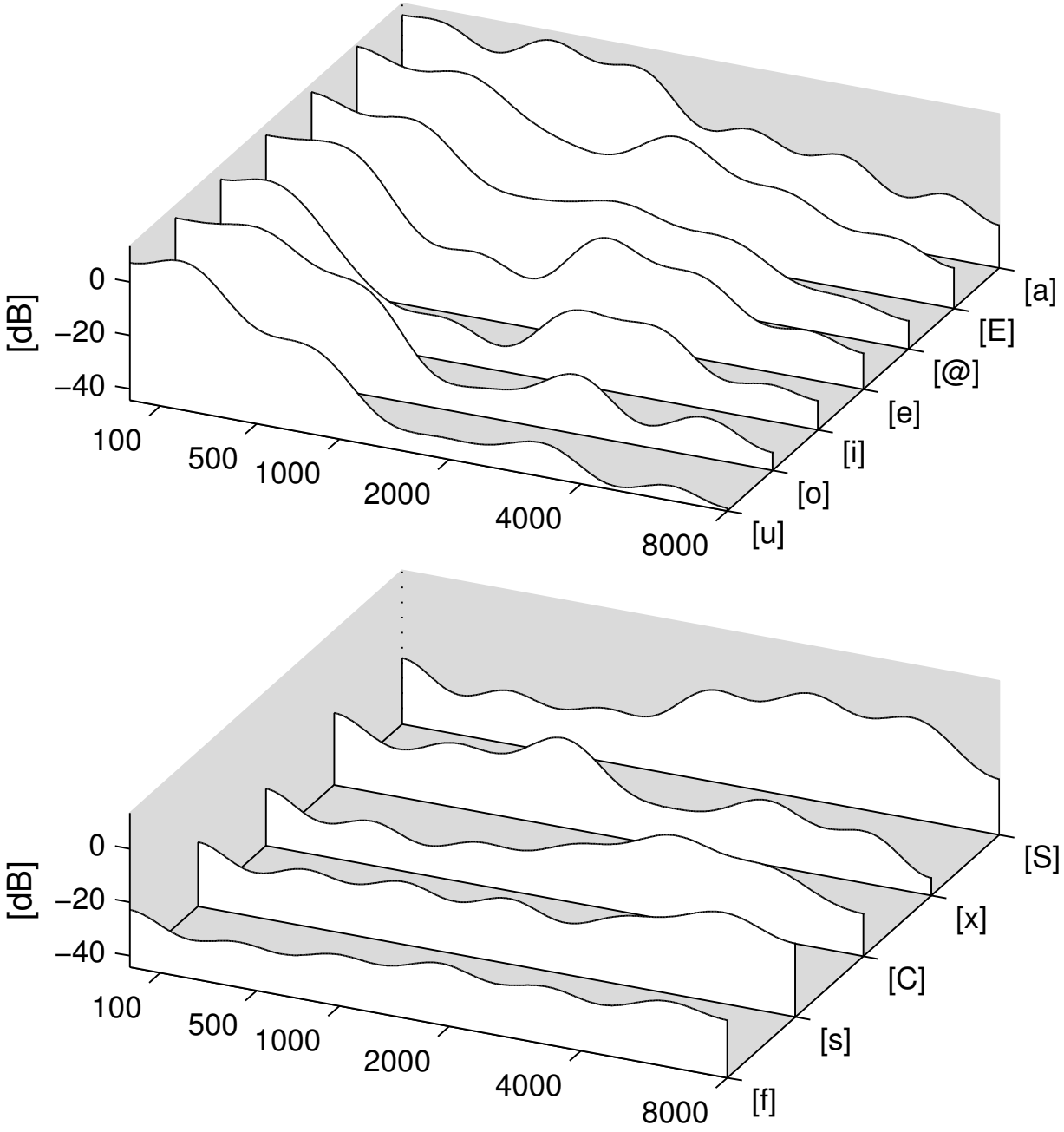
Vokale



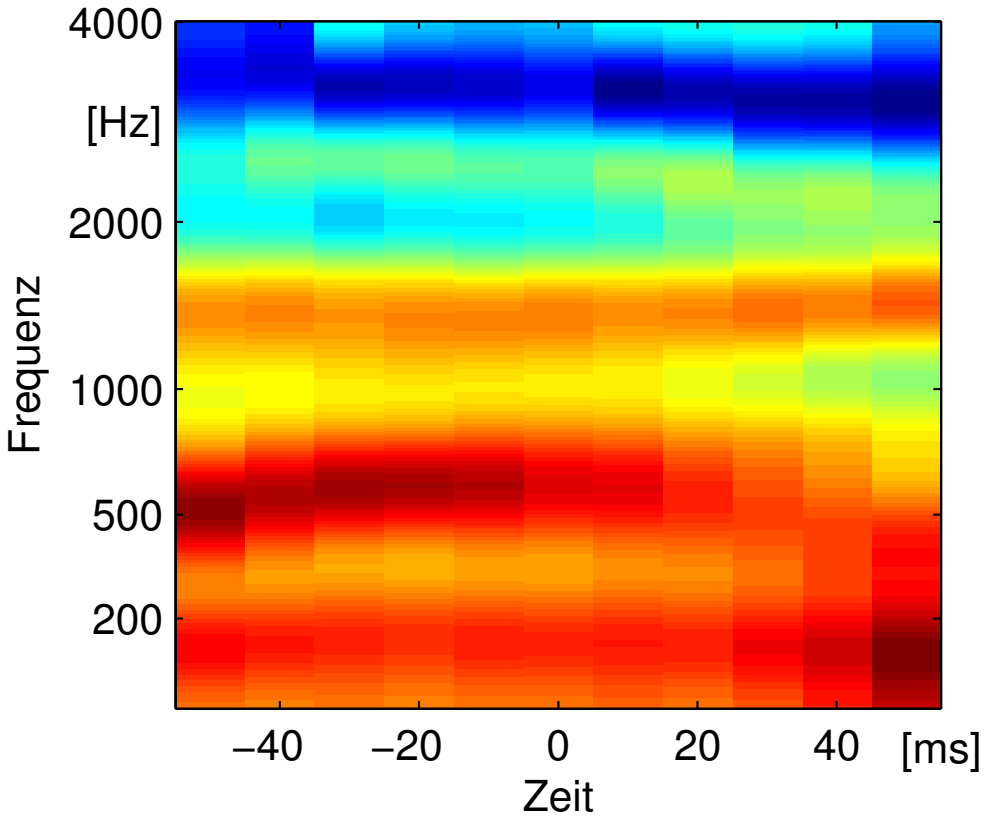
Frikative



Vokale und Frikative im Vergleich



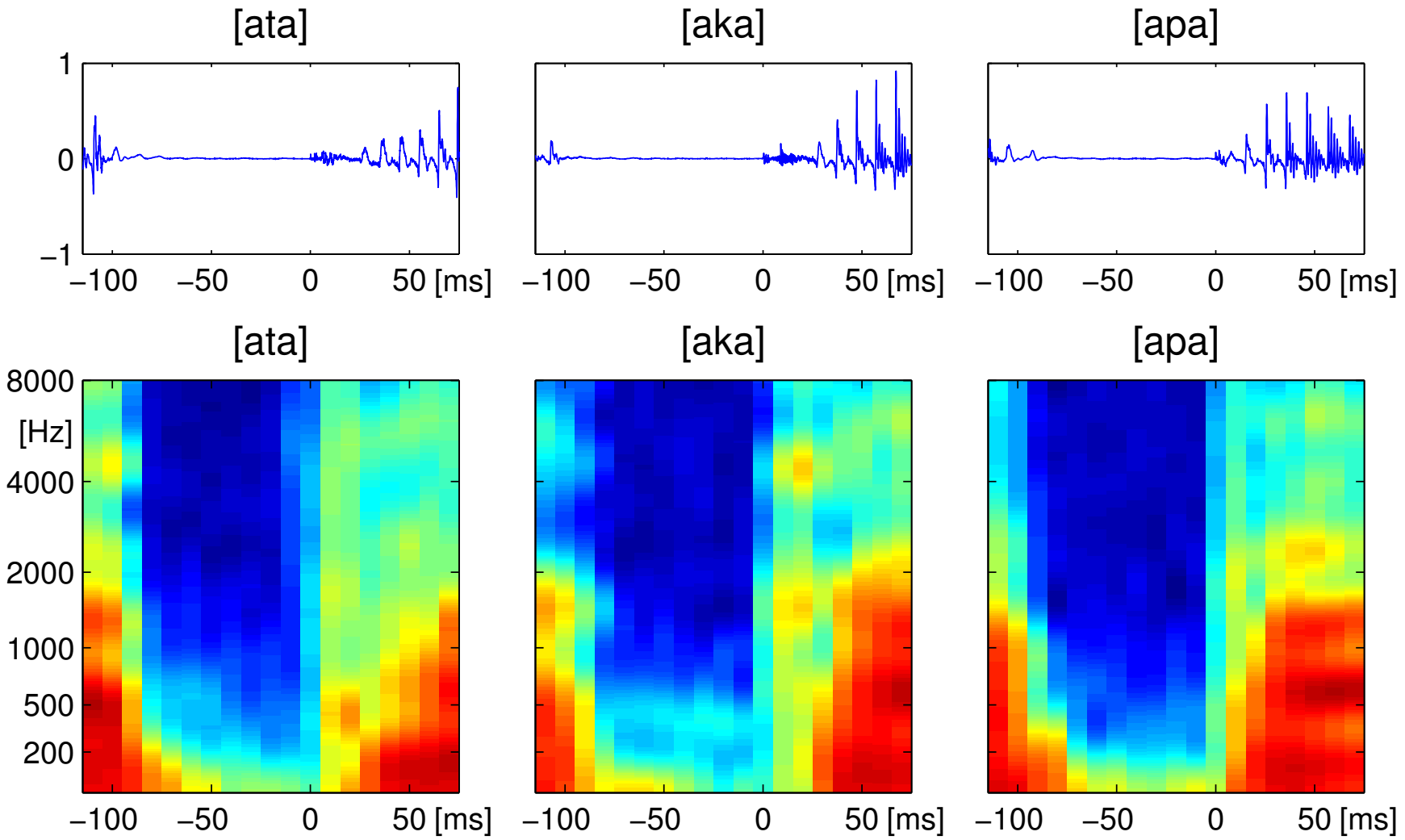
Zeitlicher Verlauf des Spektrums



Donnerstag [a]

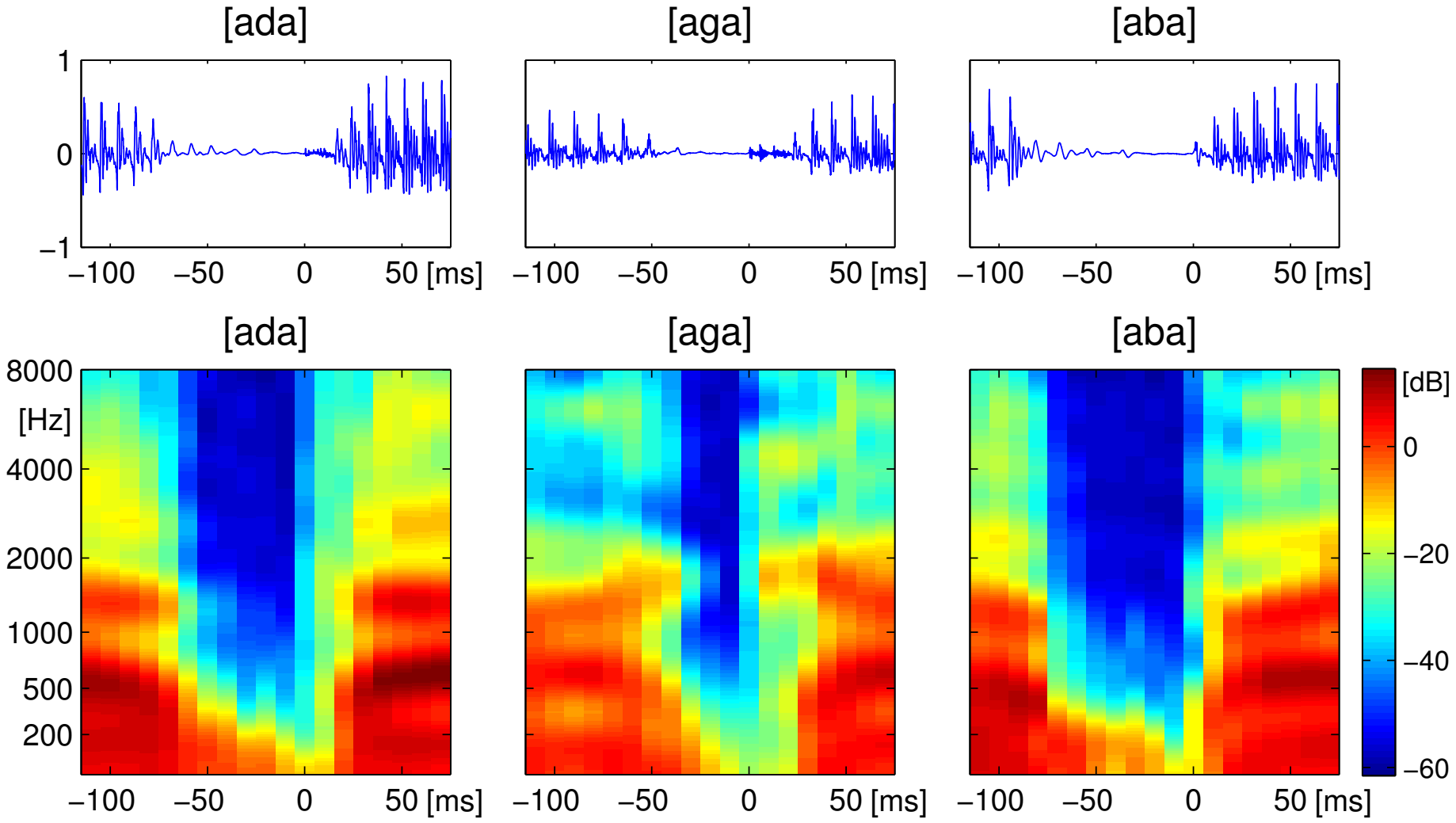
Stimmlose Plosive

Verschlusslaute [t] [k] [p]

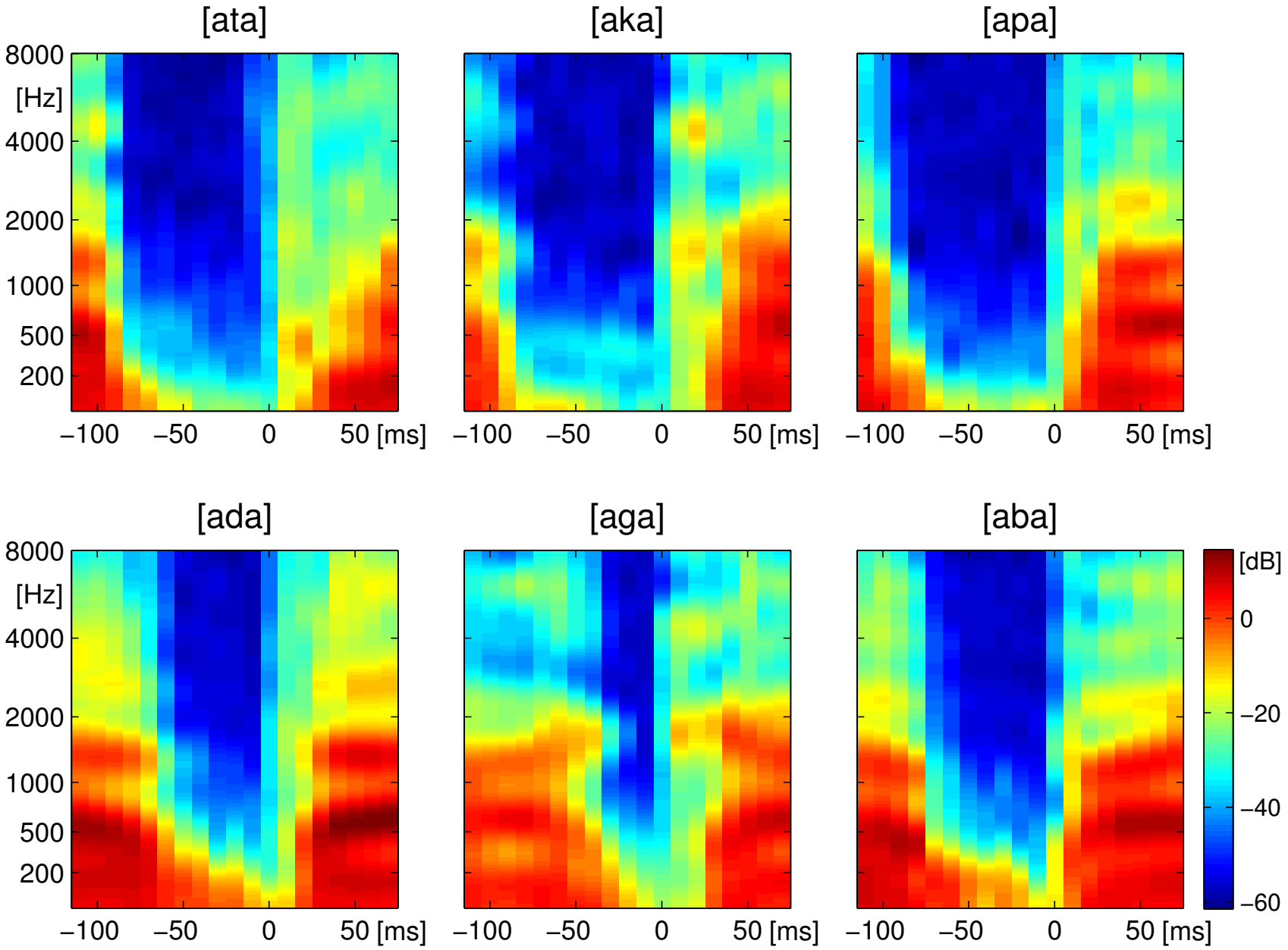


Stimmhafte Plosive

Verschlusslaute [d] [g] [b]



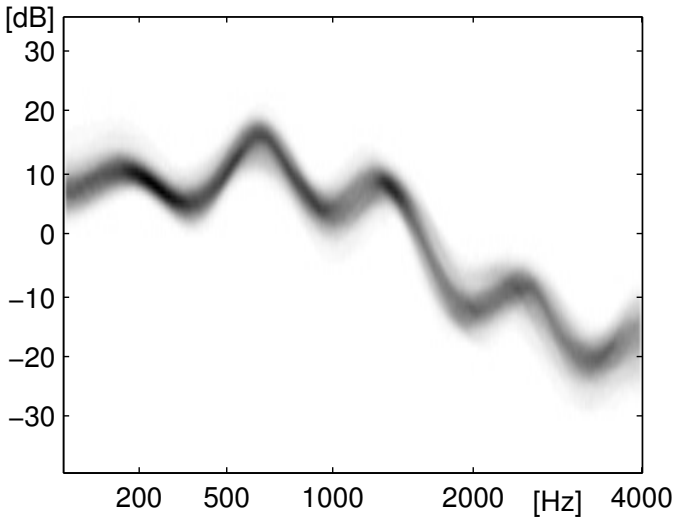
Plosive


[ata]
[aka]
[apa]
[ada]
[aga]
[aba]
8000
[Hz]
4000
2000
1000
500
200
−100
−50
0
50 [ms]
[dB]
0
−20
−40
−60

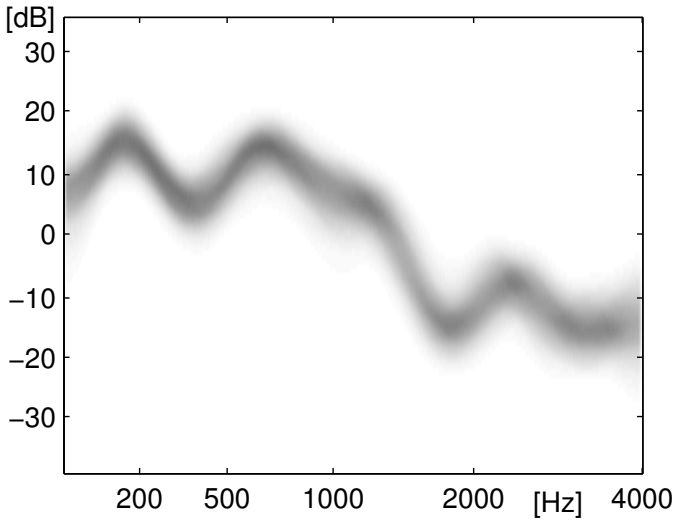
Spektrale Variabilität

Laut [aː]

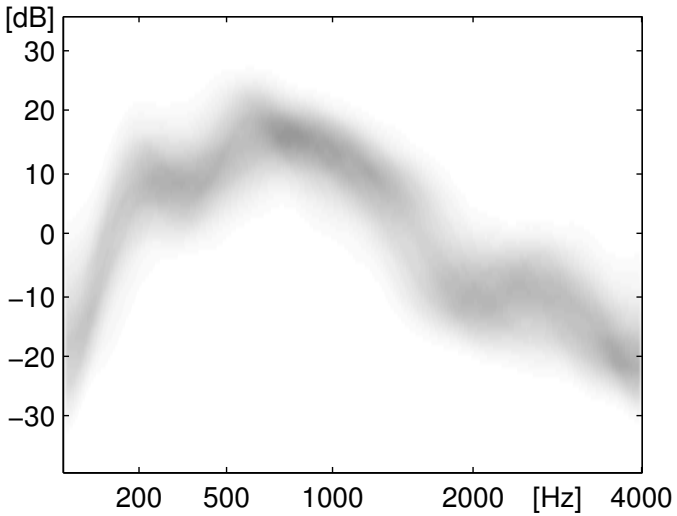
professioneller Sprecher



normaler Sprecher

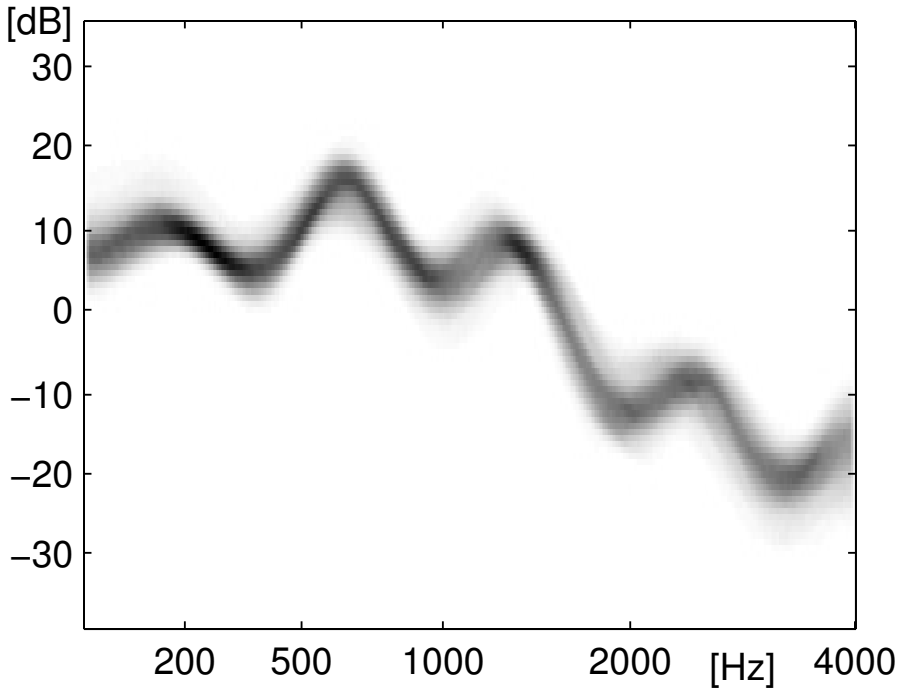


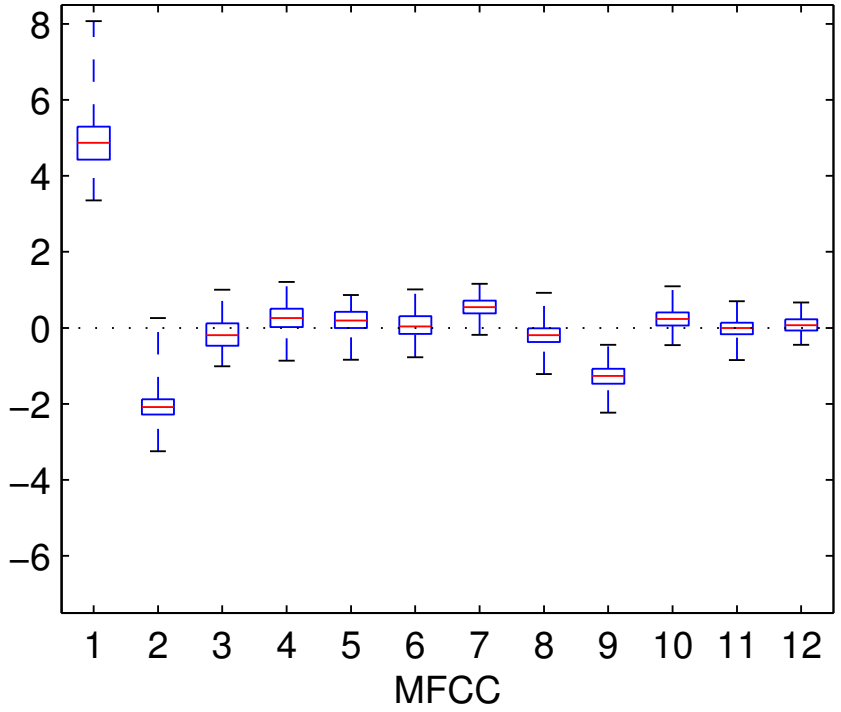
1000 Sprecher



Spektrum vs. Cepstrum

professioneller Sprecher

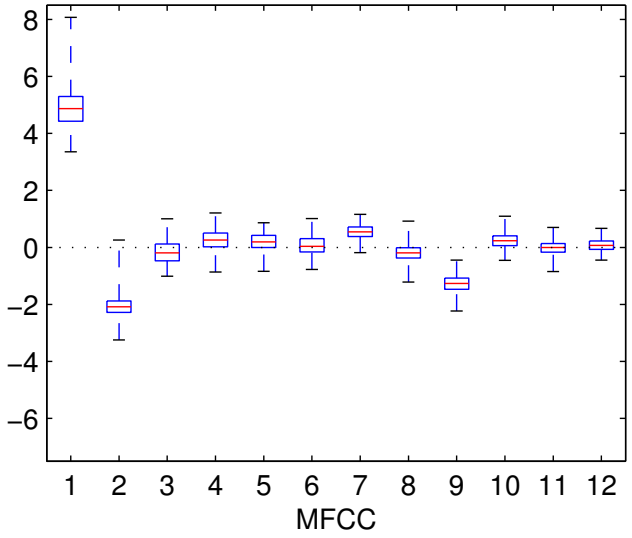




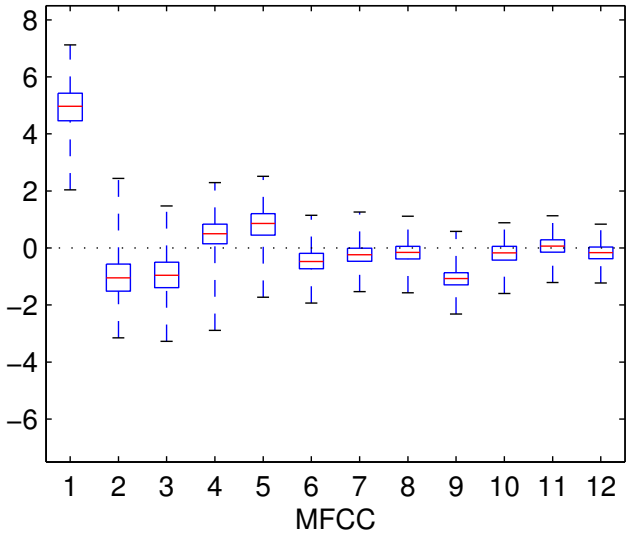
Cepstrale Variabilität

Laut [aː]

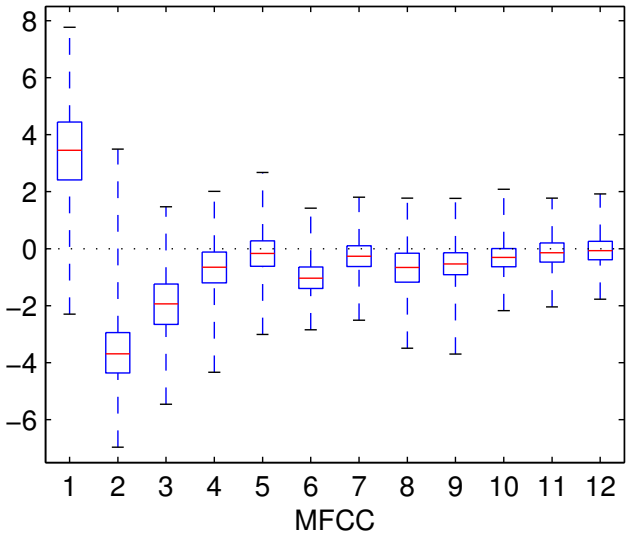
professioneller Sprecher



normaler Sprecher



1000 Sprecher



Welche Eigenschaften des Sprachsignals sind in den MFCC nicht enthalten?

Welche Eigenschaften des Sprachsignals sind nicht in den MFCC enthalten?

- die Phase
- die Feinstruktur des Spektrums
- die Grundfrequenz
- die Periodizität (stimmhaft/stimmlos)

Rekonstruktion des Signals aus den MFCC

Rekonstruktion 

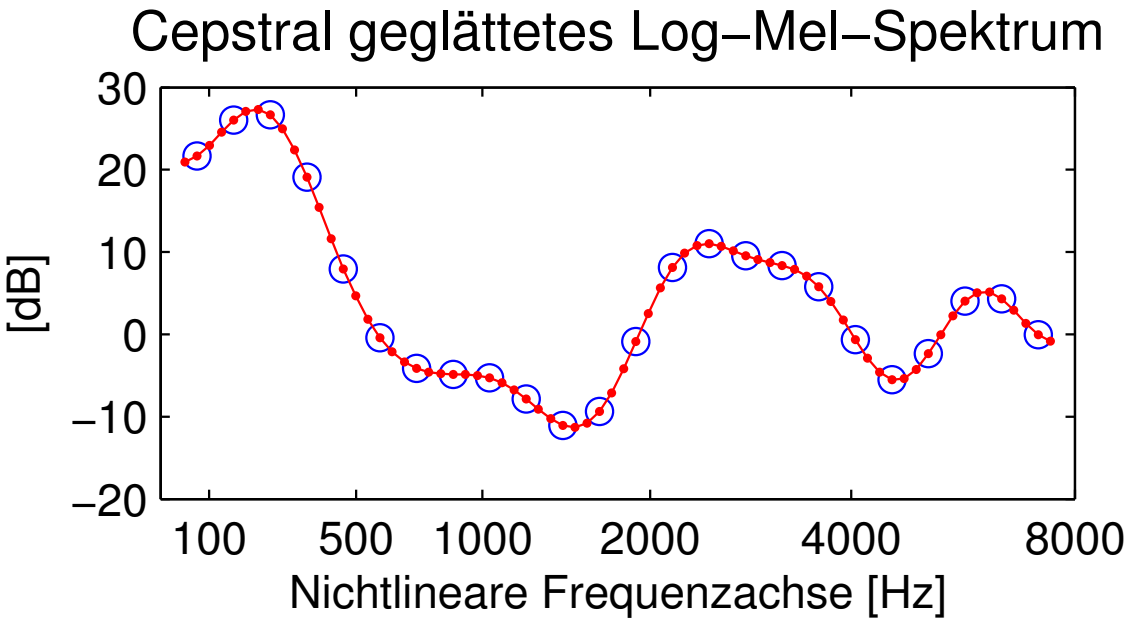
Originalsignal 

Thema der nächsten Lektion:

Spracherkennung mittels Mustervergleich

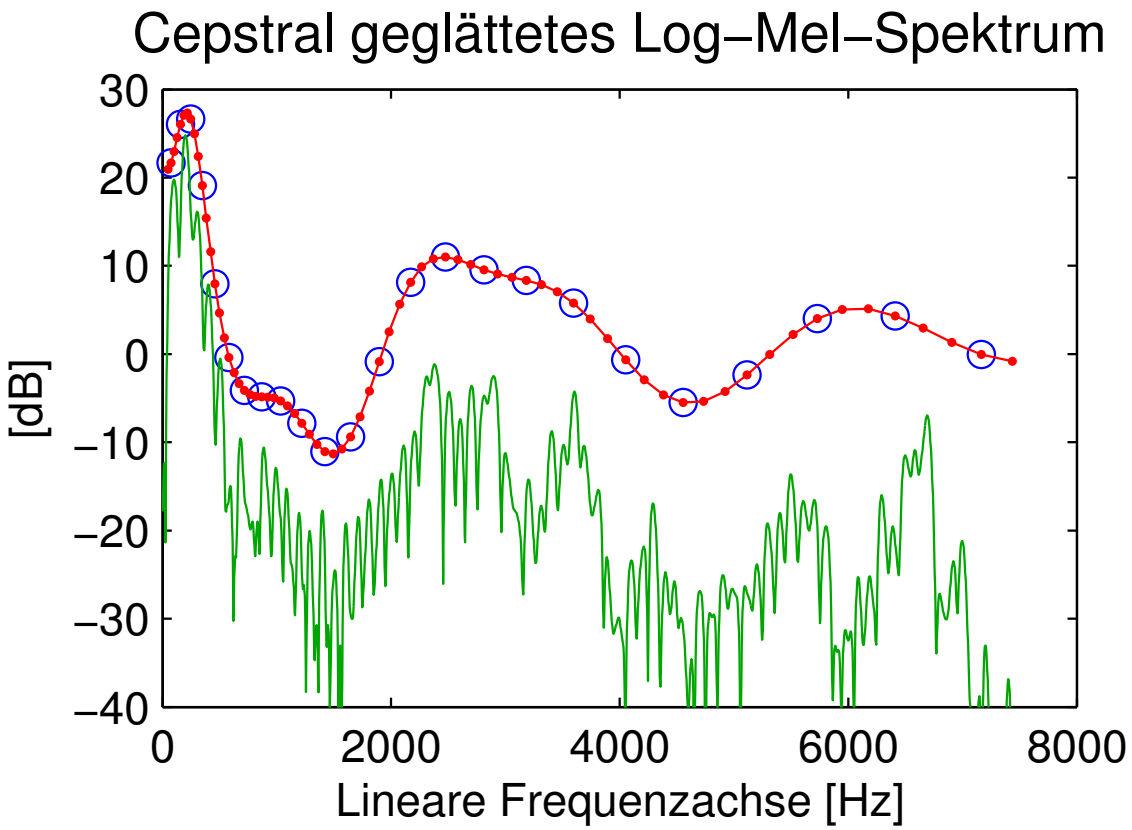
Zur Übersicht der Vorlesung *Sprachverarbeitung I* >>>

Rücktransformation: MFCC $\longrightarrow$ Log-Mel-Spektrum



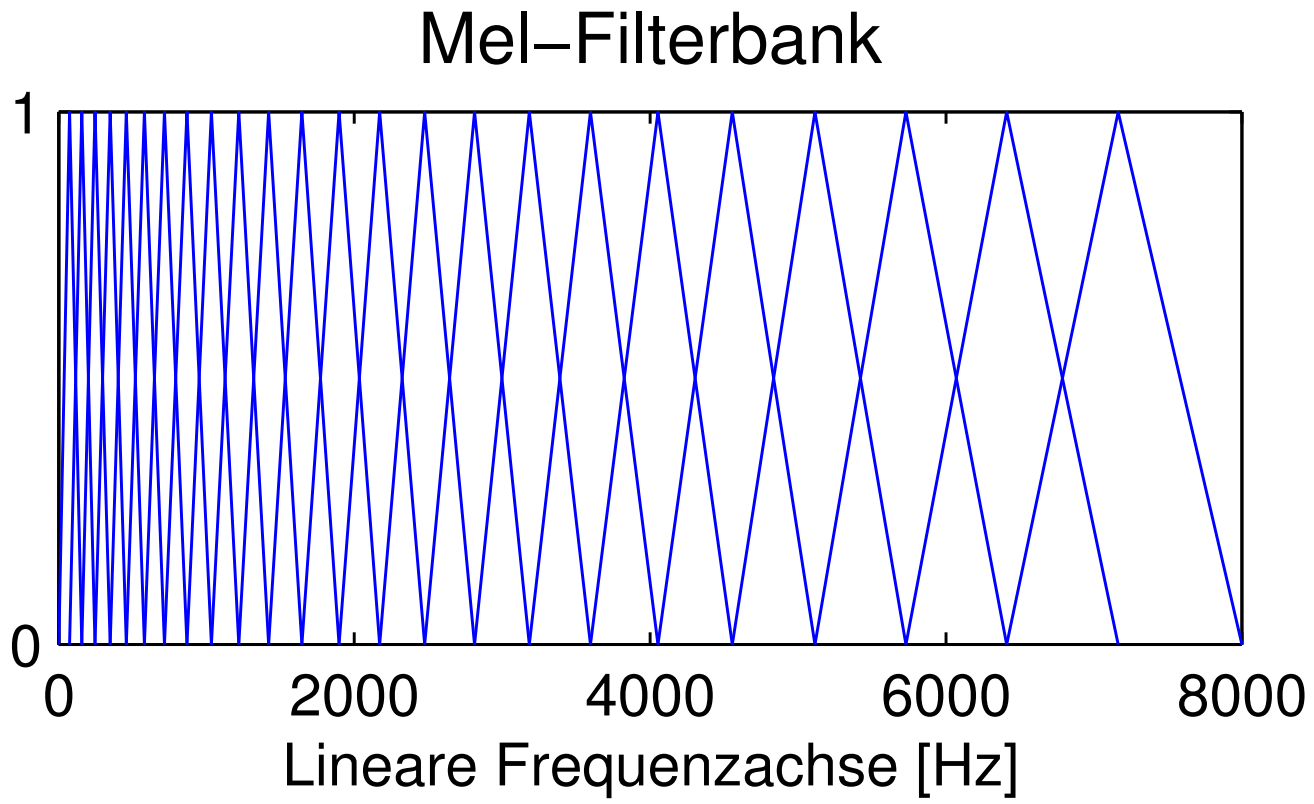
- Skala der Achse in Hertz (nicht linear)
- Interpolation (IDCT mit Zero-Padding)
- Vergleich mit Betragsspektrum des Signals?

Vergleich mit Betragsspektrum des Signals



Frage: Woher rührt der Unterschied zwischen den Spektren?

<<<


Mel−Filterbank
1
0
0
2000
4000
6000
8000
Lineare Frequenzachse [Hz]