Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Sprachverarbeitung I / 2 HS 2016

Sprachsignale: Verarbeitung und Darstellung

Buch: Kapitel 2 und 3

Beat Pfister

Sprachverarbeitung I / 2

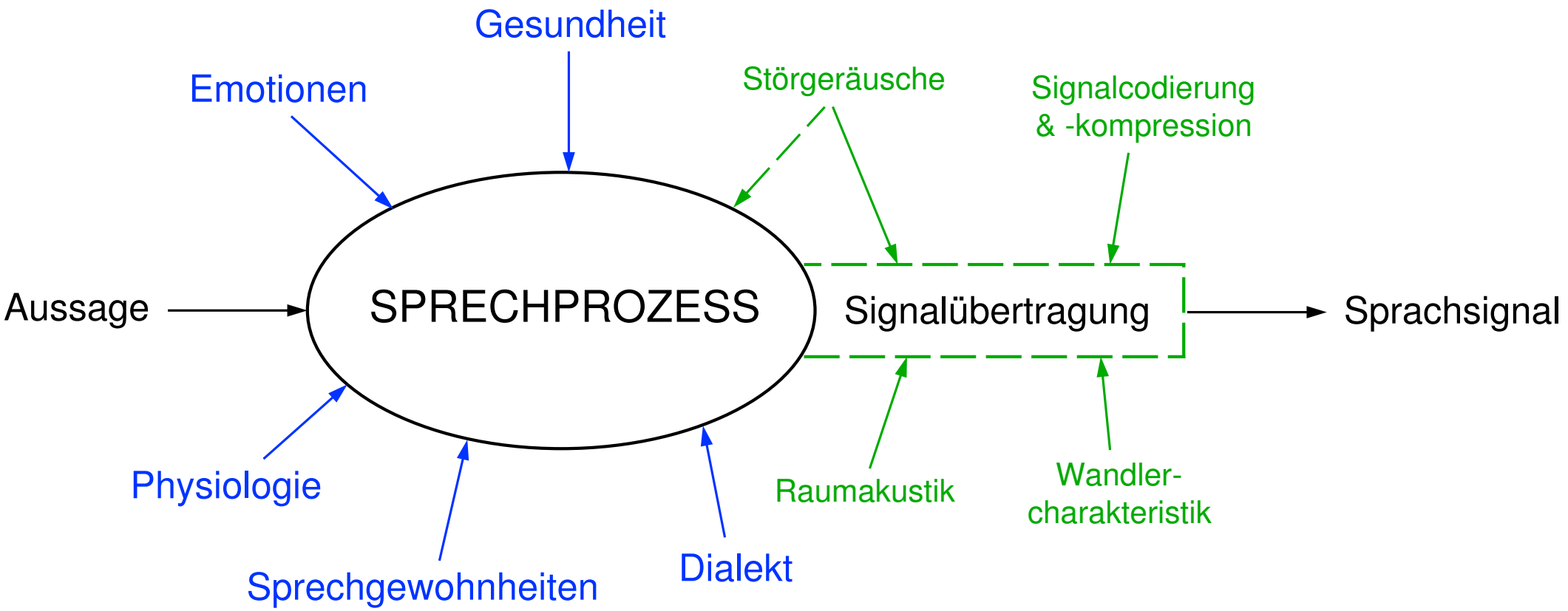
Vorlesung: Sprache $\longleftrightarrow$ Computer (Verarbeitung)

- Was steckt in einem Sprachsignal?
- Übersicht über die Sprachverarbeitung
- Digitalisierung von Sprachsignalen
- Darstellung und Eigenschaften von Sprachsignalen

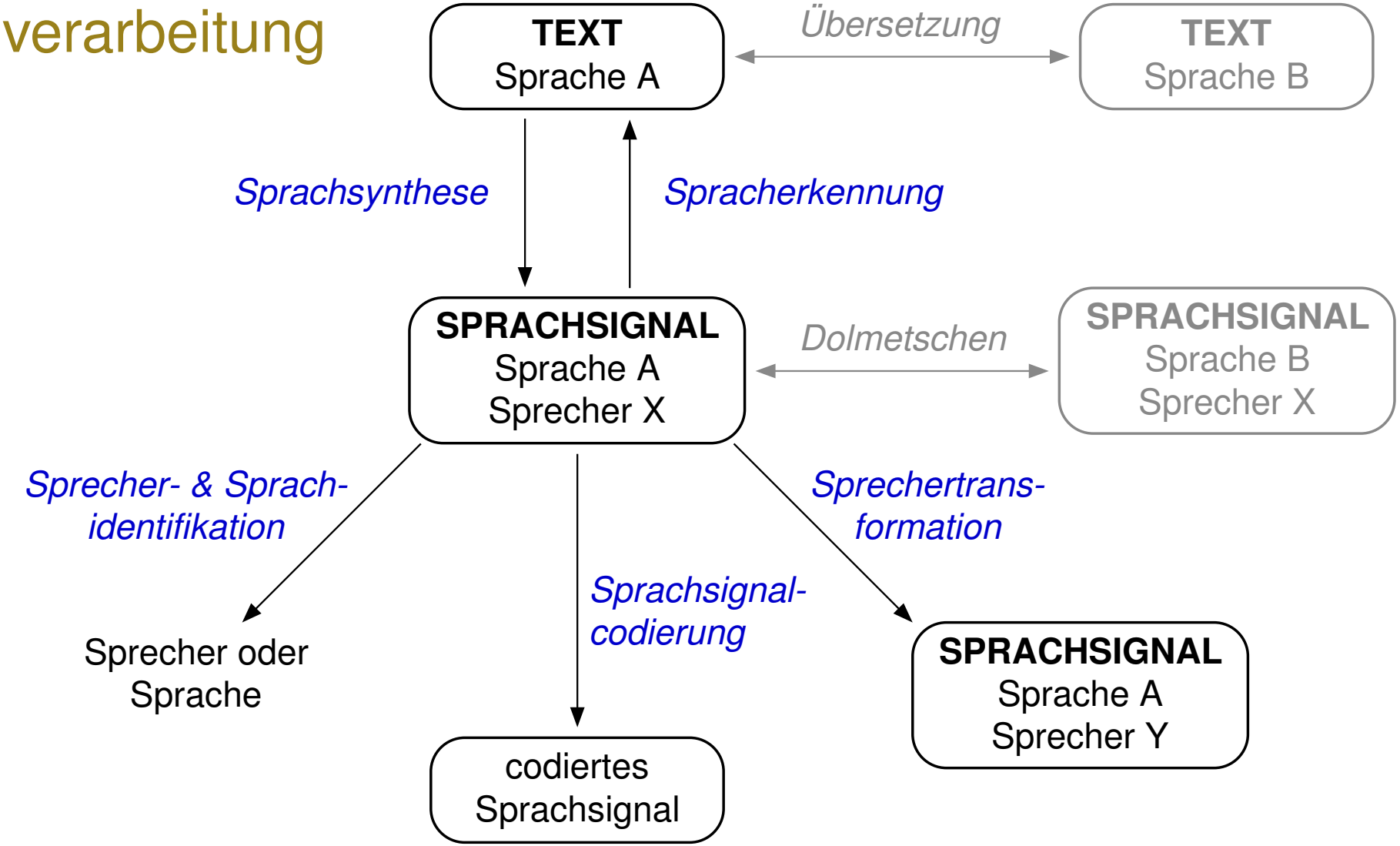
Übung: Quantisierung von Sprachsignalen

Was ist ein Sprachsignal?

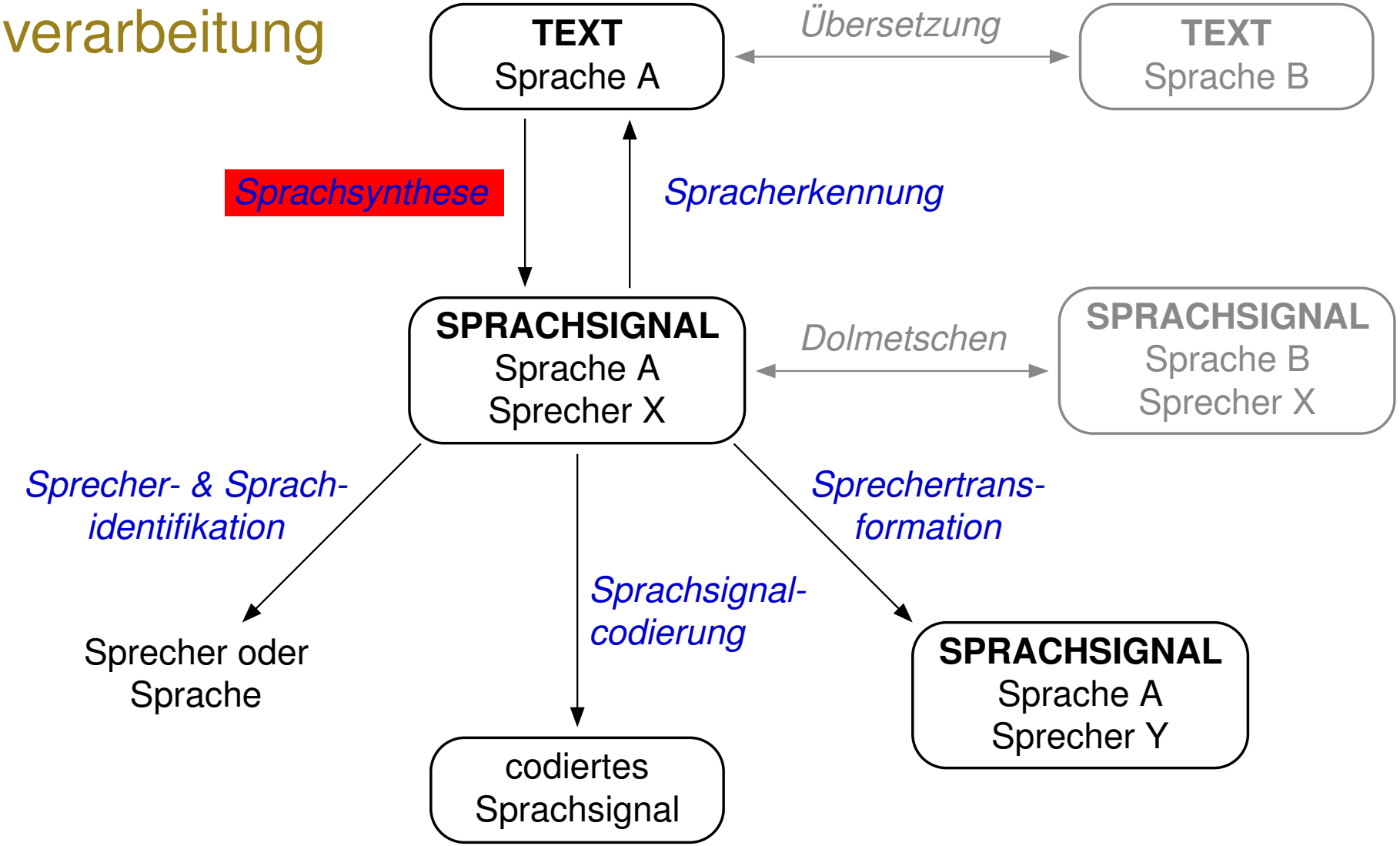

Aussage
SPRECHPROZESS
Sprachsignal


Gesundheit
Emotionen
Störgeräusche
Signalcodierung
& -kompression
Aussage
SPRECHPROZESS
Signalübertragung
Sprachsignal
Physiologie
Raumakustik
Wandler-
charakteristik
Sprechgewohnheiten
Dialekt

Teilgebiete der Sprachverarbeitung



Teilgebiete der Sprachverarbeitung



Sprachsynthese

- Sprachsynthese ab Text: TTS-Synthese (engl.: text-to-speech synthesis)

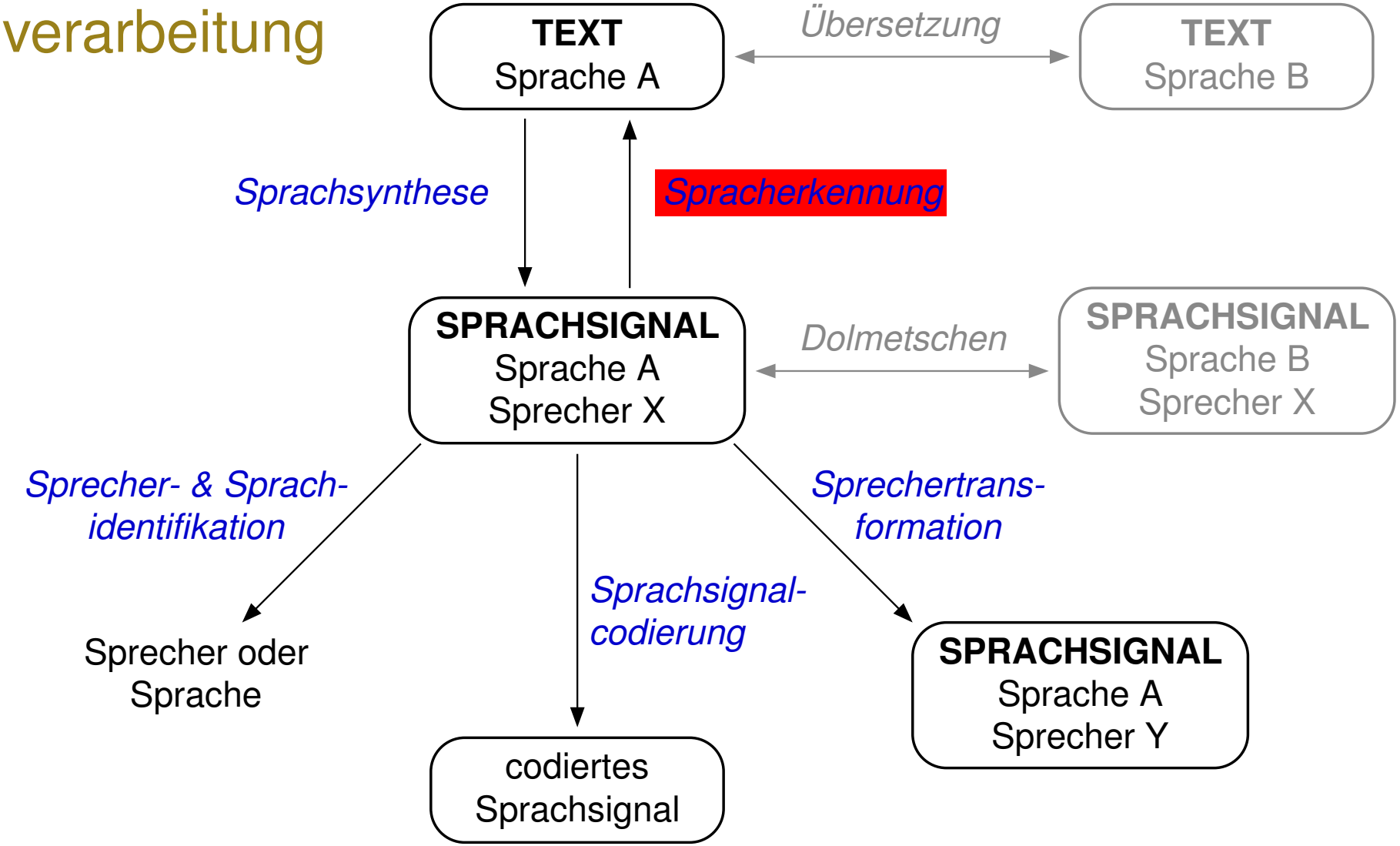
  Eingabe von Text in normaler orthographischer Schrift

- Sprachsynthese aus “Konzepten”: CTS-Synthese (engl.: concept-to-speech)

  hierarchisch strukturierte Eingabe

- Sprachausgabe: Aus- oder Wiedergabe von Sprachsignalen,
  die von einer Person aufgezeichnet worden sind

Teilgebiete der Sprachverarbeitung



Spracherkennung

Den allgemein einsetzbaren Spracherkenner gibt es (noch) nicht!

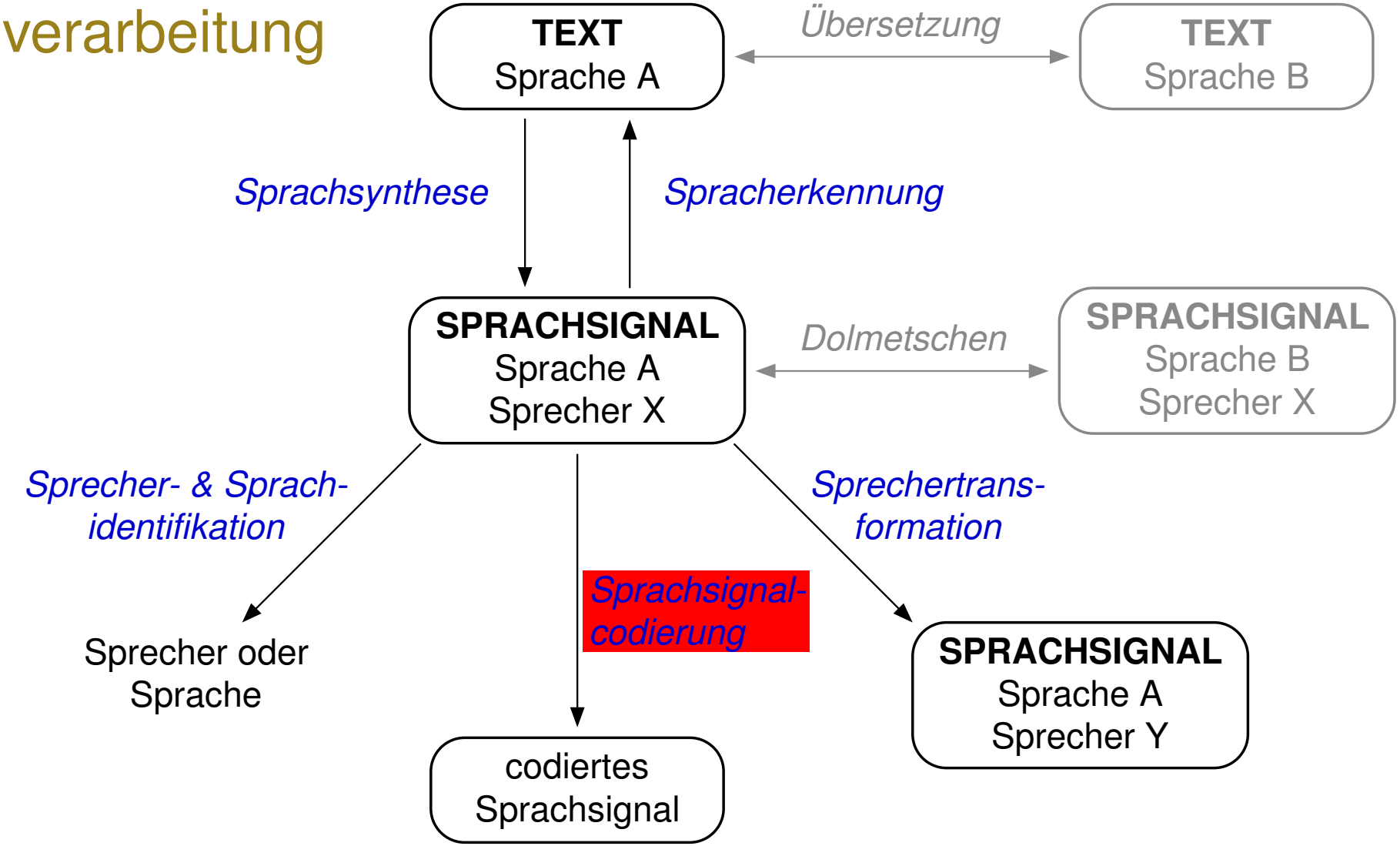
Spezielle Aufgaben:

- Erkennung einzeln gesprochener Wörter
- Spracherkenner für kleines Vokabular
- sprecherabhängige Spracherkennung

Wichtigste Ansätze:

- ⋆ Mustervergleich
- ⋆ statistischer Ansatz

Teilgebiete der Sprachverarbeitung



Codierung (digitaler Daten)

Dem Zweck einer Codierung entsprechend werden unterschieden:

Chiffrierung: Schutz der Daten vor unberechtigtem Zugriff >>>

Kanalcodierung: Anpassung der Daten an Übertragungskanal oder Speichermedium

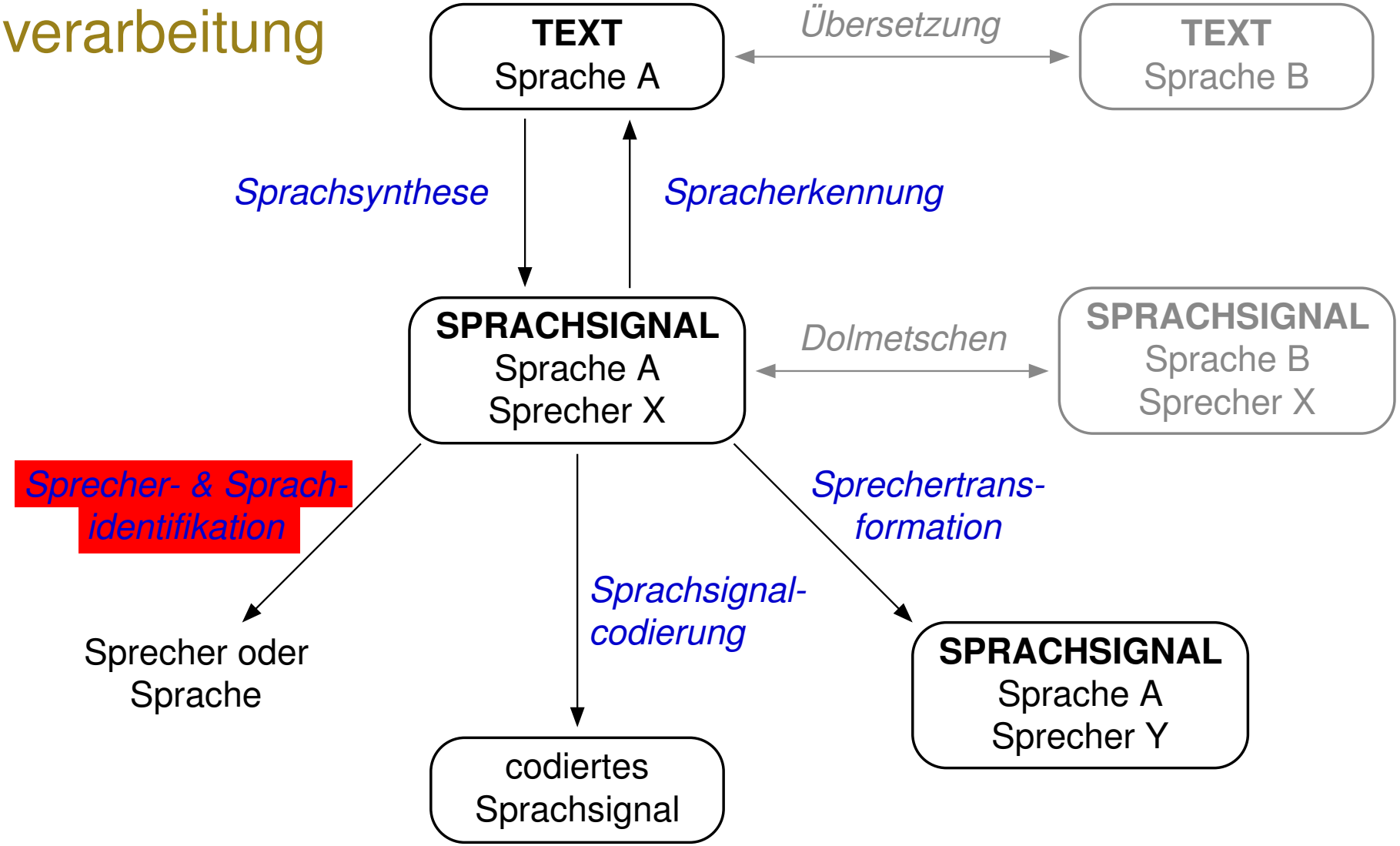
(höhere Zuverlässigkeit z.B. fehlerkorrigierende Codierung)

Quellencodierung: Kompression der Daten zwecks effizienter Übertragung oder Speicherung

- verlustlos z.B. Huffman-Codierung

  (für Sprachsignale nicht interessant)
- verlustbehaftet >>>

Teilgebiete der Sprachverarbeitung



Sprecheridentifikation “Stimme als Ausweis”

Prinzip: Vergleich eines Testsignals mit Referenzdaten

Textabhängige Verfahren:	Vergleich von Sprachmustern (Zeitnormalisation & Distanzmessung)
Textunabhängige Verfahren:	statistische Beschreibung / Modell des Sprachsignals jedes Sprechers

$\longrightarrow$ gute Resultate erzielbar!

Sprachverarbeitung

Verarbeitung von Sprachsignalen $\longrightarrow$ rein digital

Erster Schritt (vor Verarbeitung): Digitalisierung (A/D-Wandlung)

Verarbeitung

Letzter Schritt (vor Anhören): Rekonstruktion (D/A-Wandlung)

Digitalisierung von Sprachsignalen

Ziel bei der Digitalisierung: Wandlung $A \rightarrow D \rightarrow A$ ohne Qualitätsverlust!

Voraussetzungen:

- Abtastfrequenz so, dass kein Aliasing entsteht >>>
  (welche Frequenzen sind wichtig?) >>>
- Amplituden-Quantisierung genügend fein
- Rekonstruktionsfilter korrekt >>>

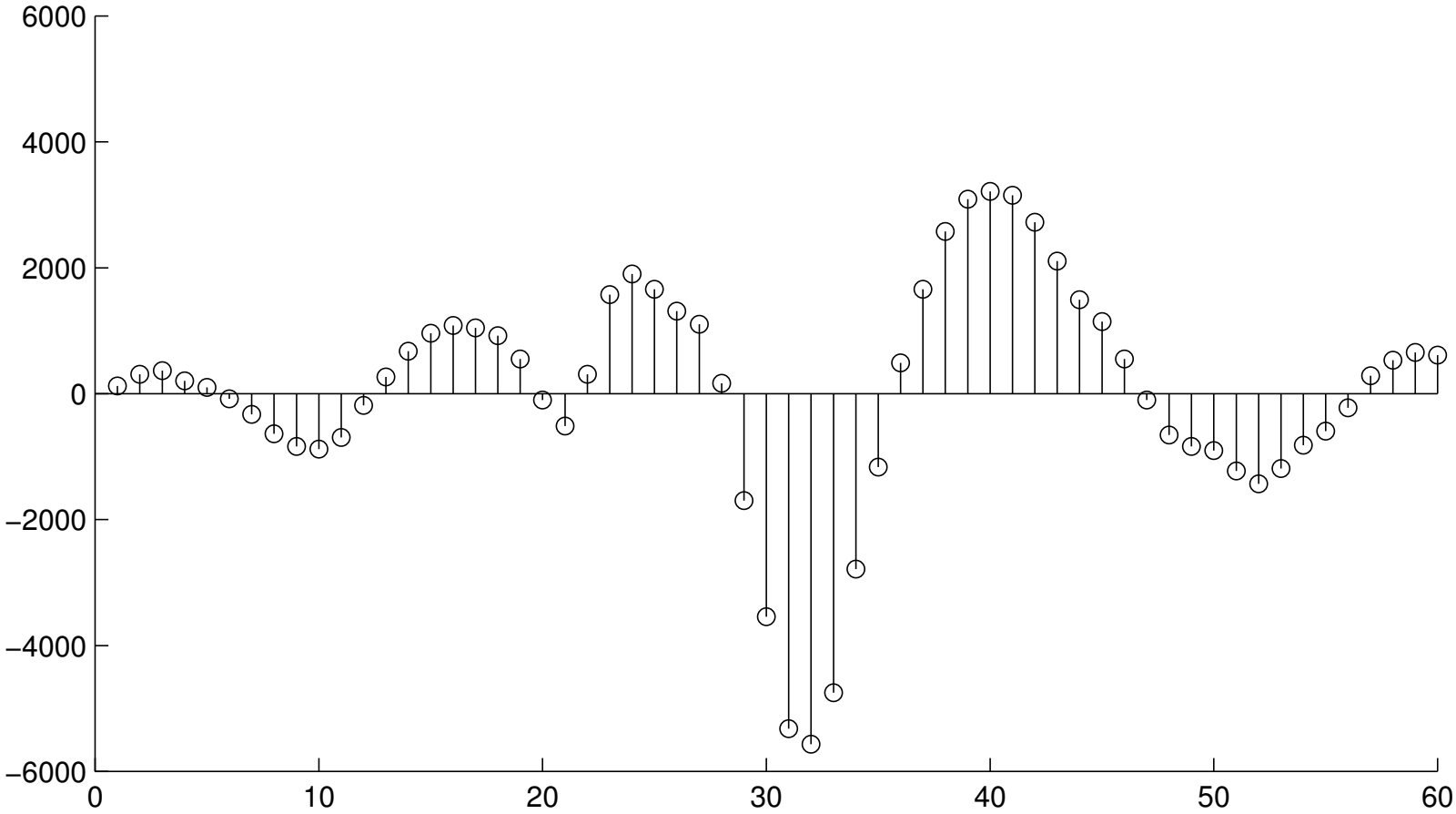
Darstellung des Sprachsignals

Diskrete Darstellung:	Abtastwerte des Schalldrucks
Oszillogramm:	Schalldruck (interpoliert) in Funktion der Zeit
Spektrum:	Betrag und Phase in Funktion der Frequenz
Spektrogramm:	Betrag in Funktion von Zeit und Frequenz
3D-Spektrogramm:	Betragspektren in Funktion der Zeit
Wirkliches Spektrogramm:	Signalkomponenten in Funktion der Zeit (Amplitude, Frequenz und Phase)

Darstellung des Sprachsignals

Diskrete Darstellung:	Abtastwerte des Schalldrucks
Oszillogramm:	Schalldruck (interpoliert) in Funktion der Zeit
Spektrum:	Betrag und Phase in Funktion der Frequenz
Spektrogramm:	Betrag in Funktion von Zeit und Frequenz
3D-Spektrogramm:	Betragspektren in Funktion der Zeit
Wirkliches Spektrogramm:	Signalkomponenten in Funktion der Zeit (Amplitude, Frequenz und Phase)

Darstellung der Abtastwerte in Funktion des Indexes

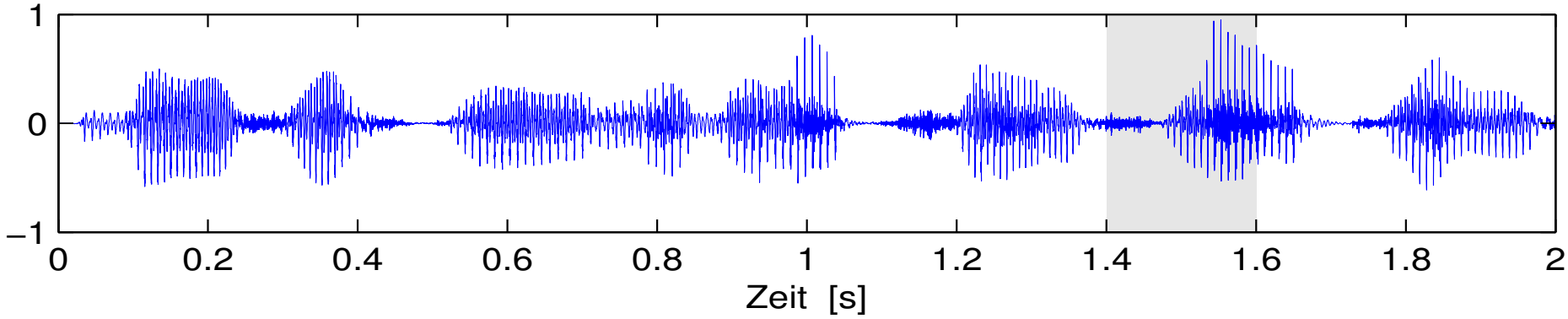


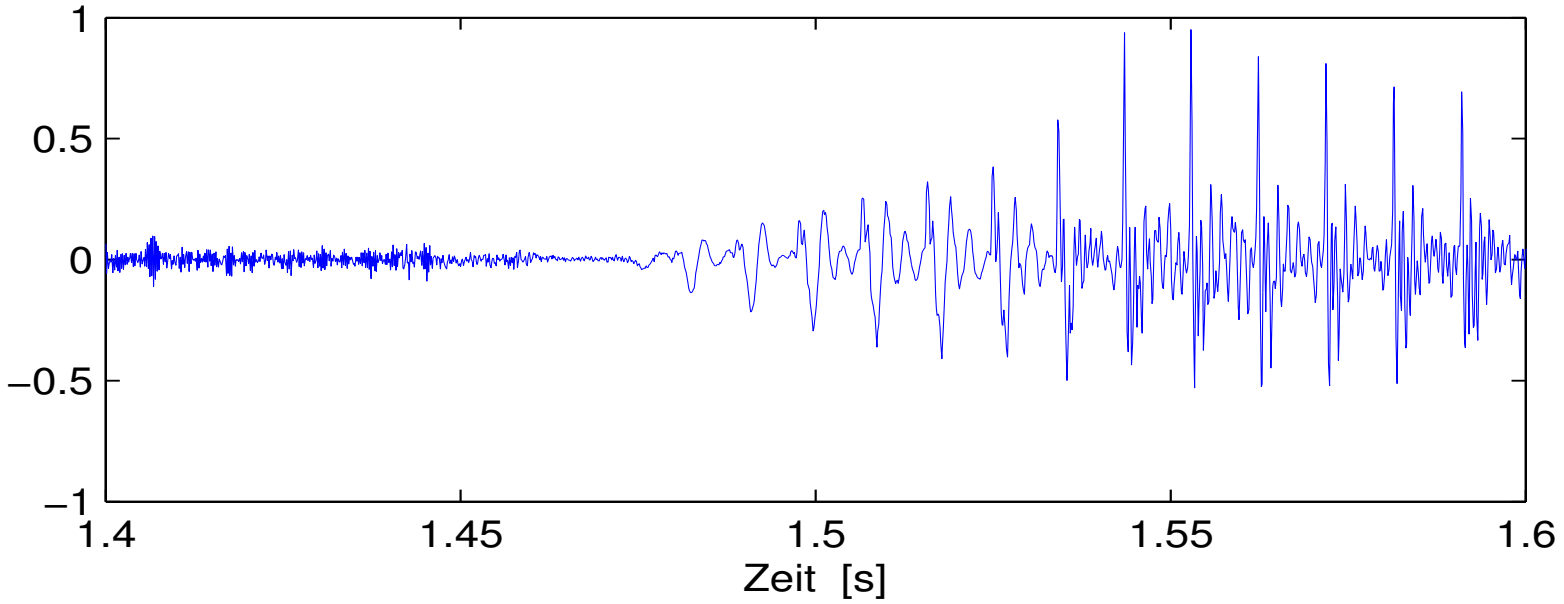
(erzeugt mit Matlab-Funktion: stem)

Darstellung des Sprachsignals

Diskrete Darstellung:	Abtastwerte des Schalldrucks
Oszillogramm:	Schalldruck (interpoliert) in Funktion der Zeit
Spektrum:	Betrag und Phase in Funktion der Frequenz
Spektrogramm:	Betrag in Funktion von Zeit und Frequenz
3D-Spektrogramm:	Betragspektren in Funktion der Zeit
Wirkliches Spektrogramm:	Signalkomponenten in Funktion der Zeit (Amplitude, Frequenz und Phase)

Zeitsignal (Oszillogramm)

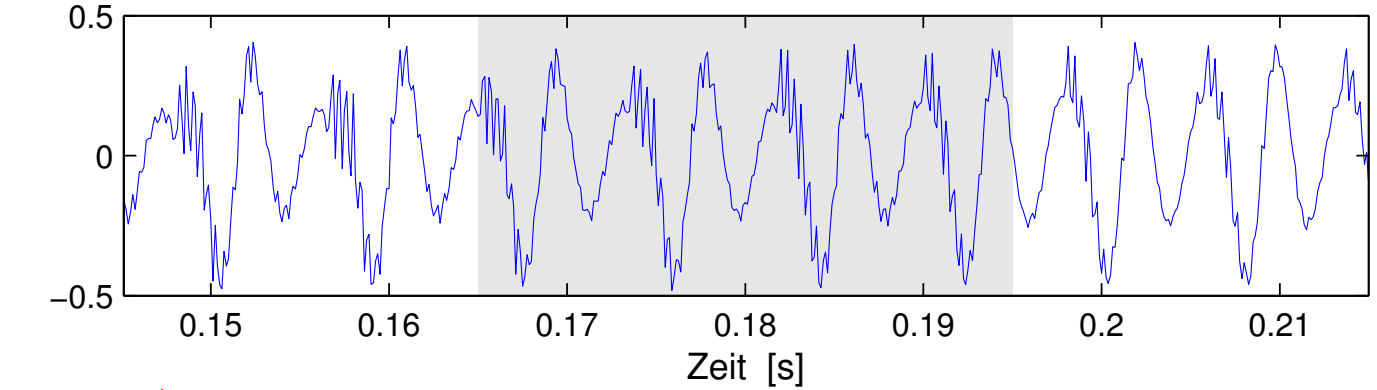


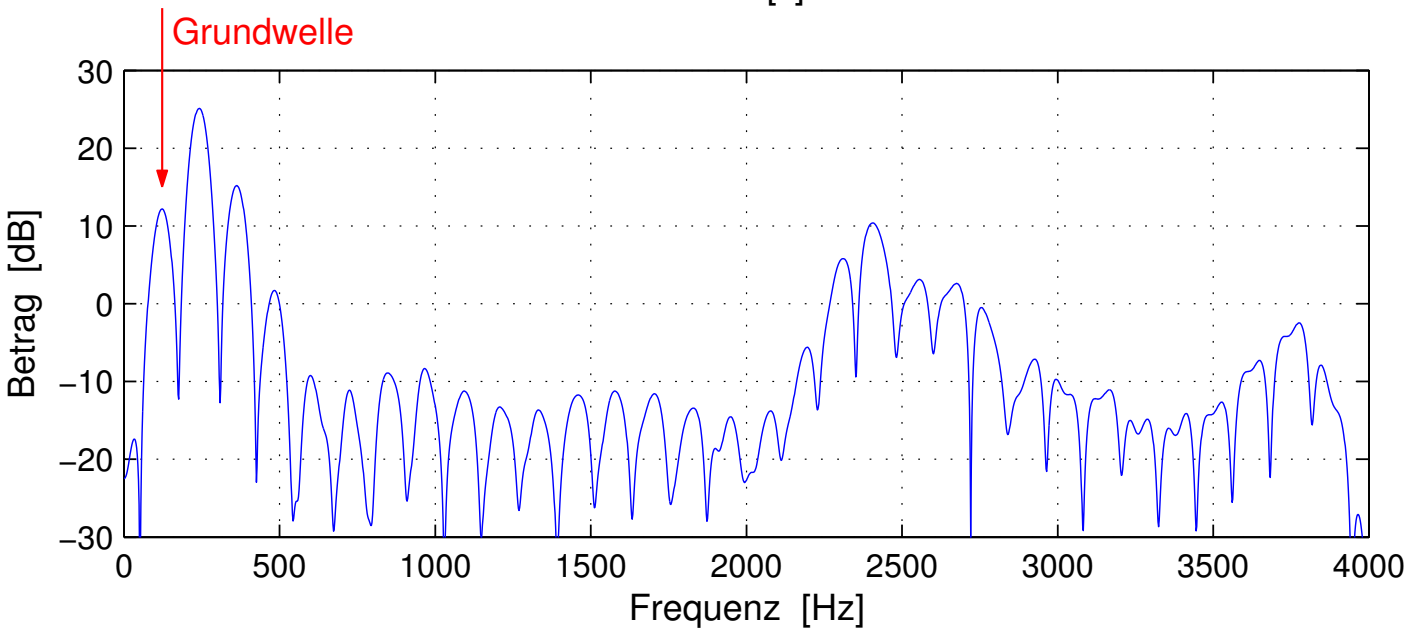


Darstellung des Sprachsignals

Diskrete Darstellung:	Abtastwerte des Schalldrucks
Oszillogramm:	Schalldruck (interpoliert) in Funktion der Zeit
Spektrum:	Betrag und Phase in Funktion der Frequenz
Spektrogramm:	Betrag in Funktion von Zeit und Frequenz
3D-Spektrogramm:	Betragspektren in Funktion der Zeit
Wirkliches Spektrogramm:	Signalkomponenten in Funktion der Zeit (Amplitude, Frequenz und Phase)

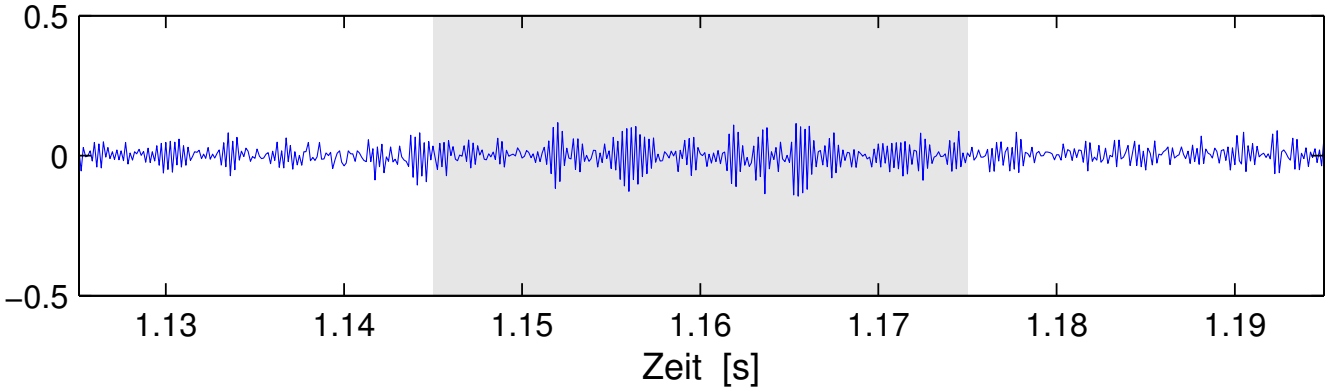
Spektrum des Sprachsignals (stimmhafter Ausschnitt)

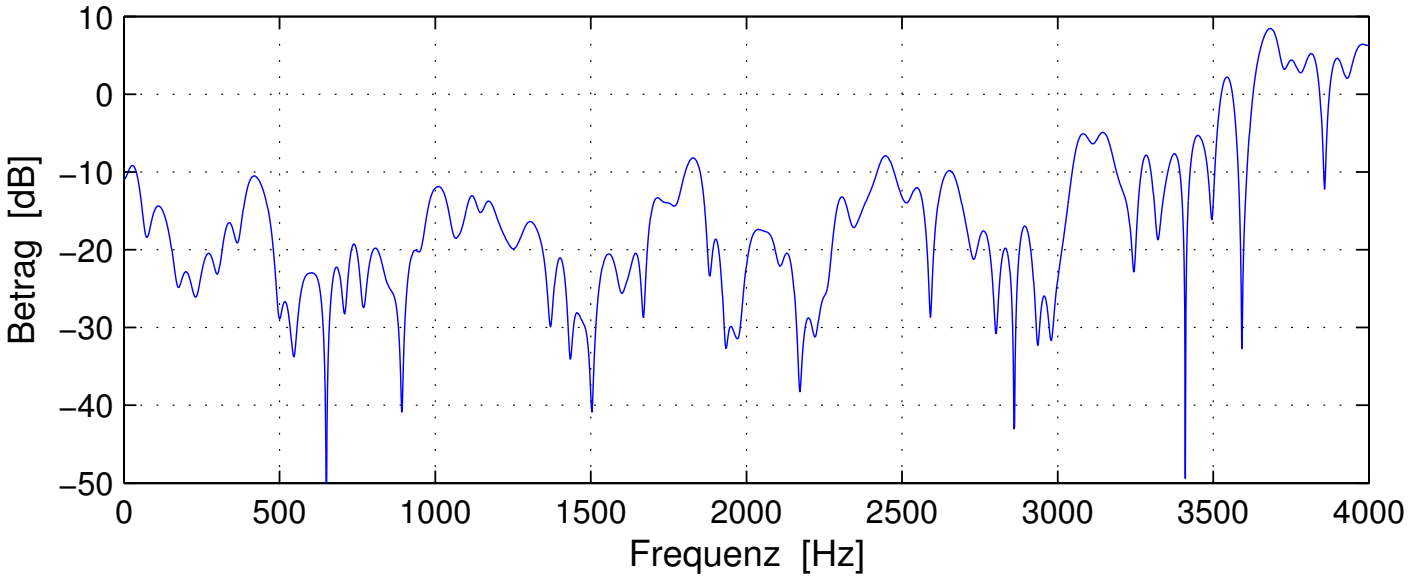




Phase?

Spektrum des Sprachsignals (stimmloser Ausschnitt)





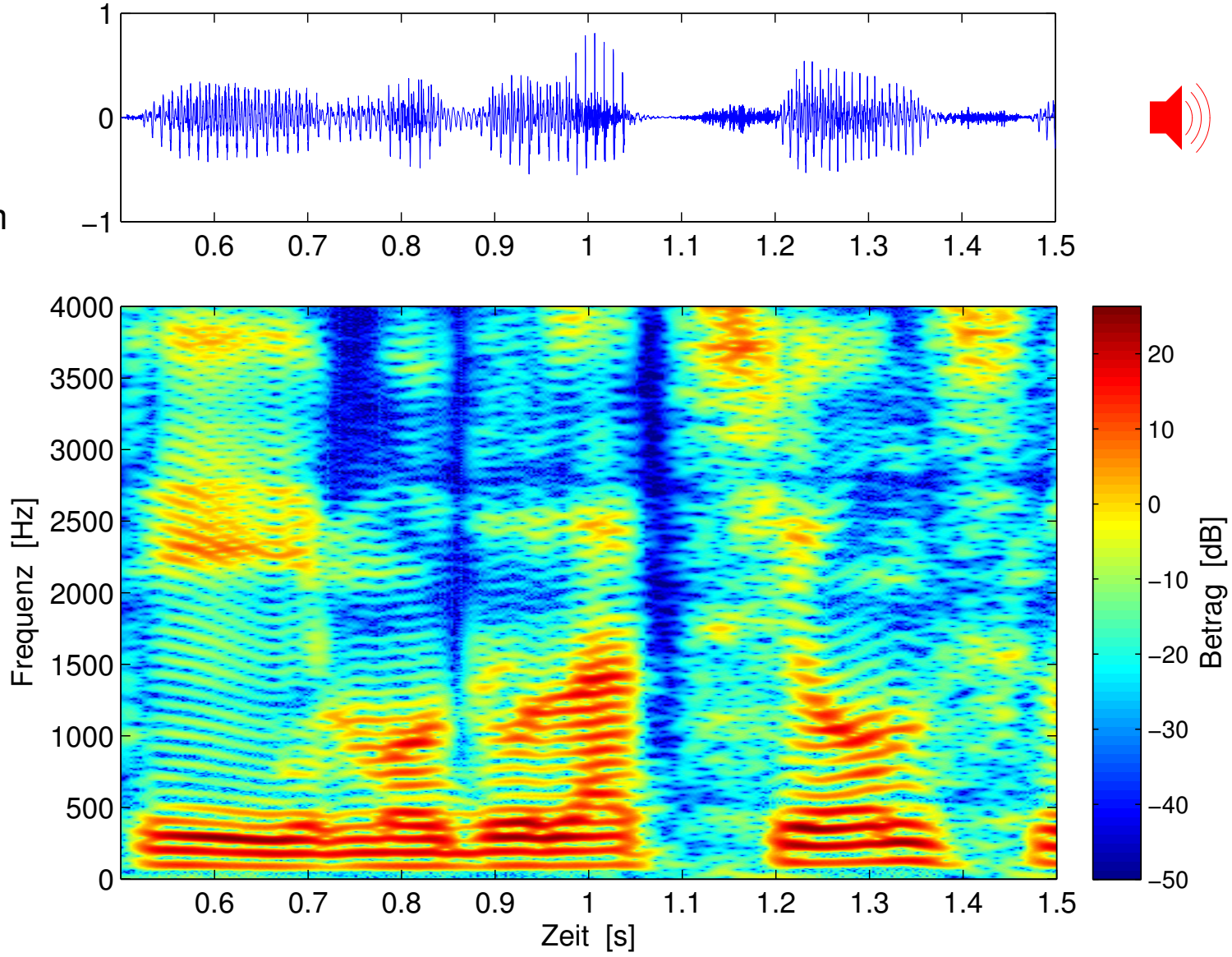
Phase?

Darstellung des Sprachsignals

Diskrete Darstellung:	Abtastwerte des Schalldrucks
Oszillogramm:	Schalldruck (interpoliert) in Funktion der Zeit
Spektrum:	Betrag und Phase in Funktion der Frequenz
Spektrogramm:	Betrag in Funktion von Zeit und Frequenz
3D-Spektrogramm:	Betragspektren in Funktion der Zeit
Wirkliches Spektrogramm:	Signalkomponenten in Funktion der Zeit (Amplitude, Frequenz und Phase)

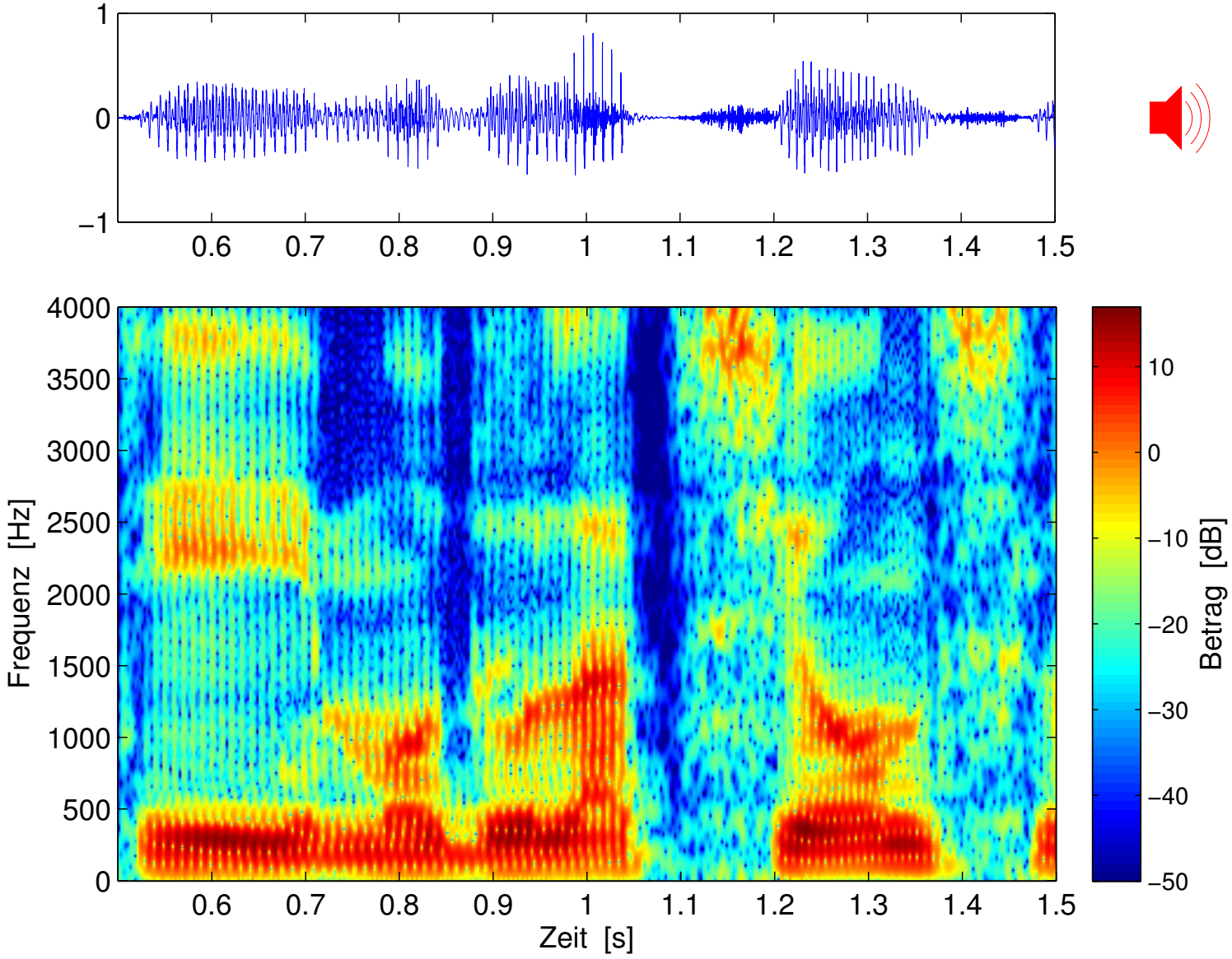
Spektrogramm

Schmalbandspektrogramm



Spektrogramm

Breitbandspektrogramm

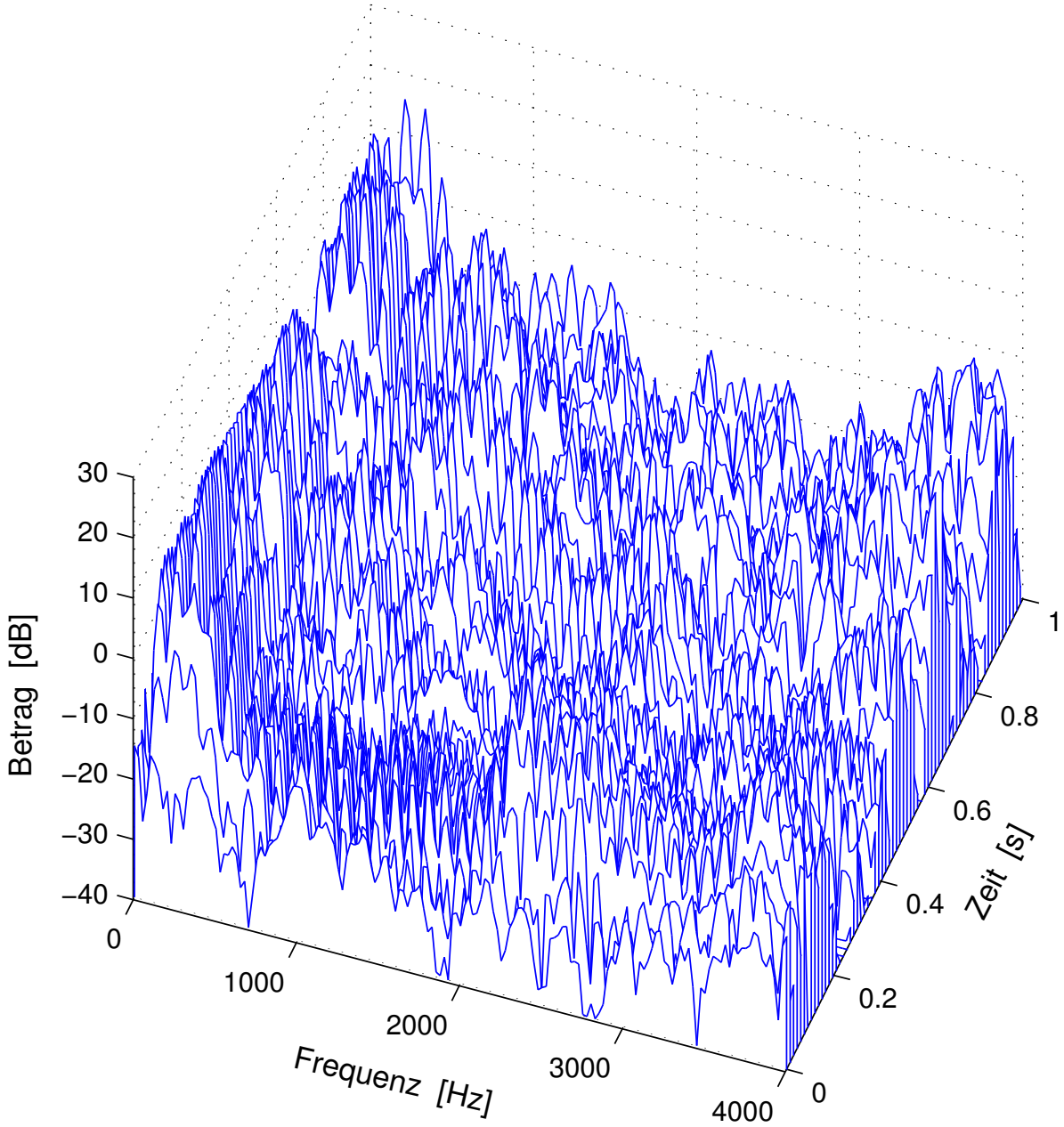


Darstellung des Sprachsignals

Diskrete Darstellung:	Abtastwerte des Schalldrucks
Oszillogramm:	Schalldruck (interpoliert) in Funktion der Zeit
Spektrum:	Betrag und Phase in Funktion der Frequenz
Spektrogramm:	Betrag in Funktion von Zeit und Frequenz
3D-Spektrogramm:	Betragspektren in Funktion der Zeit
Wirkliches Spektrogramm:	Signalkomponenten in Funktion der Zeit (Amplitude, Frequenz und Phase)

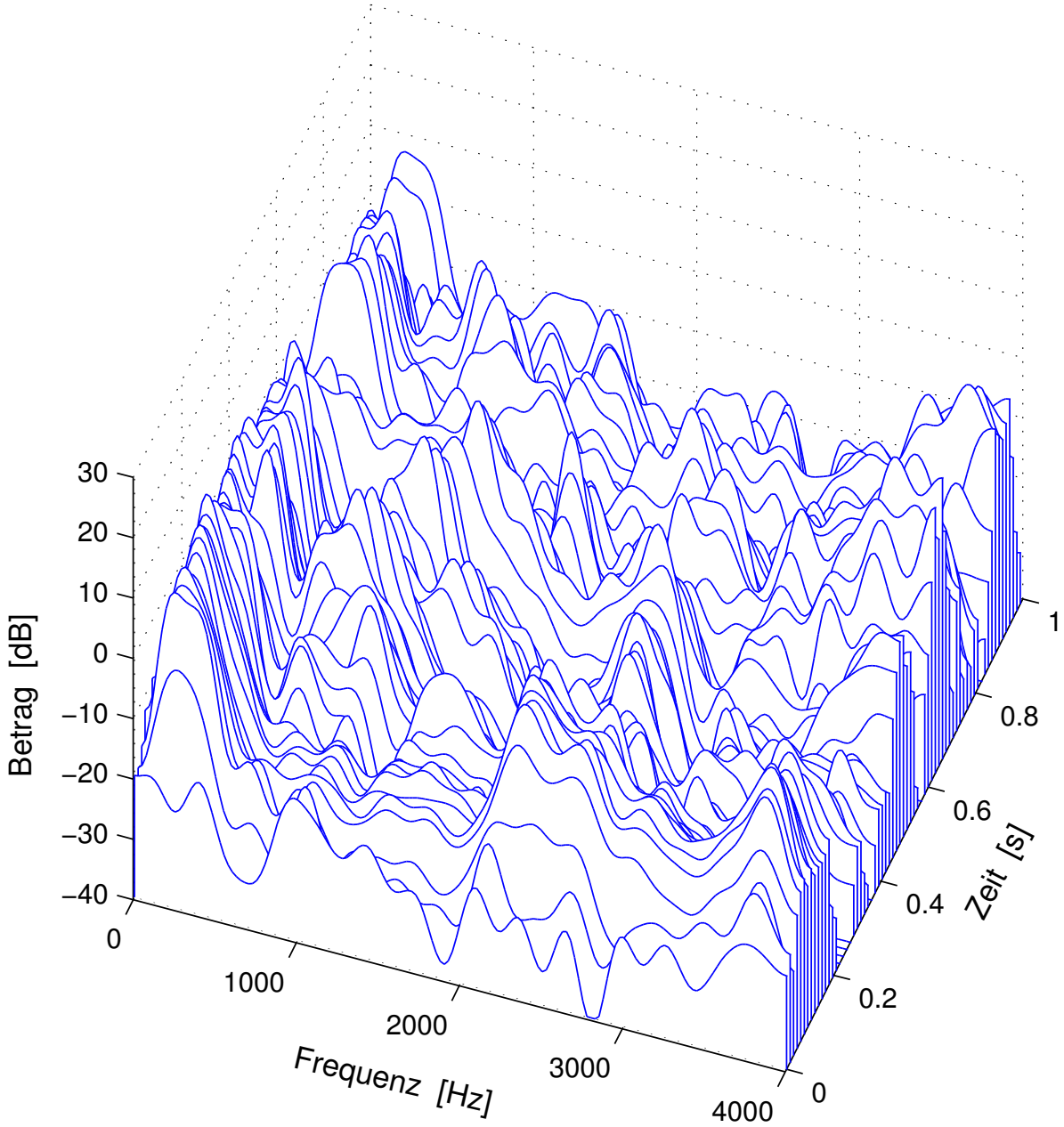
Kurzzeitspektrum

3-dimensionale Darstellung



Geglättetes Kurzzeitspektrum

3-dimensionale Darstellung



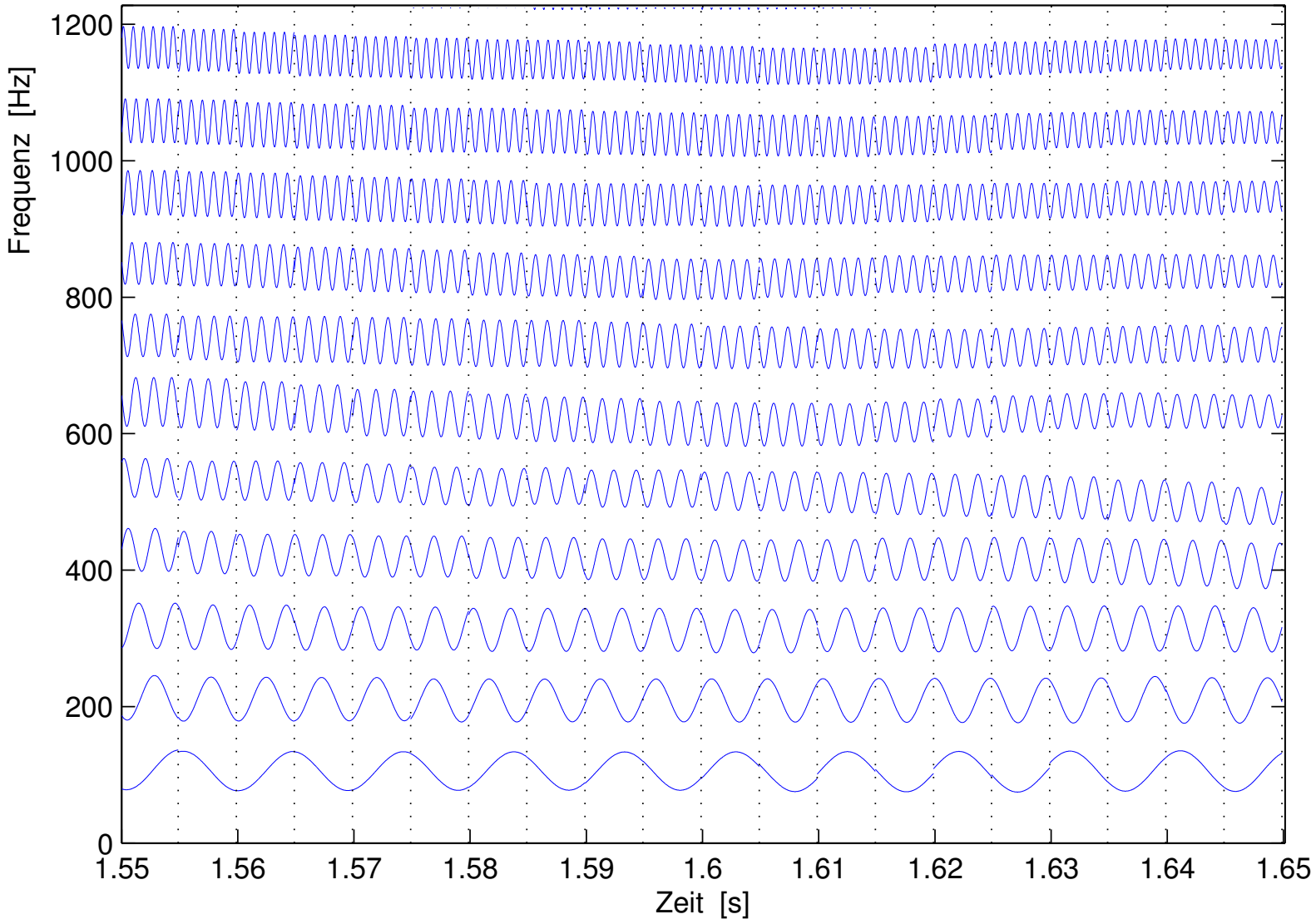
Darstellung des Sprachsignals

Diskrete Darstellung:	Abtastwerte des Schalldrucks
Oszillogramm:	Schalldruck (interpoliert) in Funktion der Zeit
Spektrum:	Betrag und Phase in Funktion der Frequenz
Spektrogramm:	Betrag in Funktion von Zeit und Frequenz
3D-Spektrogramm:	Betragspektren in Funktion der Zeit
Wirkliches Spektrogramm:	Signalkomponenten in Funktion der Zeit (Amplitude, Frequenz und Phase)

Wirkliches Spektrogramm

Spektrale Komponenten

in Funktion der Zeit



Eigenschaften / Merkmale des Sprachsignals

Aus Verarbeitung des Sprachsignals

Formanten:	Resonanzen des Vokaltraktes (zeitabhängig)
Grundfrequenz:	Frequenz der Grundwelle (zeitabhängig)
Lautdauer:	Lautgrenzen im Sprachsignal
Lautintensität:	lokale Signalleistung / wahrgenommene Lautheit

Eigenschaften / Merkmale des Sprachsignals

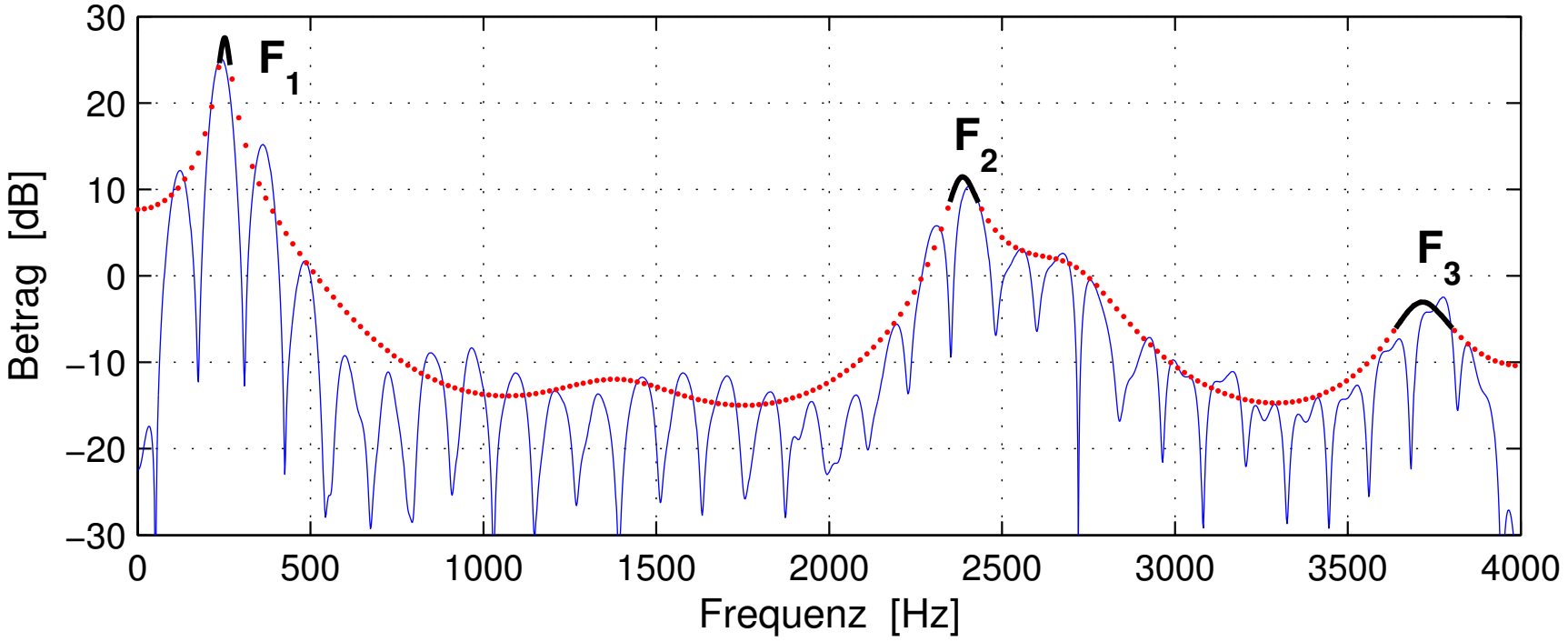
Aus Verarbeitung des Sprachsignals

Formanten:	Resonanzen des Vokaltraktes (zeitabhängig)
Grundfrequenz:	Frequenz der Grundwelle (zeitabhängig)
Lautdauer:	Lautgrenzen im Sprachsignal
Lautintensität:	lokale Signalleistung / wahrgenommene Lautheit

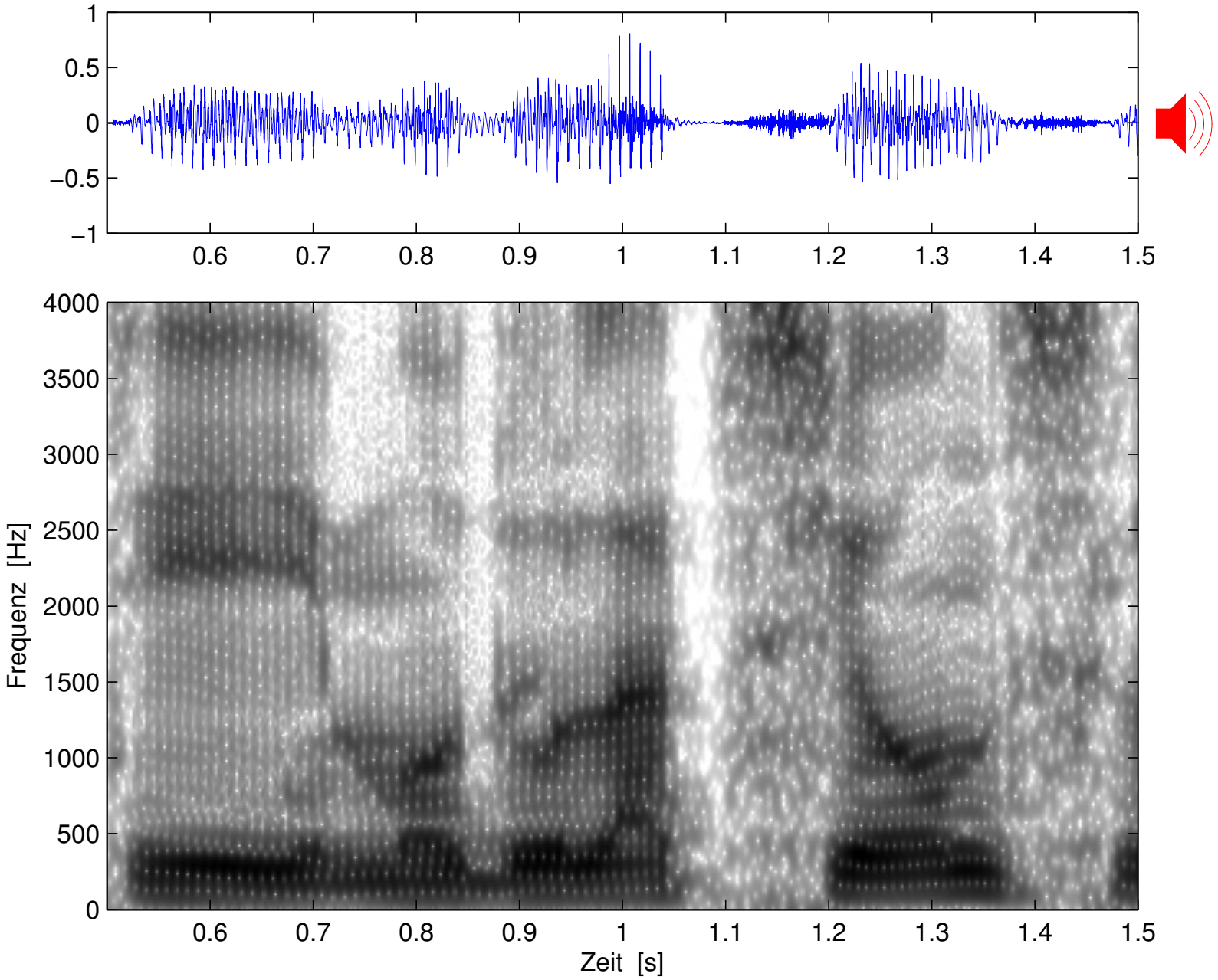
Formanten

Resonanzen des Vokaltraktes $\longrightarrow$ Frequenzbereiche mit mehr Energie

(hauptsächlich bei Vokalen)



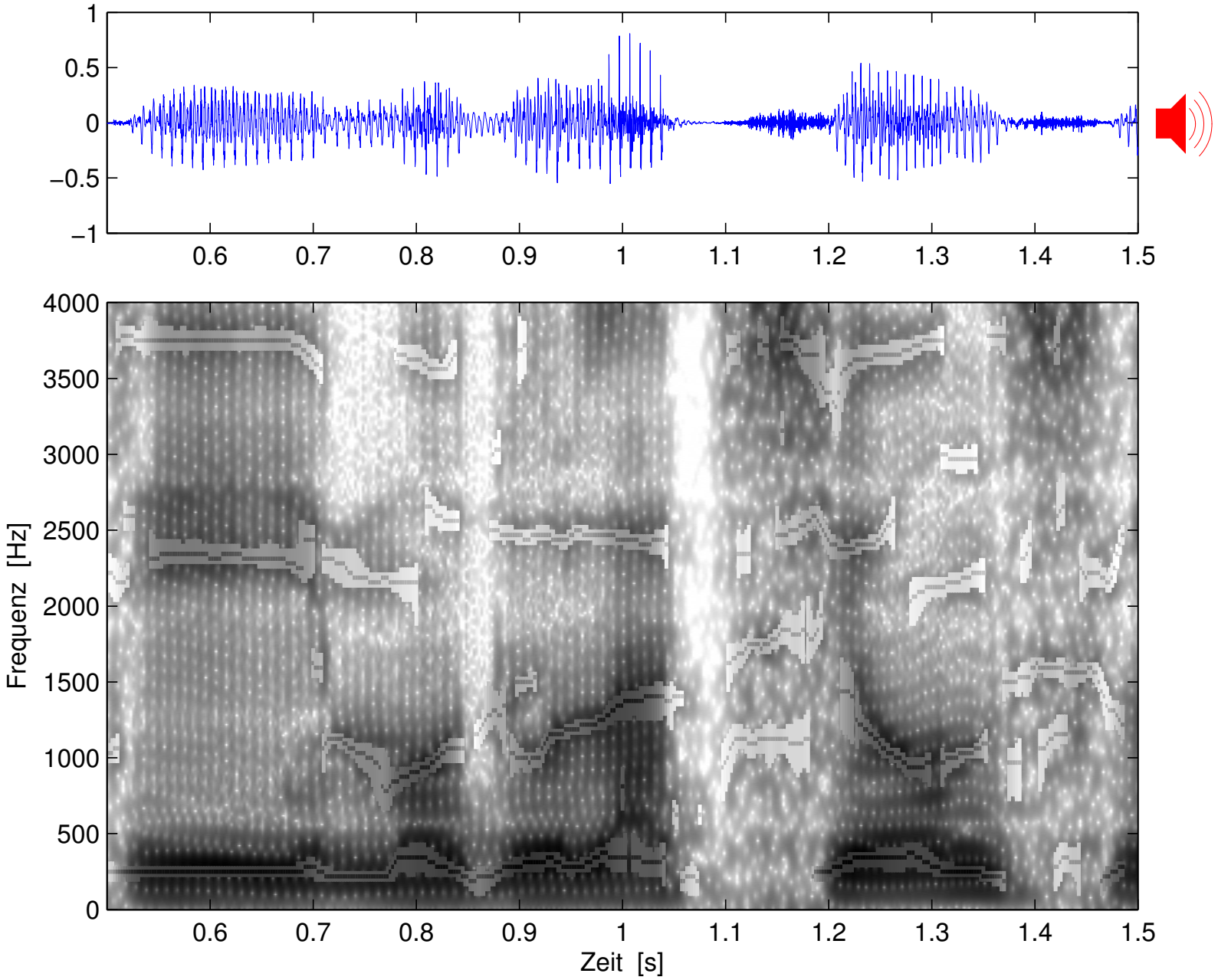
Breitband-spektrogramm



Formanten

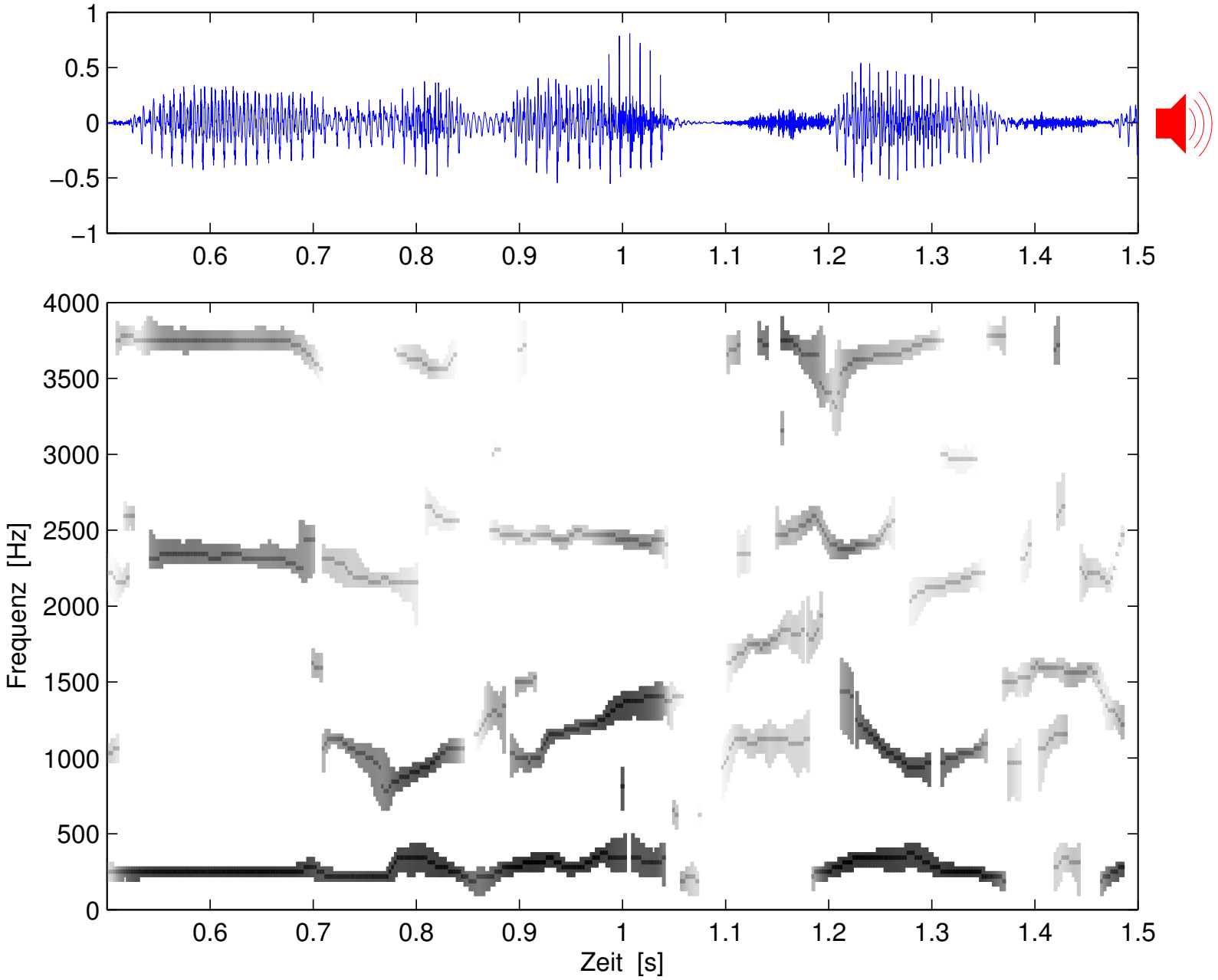
auf Breitband-

spektrogramm



Formanten

- Mittenfrequenz
- 3dB-Bandbreite
- Amplitude

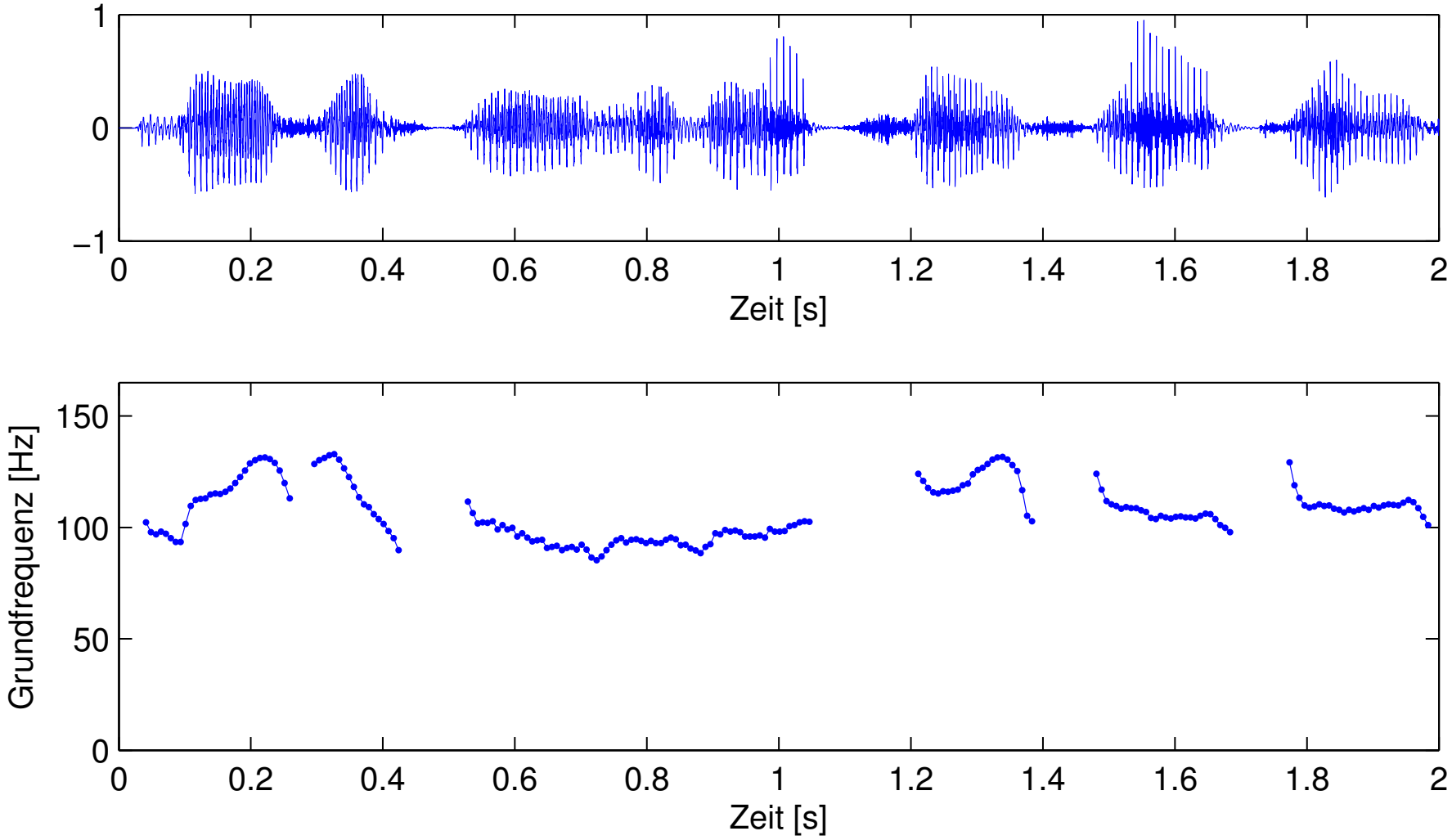


Eigenschaften / Merkmale des Sprachsignals

Aus Verarbeitung des Sprachsignals

Formanten:	Resonanzen des Vokaltraktes (zeitabhängig)
Grundfrequenz:	Frequenz der Grundwelle (zeitabhängig)
Lautdauer:	Lautgrenzen im Sprachsignal
Lautintensität:	lokale Signalleistung / wahrgenommene Lautheit

Grundfrequenz zeitlicher Verlauf

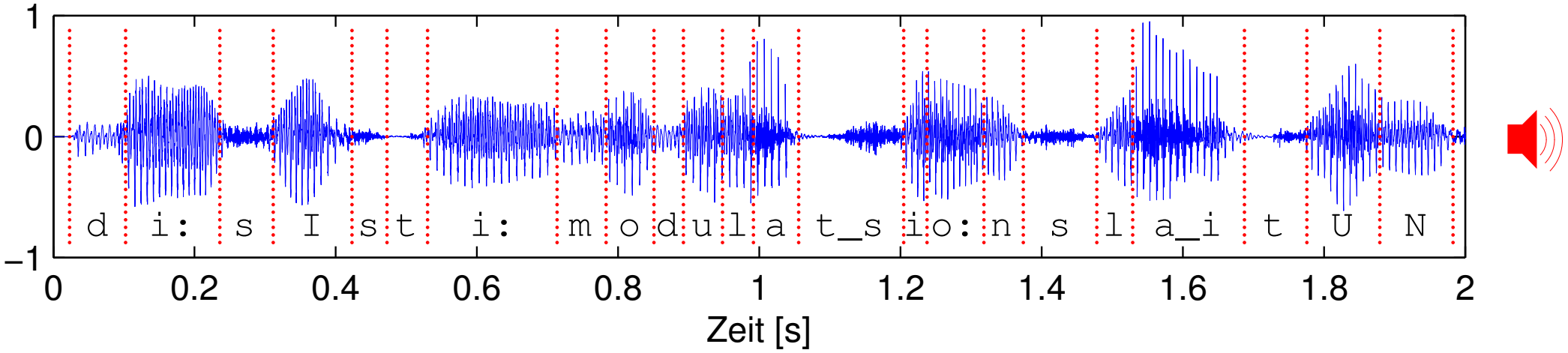


Eigenschaften / Merkmale des Sprachsignals

Aus Verarbeitung des Sprachsignals

Formanten:	Resonanzen des Vokaltraktes (zeitabhängig)
Grundfrequenz:	Frequenz der Grundwelle (zeitabhängig)
Lautdauer:	Lautgrenzen im Sprachsignal
Lautintensität:	lokale Signalleistung / wahrgenommene Lautheit

Lautgrenzen markiert im Oszillogramm



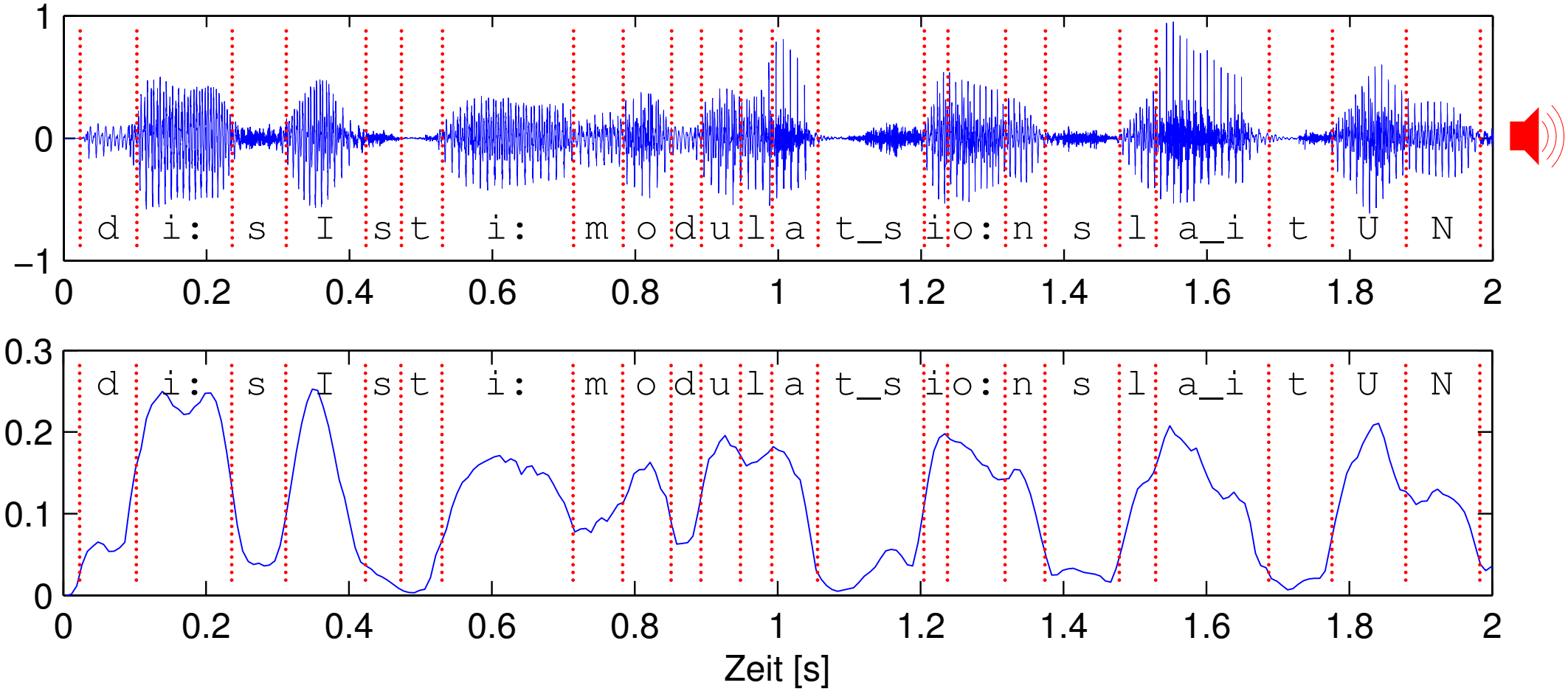
(manuell gesetzte Lautgrenzen)

Eigenschaften / Merkmale des Sprachsignals

Aus Verarbeitung des Sprachsignals

Formanten:	Resonanzen des Vokaltraktes (zeitabhängig)
Grundfrequenz:	Frequenz der Grundwelle (zeitabhängig)
Lautdauer:	Lautgrenzen im Sprachsignal
Lautintensität:	lokale Signalleistung / wahrgenommene Lautheit

Intensität des Sprachsignals



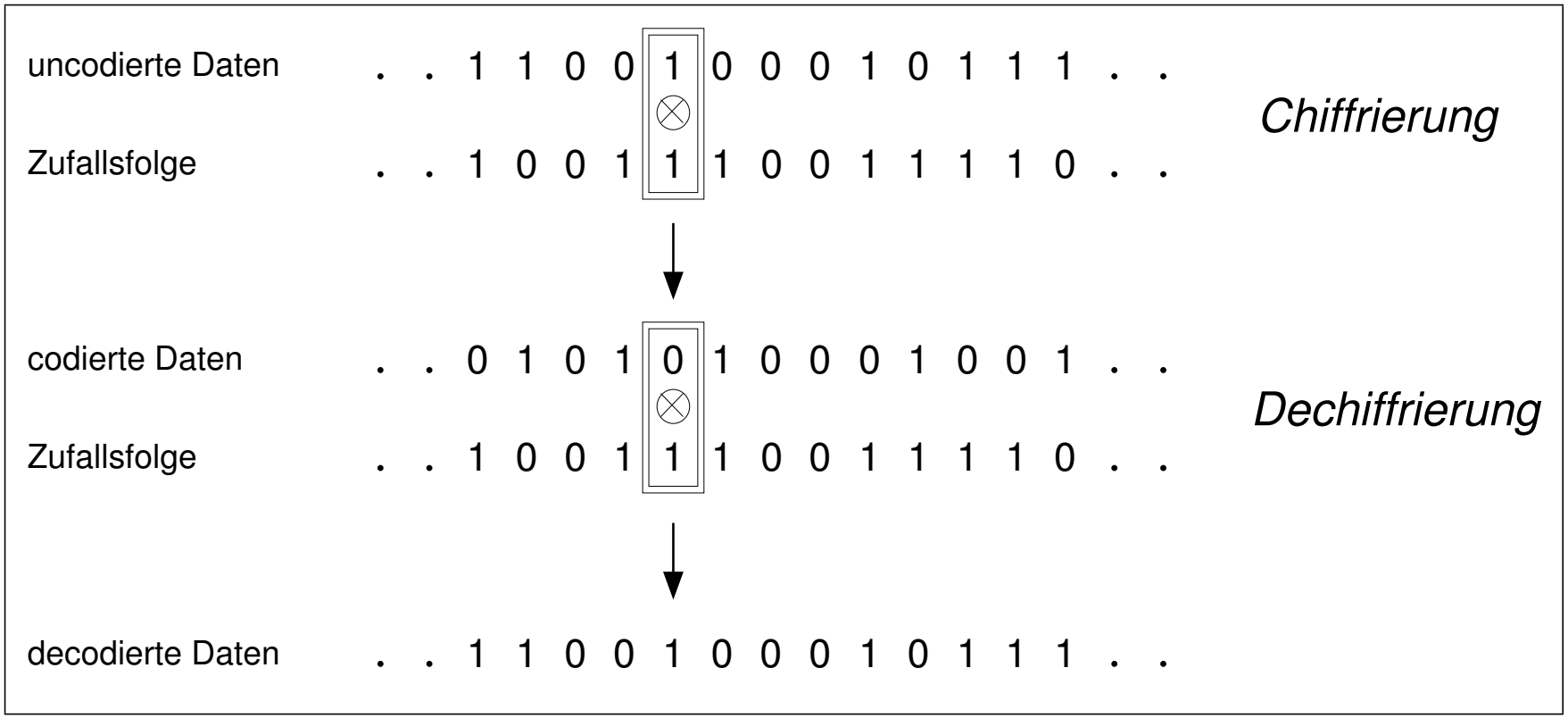
Merke: Signalleistung $\neq$ wahrgenommene Lautheit !

Thema der nächsten Lektion:

Analyse des Sprachsignals

Zur Übersicht der Vorlesung *Sprachverarbeitung I* >>>

Prinzip der Datenchiffrierung



<<<

Quellencodierung von Sprachsignalen

Signalformcodierung: Approximation des zeitlichen Verlaufs

- weniger Bits notwendig
- Codierungsfehler möglichst wenig hörbar

>>>

Sprachsignalmodellierung (Vocoder):

Einsatz von Wissen über die Erzeugung von Sprachsignalen

$\longrightarrow$ Trennung von Tonerzeugung und Klangformung
z.B. LPC (linear predictive coding)

- Tonquellen: Pulsfolge & Rauschen
- Klangformung: digitales Filter

Beispiele von Sprachcodierungen

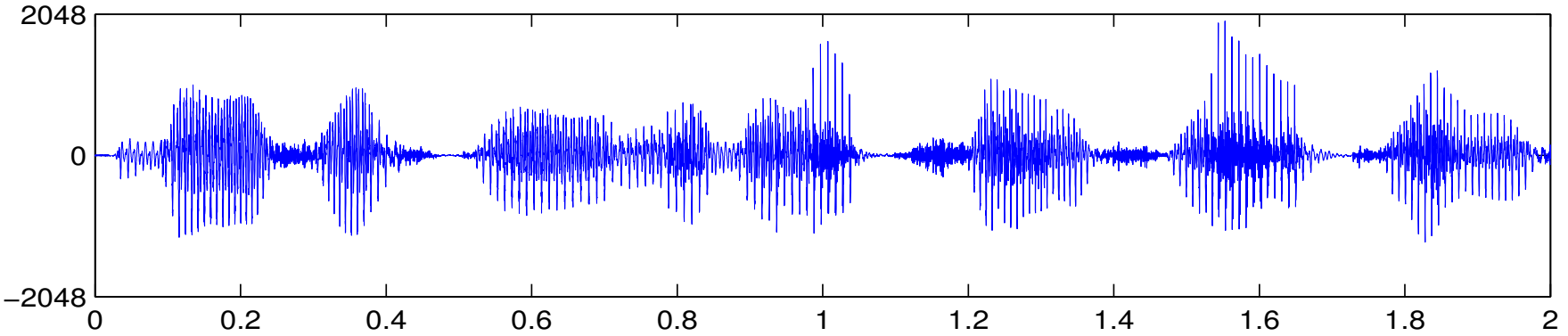
	Verfahren	Abtastfreq.	Bitrate	
Originalsignal	16-Bit-PCM	16 kHz	256 kBit/s	
Referenzsignal	12-Bit-PCM	8 kHz	96 kBit/s	
Signalquantisierung	8-Bit-log-PCM	8 kHz	64 kBit/s	
Signalformcodierung	ADPCM	8 kHz	32 kBit/s	
Sprachmodellierung	LPC	8 kHz	$\approx$ 6 kBit/s	
Sprachmodellierung	LPC	8 kHz	$\approx$ 3 kBit/s	

<<<

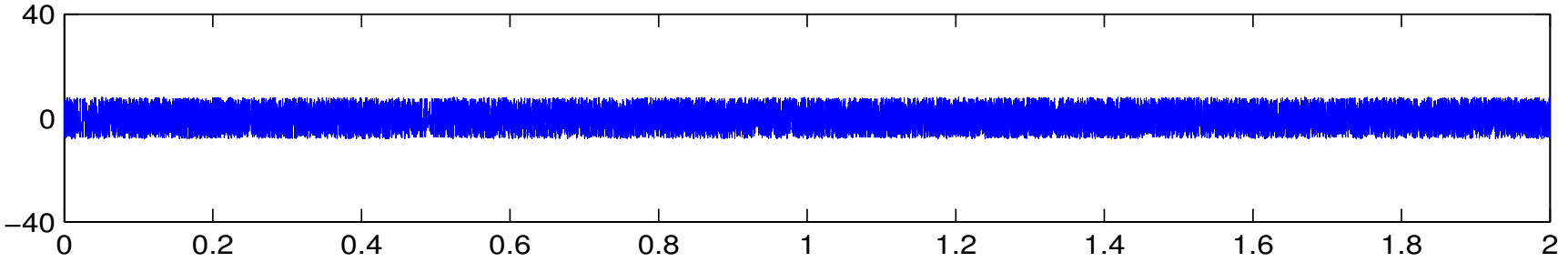
Logarithmische 8-Bit-Quantisierung

Datenreduktion: 33 %

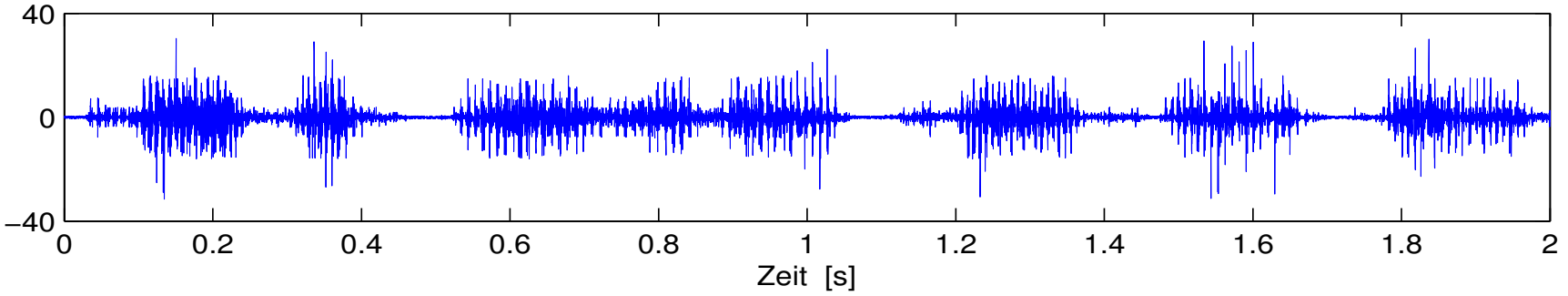
12 Bits



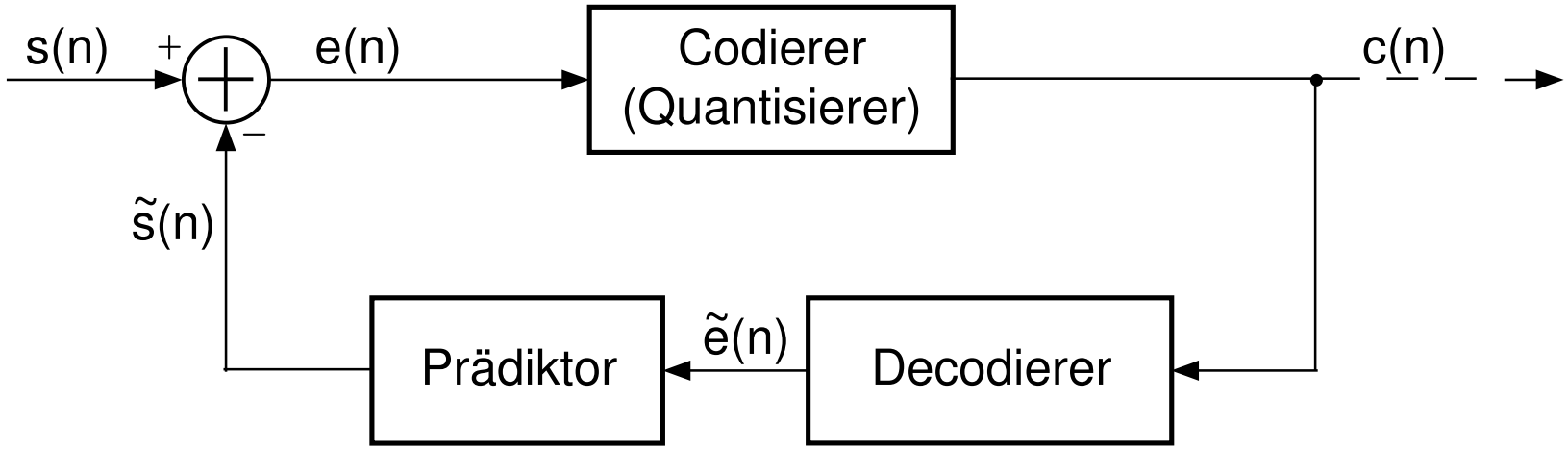
Fehlersignal
lin. 8-Bit-Q.



Fehlersignal
log. 8-Bit-Q.



Differenz-Codierer



Beispiel: ADPCM (adaptive differential pulse code modulation)

<<<

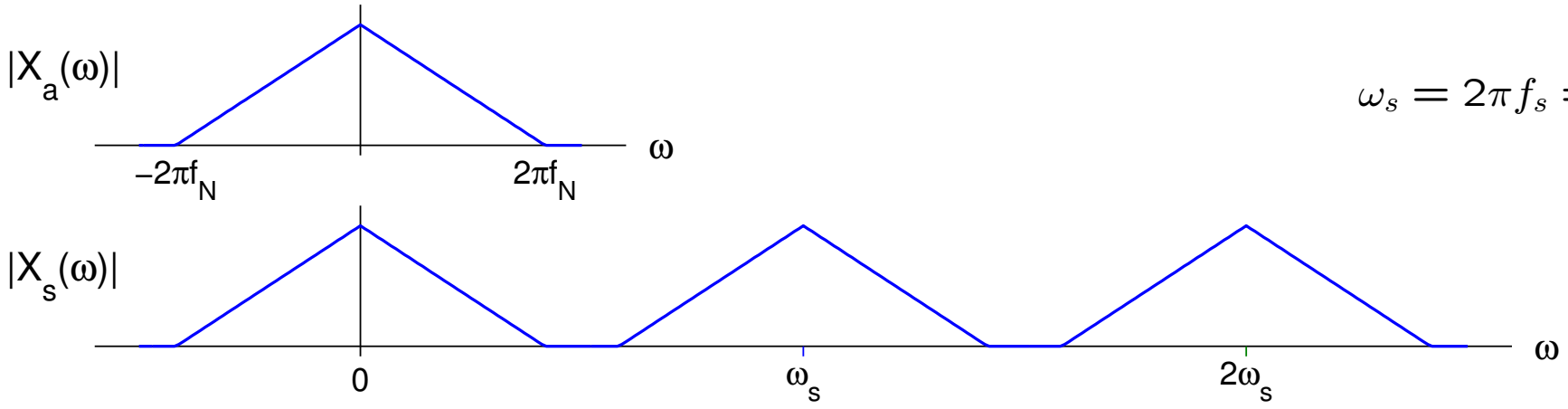
Analoges vs. abgetastetes Signal

Ein abgetastetes Signal $x_s(t)$ entsteht durch Multiplikation des analogen Signals $x_a(t)$ mit einer Pulsfolge $s(t)$:

$$x_s(t) = x_a(t) \cdot s(t) = x_a(t) \cdot \sum_{n=-\infty}^{\infty} \delta(t - nT_s)$$

$$\downarrow$$

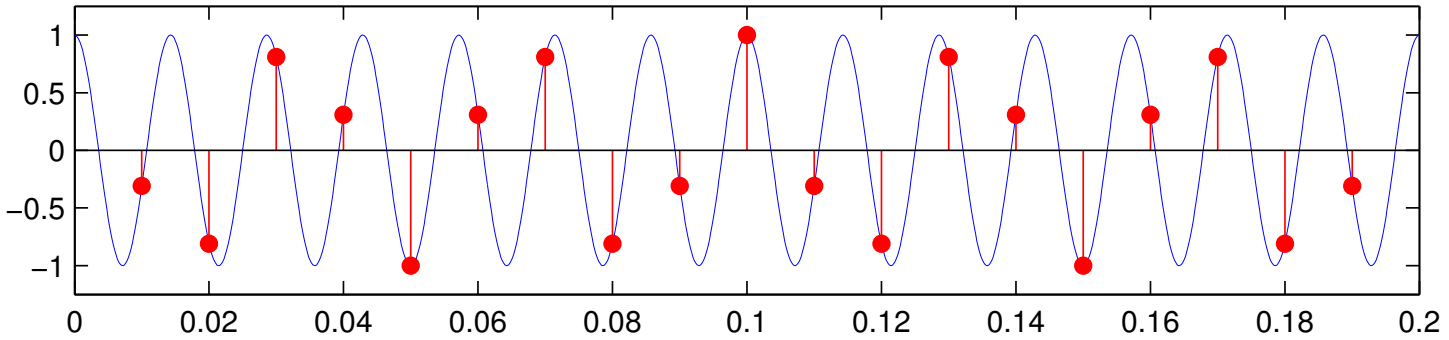
$$X_s(\omega) = \frac{1}{2\pi} X_a(\omega) * S(\omega) = \frac{\omega_s}{2\pi} X_a(\omega) * \sum_{k=-\infty}^{\infty} \delta(\omega - k\omega_s)$$



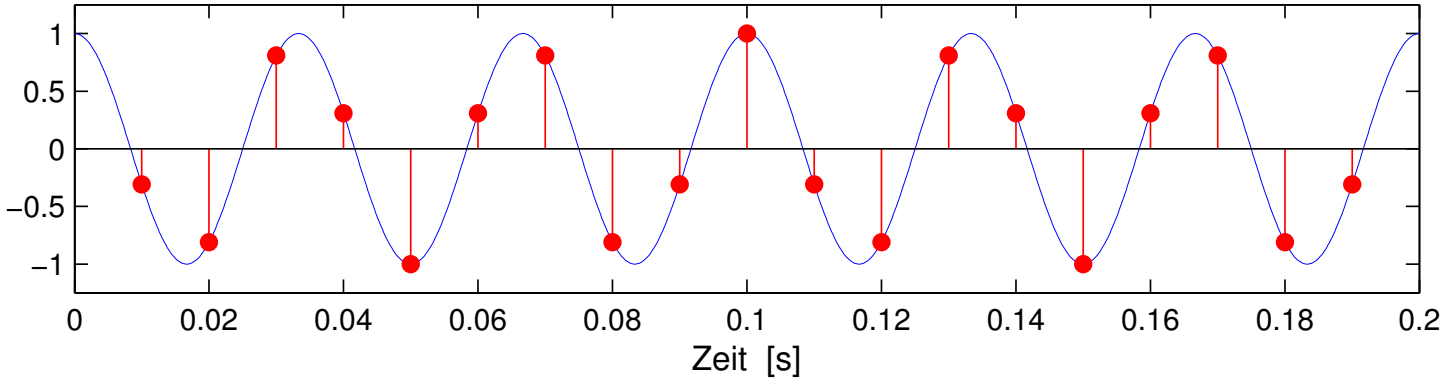
$$\omega_s = 2\pi f_s = \frac{2\pi}{T_s}$$

Aliasing-Effekt: Abtastung des Signals mit $f_s = 100$ Hz

Sinus 70 Hz:



Sinus 30 Hz:



Wie wirkt sich Aliasing bei Sprachsignalen aus?

Aliasing-Effekt bei Sprachsignalen

Breitbandiges Sprachsignal ($\approx$ analoges Signal) 

Sprachsignal mit 8 kHz abgetastet (ohne Anti-Aliasing-Filter) 

Mit Anti-Aliasing-Filter gefiltertes Sprachsignal mit 8 kHz abgetastet 

<<<

Welche Frequenzen sind in einem Sprachsignal wichtig?

Breitbandiges Sprachsignal: 0 – 16 kHz

Gefiltert mit Tiefpass:

0 – 6.8 kHz

0 – 3.4 kHz

0 – 1.7 kHz

0 – 1.0 kHz

Gefiltert mit Bandpass:

1 – 2.0 kHz

2 – 3.5 kHz

3.5 – 6 kHz

6.0 – 9 kHz

<<<

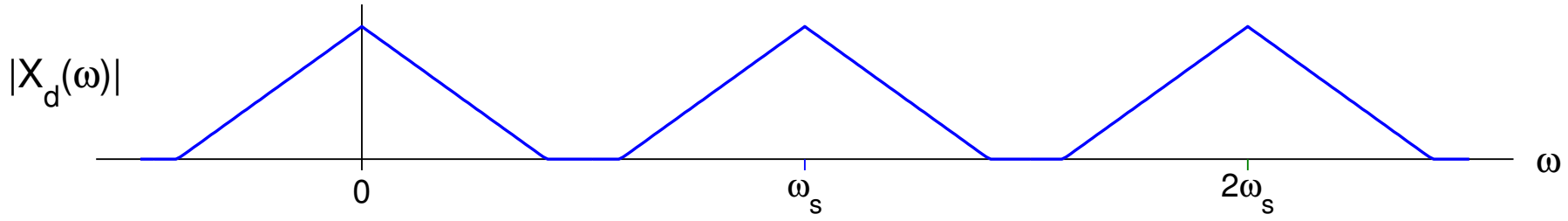
Spektrum des digitalen Signals

Digitales Signal $s(n)$ ist nur für $t = nT_s$ definiert!

Aus Annahme $x_d(t) = s(n)$ für $t = nT_s$

$x_d(t) = 0$ für $t \neq nT_s$

folgt:



Aus dem digitalen Signal $x_d(t)$ kann das analoge Signal $x_a(t)$ mit einem idealen Tiefpass der Grenzfrequenz $f_s/2$ exakt rekonstruiert werden.

Rekonstruktion des analogen Signals

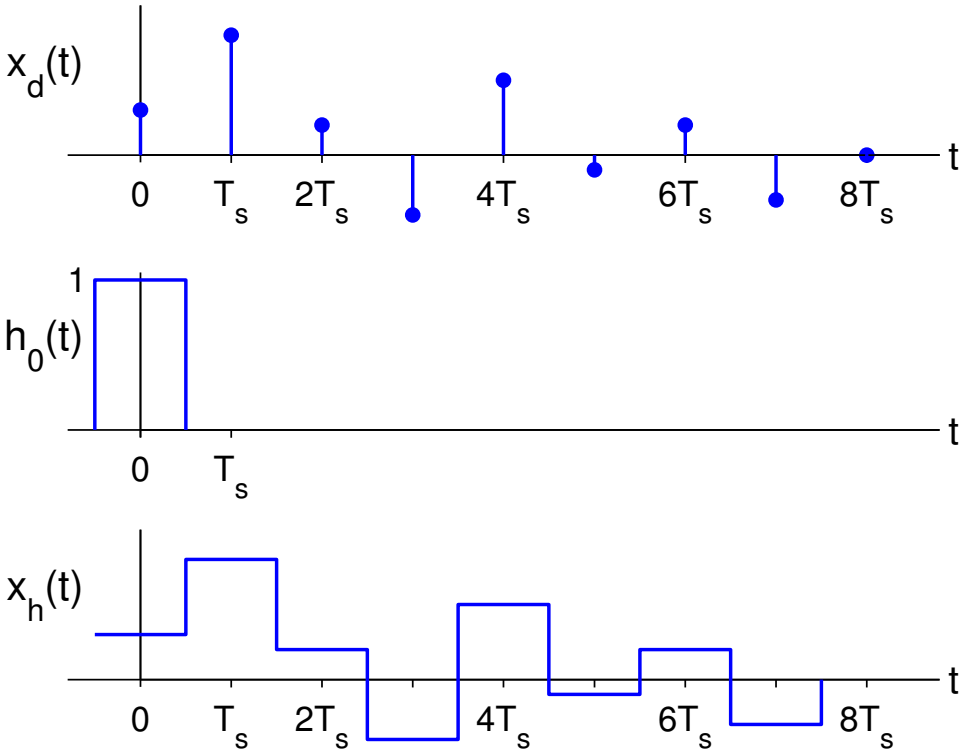
Problem: δ-Funktion unpraktisch!

Ausweg: Faltung von $x_d(t)$ mit $h_o(t)$ wobei

$$h_o(t) = \begin{cases} 1 & \text{für } -T_s/2 < t < +T_s/2 \\ 0 & \text{sonst} \end{cases}$$

$\longrightarrow$ Treppenfunktion

Frage: Was ist die Konsequenz?

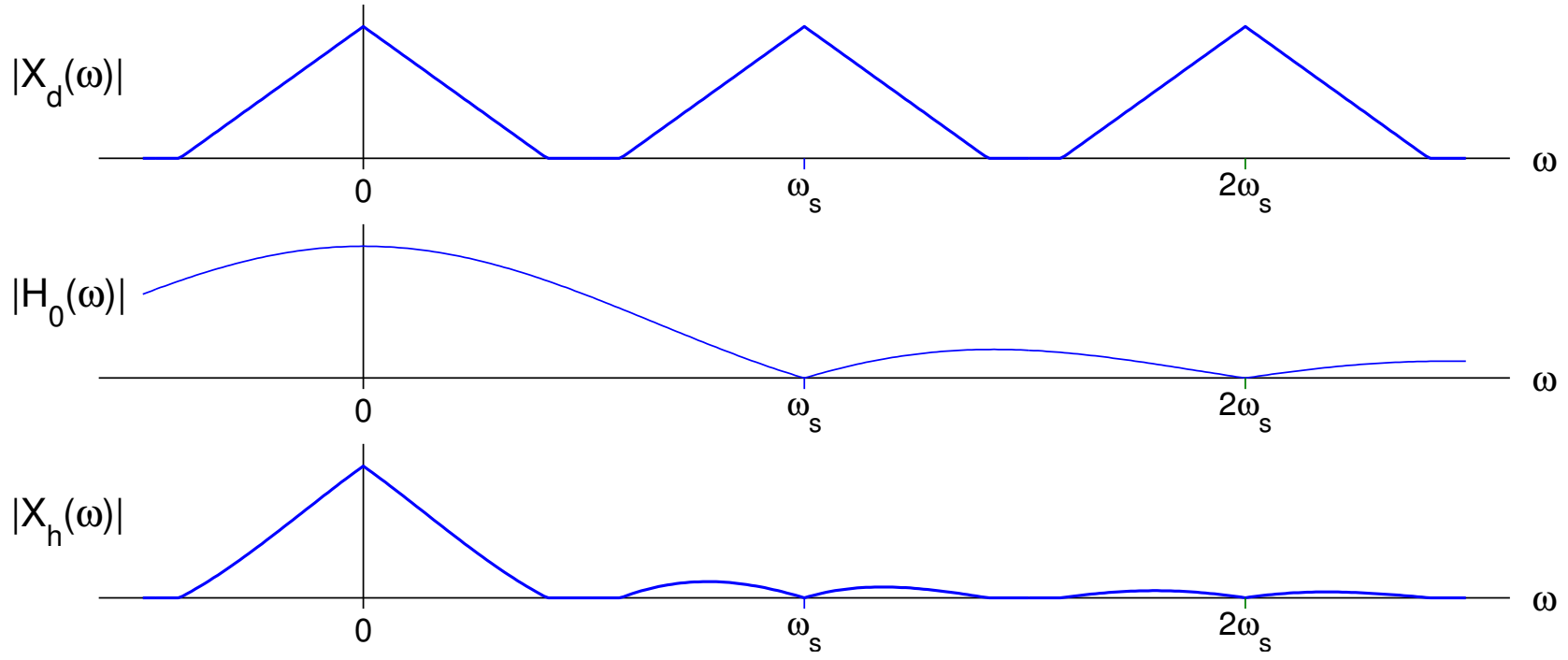


Abgetastestes Signal falten mit: $h_o(t) = \begin{cases} 1 & \text{für } -T_s/2 < t < +T_s/2 \\ 0 & \text{sonst} \end{cases}$

$$x_h(t) = x_d(t) * h_o(t) = \{x_a(t) \cdot s(t)\} * h_o(t)$$

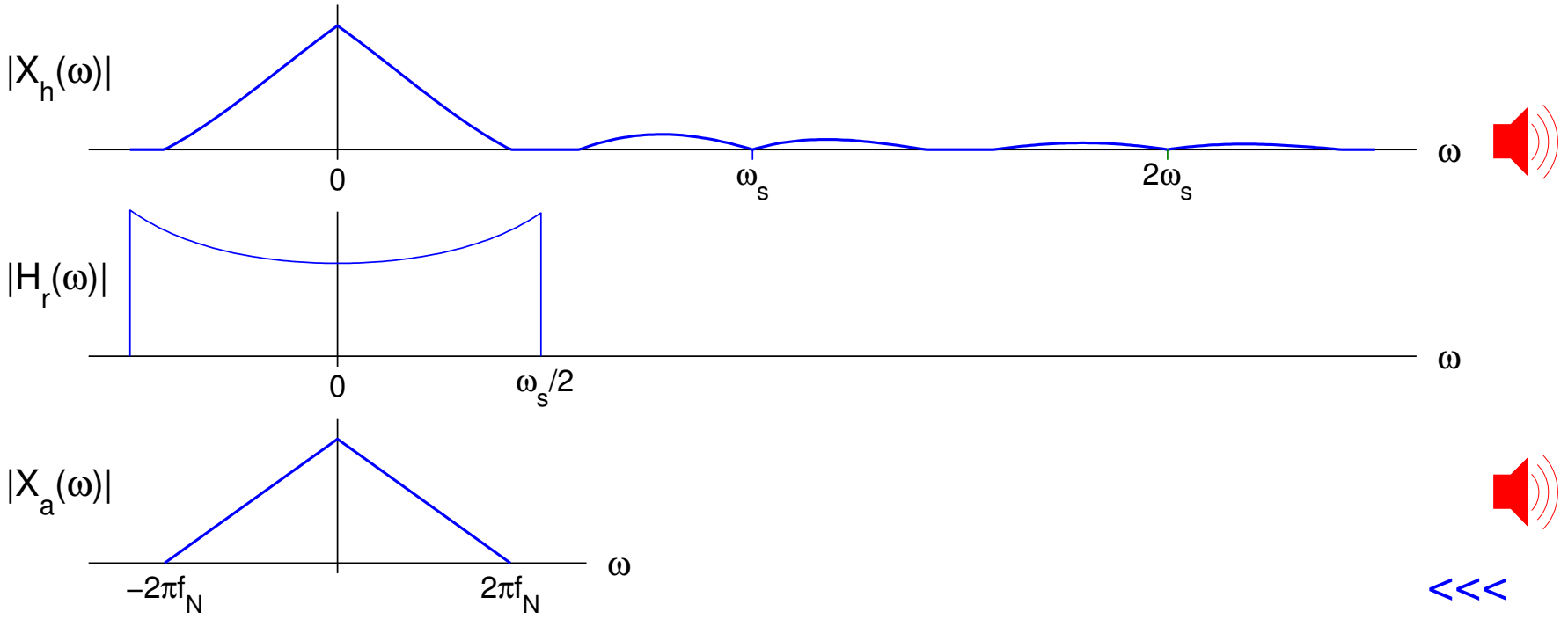
$$\updownarrow$$

$$X_h(\omega) = X_d(\omega) \cdot H_o(\omega) = \tfrac{1}{2\pi}\{X_a(\omega) * S(\omega)\} \cdot H_o(\omega)$$



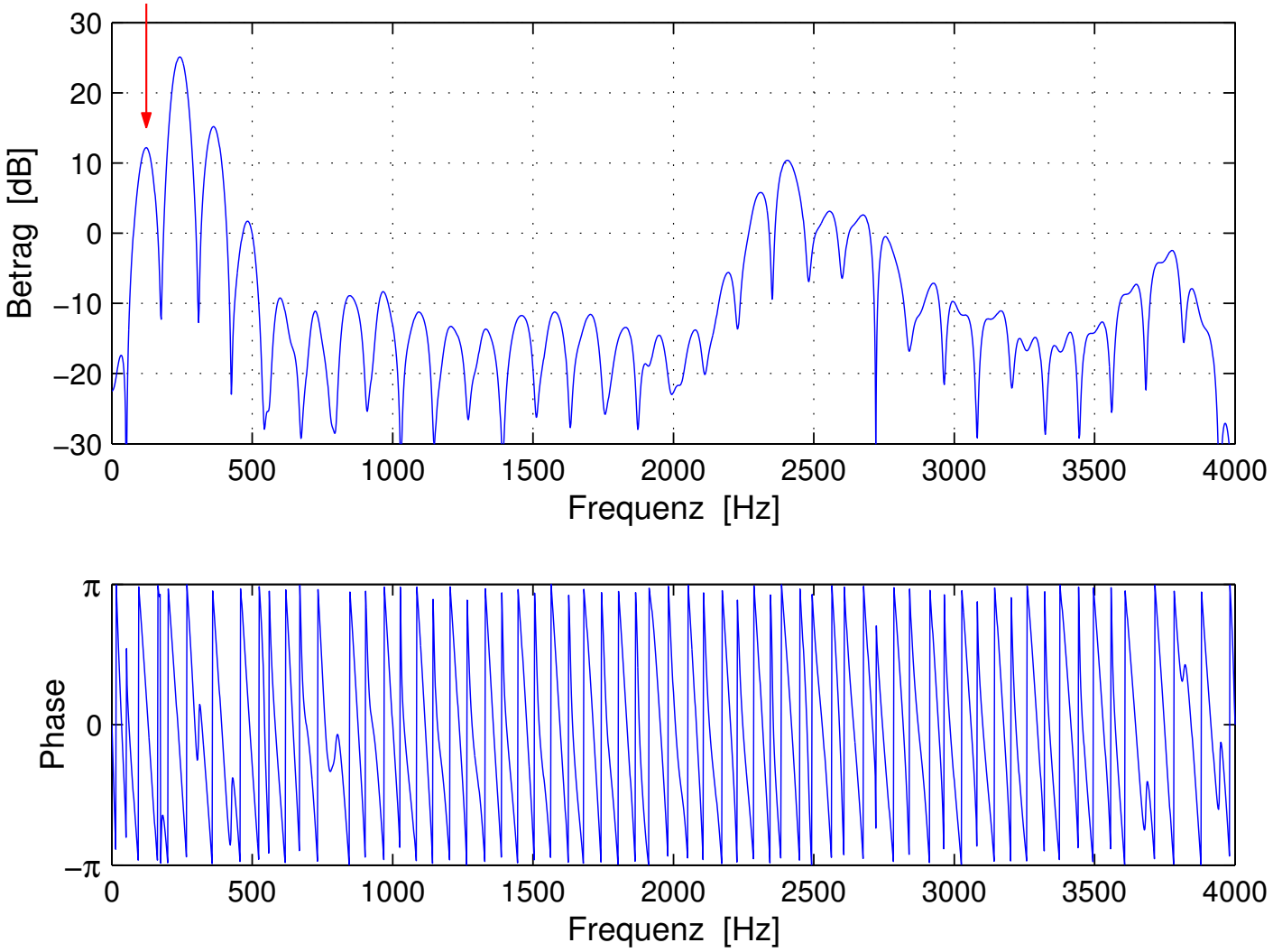
Ideales Rekonstruktionsfilter:

$$H_r(\omega) = \begin{cases} 1/H_o(\omega) & \text{für } -\omega_s/2 < \omega < +\omega_s/2 \\ 0 & \text{sonst} \end{cases}$$



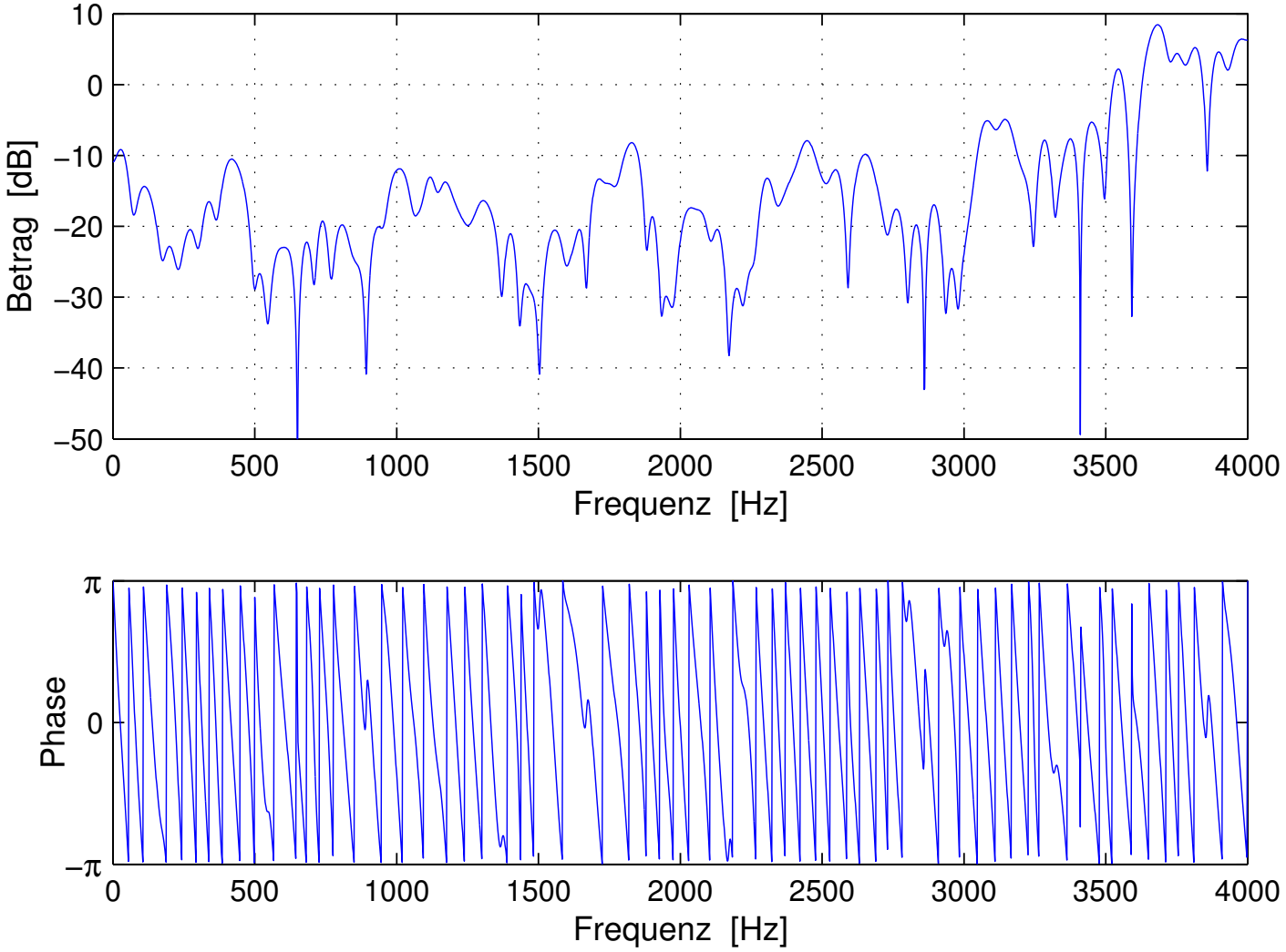
<<<

Spektrum des Sprachsignals (stimmhafter Ausschnitt)



<<<

Spektrum des Sprachsignals (stimmloser Ausschnitt)



<<<