ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Sprachverarbeitung I / 5 HS 2016

Homomorphe Analyse: Cepstrum

Buch: Kapitel 4.6

Beat Pfister

Sprachverarbeitung I / 5

Vorlesung: Homomorphe Systeme / Analyse

- DFT-Cepstrum
- Mel-Cepstrum

Übung: LPC-Analyse und -Synthese von Sprachsignalen

Das Superpositionsprinzip

lineare Systeme:

$$
\begin{aligned}
T\{x_1(n) + x_2(n)\} &= T\{x_1(n)\} + T\{x_2(n)\} \\
T\{c \cdot x(n)\} &= c \cdot T\{x(n)\}
\end{aligned}
$$

Das verallgemeinerte Superpositionsprinzip

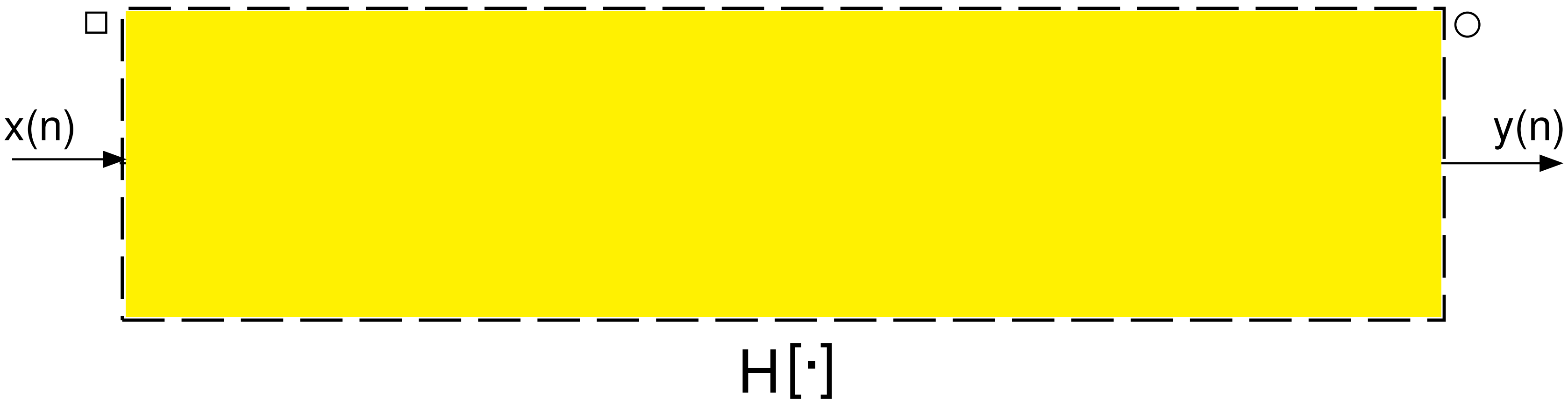
lineare Systeme:

$$\begin{aligned} T\{x_1(n) + x_2(n)\} &= T\{x_1(n)\} + T\{x_2(n)\} \\ T\{c \cdot x(n)\} &= c \cdot T\{x(n)\} \end{aligned}$$

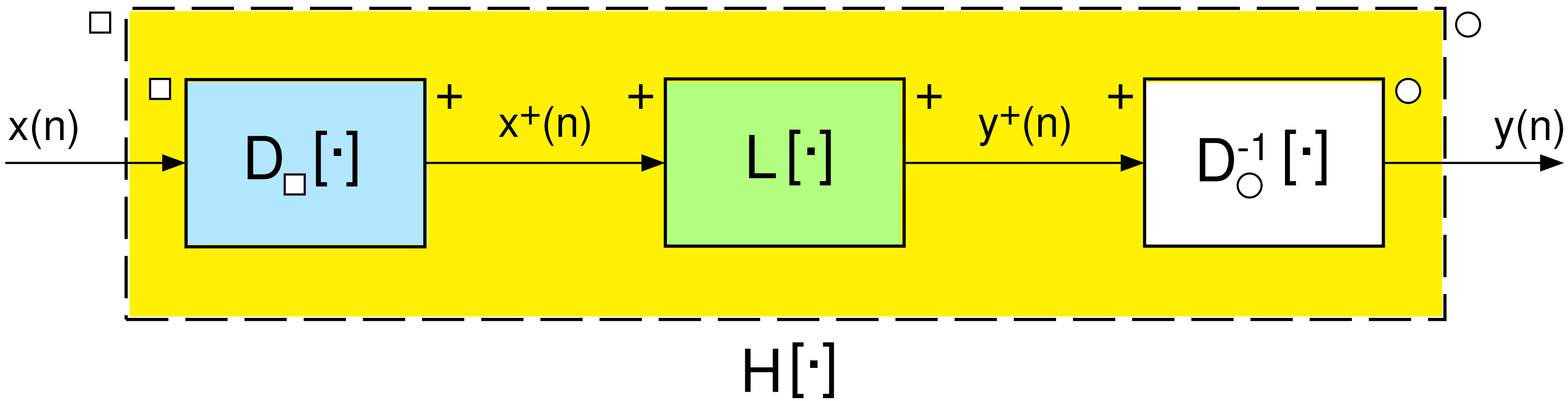
homomorphe Systeme:

$$\begin{aligned} H\{x_1(n) \square x_2(n)\} &= H\{x_1(n)\} \circ H\{x_2(n)\} \\ H\{c \diamond x(n)\} &= c \triangleleft H\{x(n)\} \end{aligned}$$

Homomorphes System



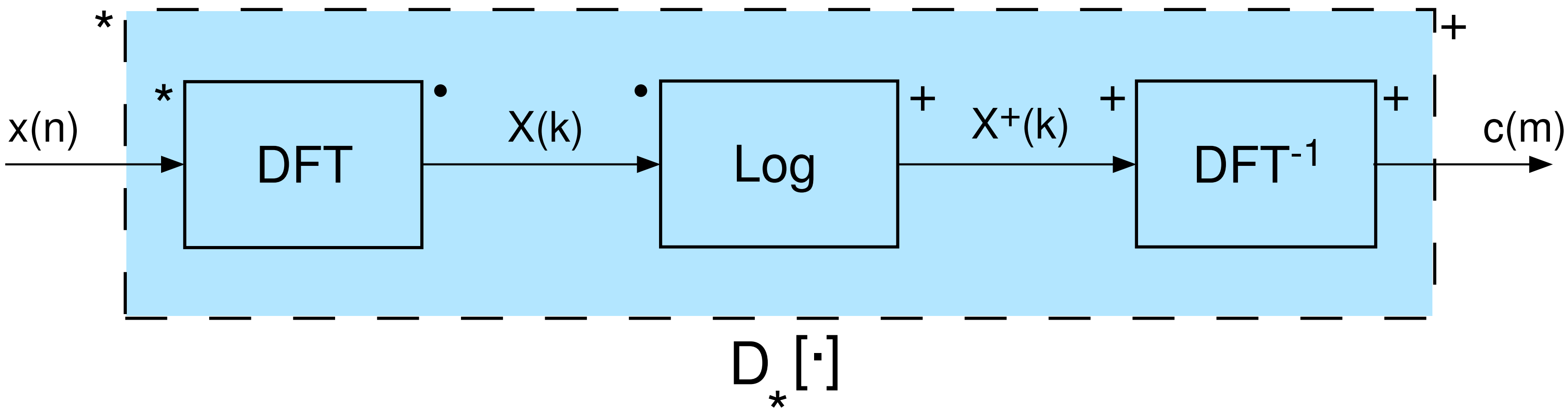
Homomorphes System



Kanonische Darstellung eines homomorphen Systems $H[\cdot]$:

$D_\square[\cdot]$:	charakteristisches System für die Eingangsoperation □
$L[\cdot]$:	*lineares* System (z.B. lineares Filter)
$D_\circ^{-1}[\cdot]$:	inverses charakteristisches System für die Ausgangsoperation ○

Charakteristisches System für die Faltung



DFT: $X(k) = \sum_{n=0}^{N-1} x(n) e^{-j(2\pi/N)kn}$ $\longrightarrow$ lineares Spektrum

LOG: $X^{+}(k) = \log\{X(k)\}$ $\longrightarrow$ logarith. Spektrum

DFT^{-1}: $c(m) = \frac{1}{N} \sum_{k=0}^{N-1} X^{+}(k) e^{j(2\pi/N)km}$ $\longrightarrow$ komplexes Cepstrum

Reelles Cepstrum

Variante des charakteristischen Systems für die Faltung, wobei das Betragsspektrum eingesetzt wird.

DFT: $$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j(2\pi/N)kn}$$

$$|X(k)| = \sqrt{X(k) \cdot X^*(k)} \qquad \longrightarrow \text{ lin. Betragsspektrum}$$

LOG: $$X^+(k) = \log\{|X(k)|\} \qquad \longrightarrow \text{ log. Betragsspektrum}$$

DFT^{-1}: $$c(m) = \frac{1}{N} \sum_{k=0}^{N-1} X^+(k)\, e^{j(2\pi/N)km} \qquad \longrightarrow \text{ reelles Cepstrum}$$

Reelles Cepstrum

Variante des charakteristischen Systems für die Faltung, wobei das Betragsspektrum eingesetzt wird.

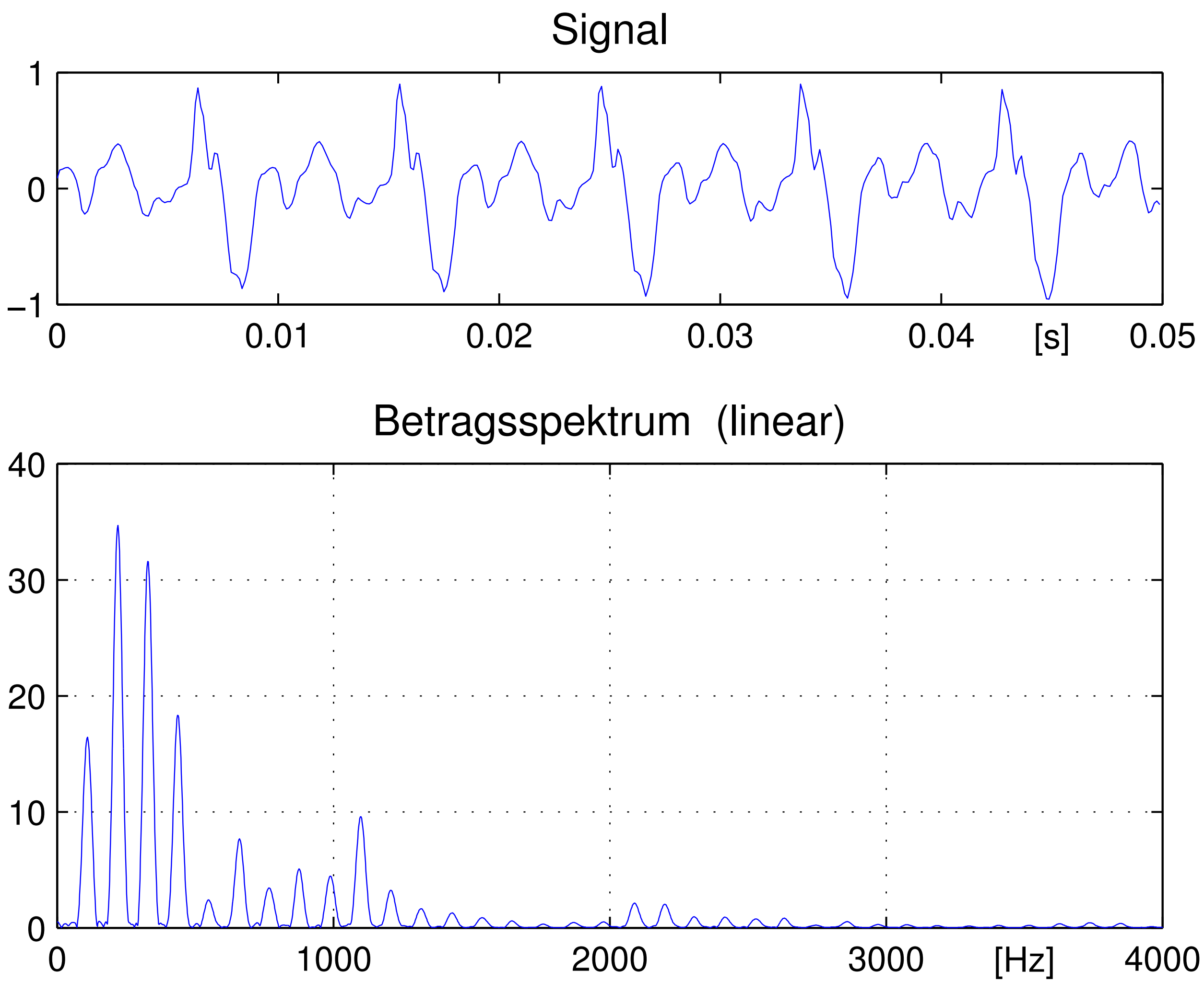
DFT: $X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j(2\pi/N)kn}$

$|X(k)| = \sqrt{X(k) \cdot X^*(k)}$ $\longrightarrow$ lin. Betragsspektrum

LOG: $X^+(k) = \log\{|X(k)|\}$ $\longrightarrow$ log. Betragsspektrum

DFT^{-1}: $c(m) = \frac{1}{N} \sum_{k=0}^{N-1} X^+(k)\, e^{j(2\pi/N)km}$ $\longrightarrow$ reelles Cepstrum

DFT



Signal
1
0
−1
0
0.01
0.02
0.03
0.04
[s]
0.05
Betragsspektrum (linear)
40
30
20
10
0
0
1000
2000
3000
[Hz]
4000

Reelles Cepstrum

Variante des charakteristischen Systems für die Faltung, wobei das Betragsspektrum eingesetzt wird.

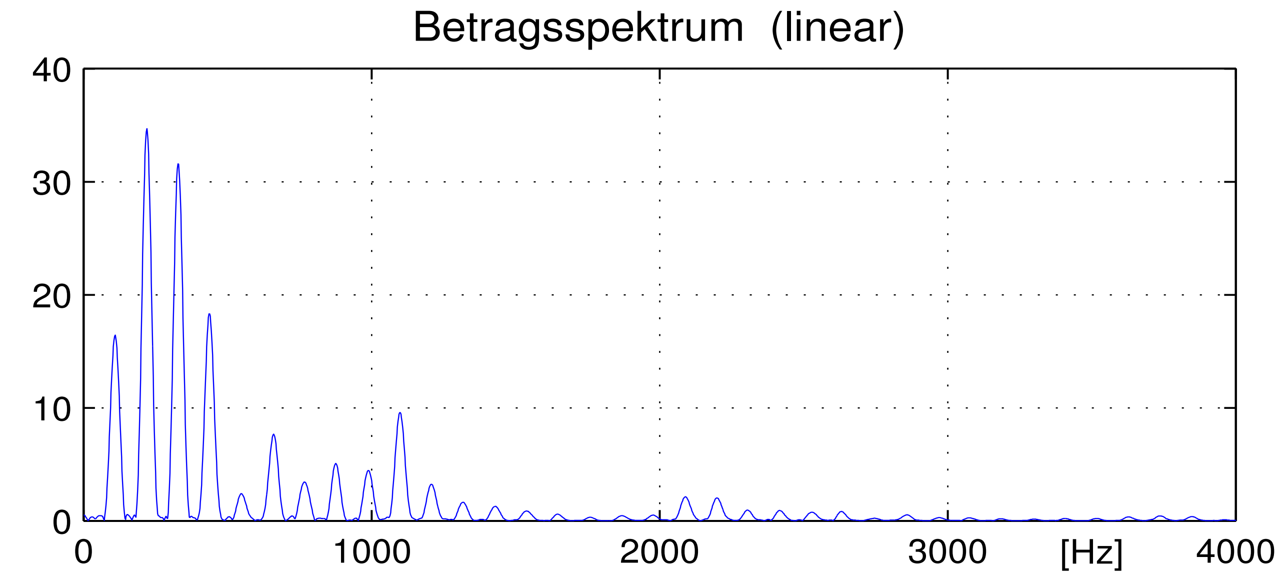
DFT: $$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j(2\pi/N)kn}$$

$$|X(k)| = \sqrt{X(k) \cdot X^*(k)} \qquad \longrightarrow \text{ lin. Betragsspektrum}$$

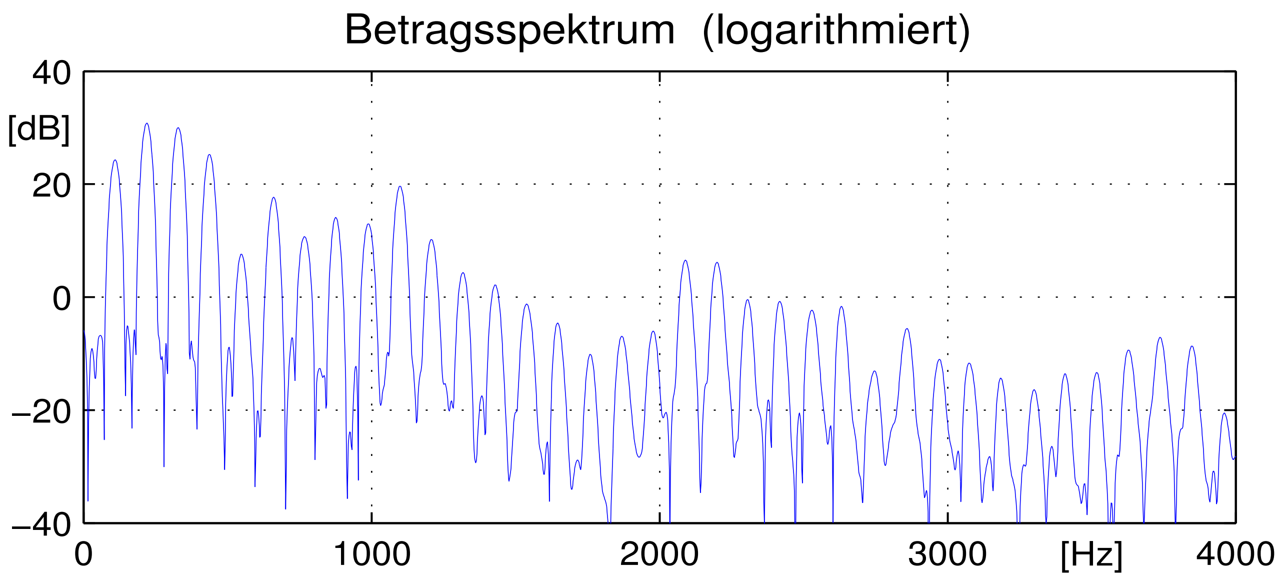
LOG: $$X^+(k) = \log\{|X(k)|\} \qquad \longrightarrow \text{ log. Betragsspektrum}$$

DFT^{-1}: $$c(m) = \frac{1}{N} \sum_{k=0}^{N-1} X^+(k)\, e^{j(2\pi/N)km} \qquad \longrightarrow \text{ reelles Cepstrum}$$

LOG



Betragsspektrum (linear)
40
30
20
10
0
0
1000
2000
3000
[Hz]
4000



Betragsspektrum (logarithmiert)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

Reelles Cepstrum

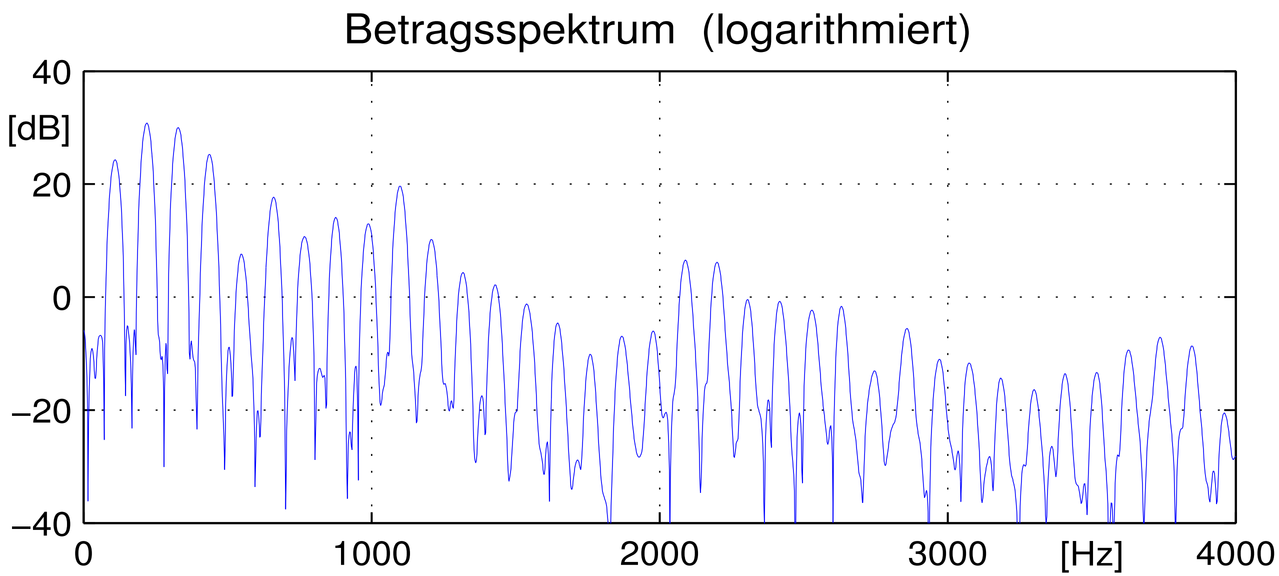
Variante des charakteristischen Systems für die Faltung, wobei das Betragsspektrum eingesetzt wird.

DFT:
$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j(2\pi/N)kn}$$

$$|X(k)| = \sqrt{X(k) \cdot X^*(k)} \longrightarrow \text{lin. Betragsspektrum}$$

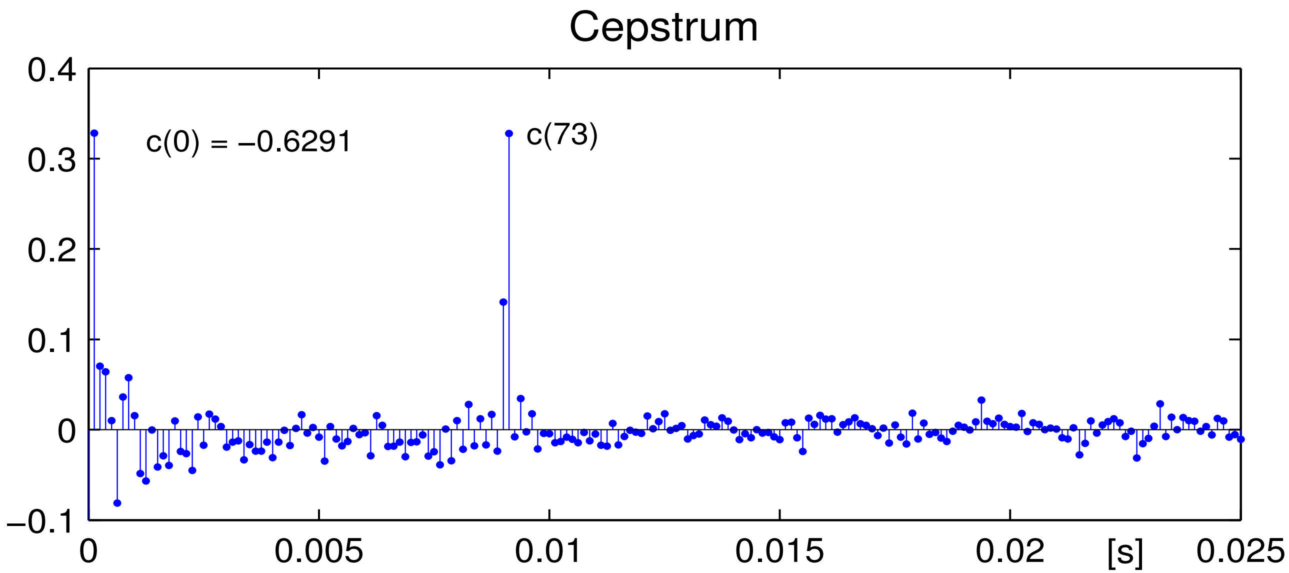
LOG:
$$X^+(k) = \log\{|X(k)|\} \longrightarrow \text{log. Betragsspektrum}$$

DFT^{-1}:
$$c(m) = \frac{1}{N} \sum_{k=0}^{N-1} X^+(k)\, e^{j(2\pi/N)km} \longrightarrow \text{reelles Cepstrum}$$



Betragsspektrum (logarithmiert)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

DFT^{-1}



Cepstrum
0.4
0.3
0.2
0.1
0
−0.1
c(0) = −0.6291
c(73)
0
0.005
0.01
0.015
0.02
[s]
0.025

Was sagt uns das Cepstrum ?

Merke: log. Spektrum und Cepstrum bilden ein Fourier-Paar

$$X^{+}(k) \;\bullet\!\!-\!\!\circ\; c(m)$$

Was ergibt die DFT von $c(i)$? >>>

Was ergibt die DFT von $c(0)$? >>>

Was ergibt die DFT von $[c(0)..c(i)]$? >>>

Was steckt im reellen Cepstrum ?

Cepstrum: Ausgangssequenz des charakt. Systems für die Faltung

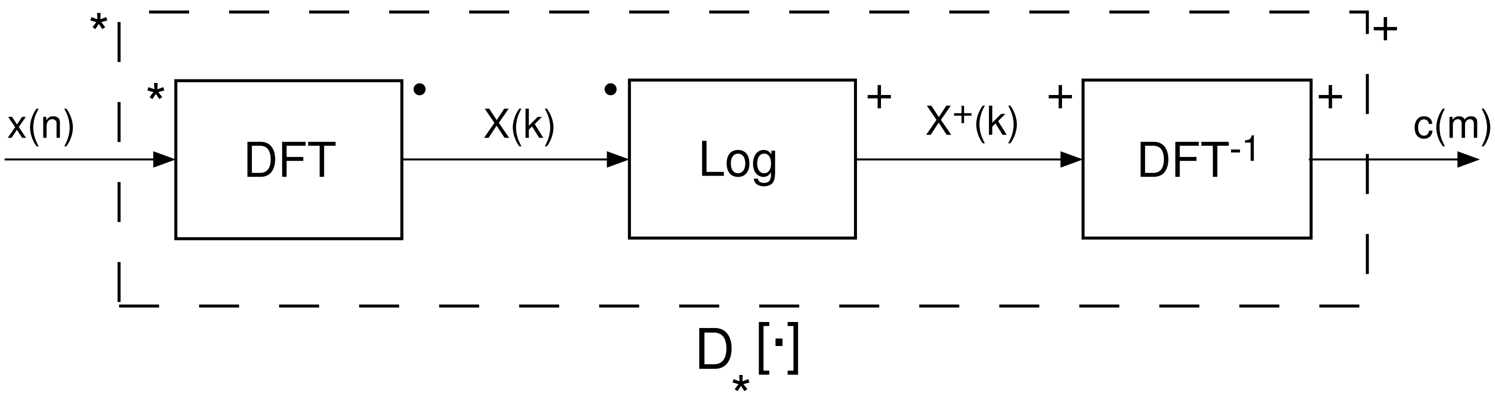
- Signalenergie
- Signalperiode

Was steckt im reellen Cepstrum ?

Cepstrum: Ausgangssequenz des charakt. Systems für die Faltung

- Signalenergie
- Signalperiode

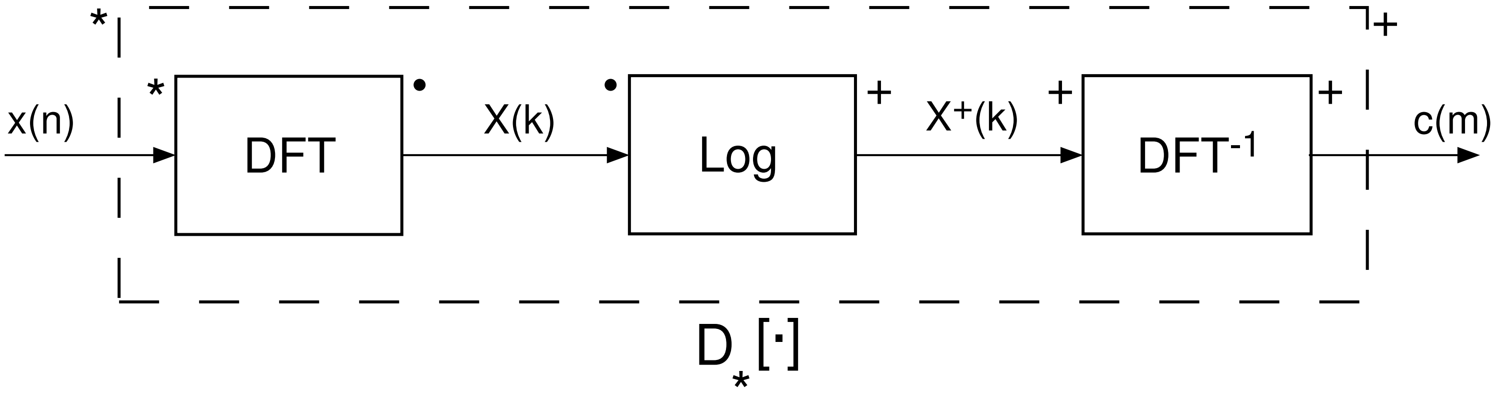
Charakteristisches System für die Faltung



Multiplikation der Eingangssequenz $x(n)$ mit der Konstanten a

Auswirkung auf Cepstrum ?

Charakteristisches System für die Faltung



Multiplikation der Eingangssequenz $x(n)$ mit der Konstanten a

$c(0)$	$\longrightarrow$	$+\log(a)$	$\longrightarrow$	Log. der Signalenergie
$c(m), \quad m > 0$	$\longrightarrow$	unverändert	$\longrightarrow$	leistungsunabhängig

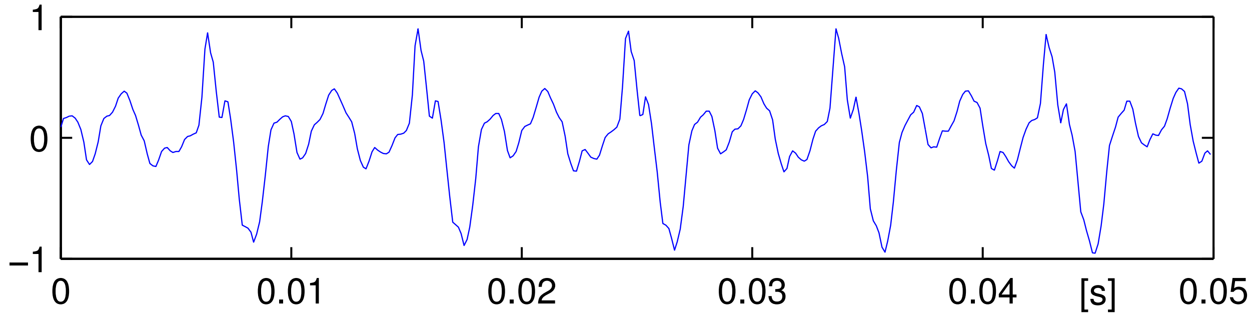
Was steckt im reellen Cepstrum ?

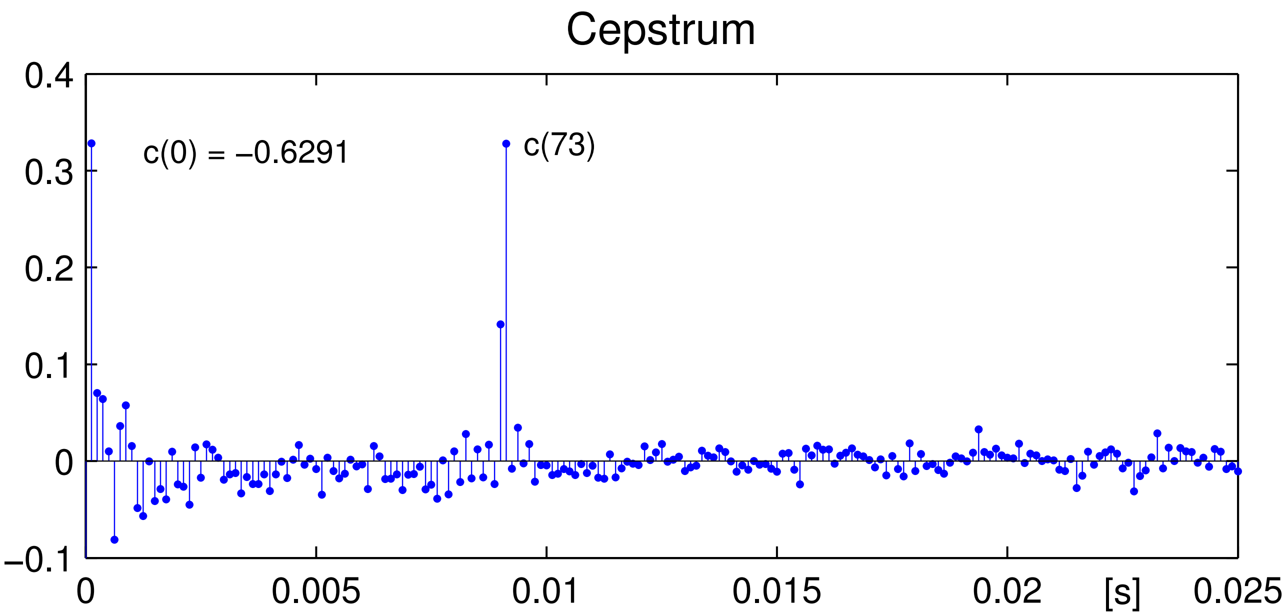
Cepstrum: Ausgangssequenz des charakt. Systems für die Faltung

- Signalenergie
- Signalperiode

Cepstrum

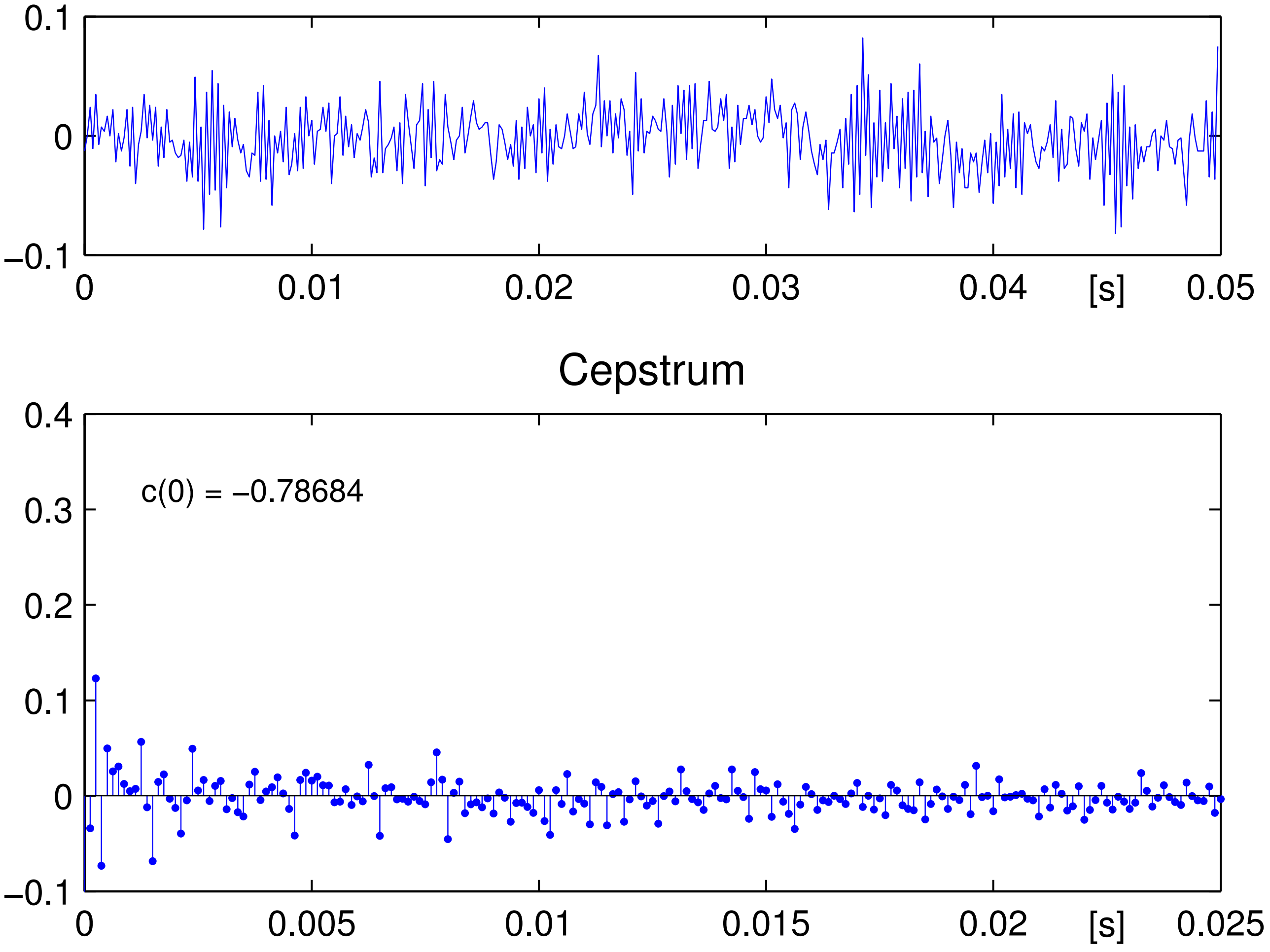
stimmhaftes Sprachsignal





Cepstrum

stimmloses Sprachsignal



Kann man mit dem Cepstrum
die Periode eines Sinussignals bestimmen ?

>>>

Kann man mit dem Cepstrum
die Periode eines Sinussignals bestimmen ?

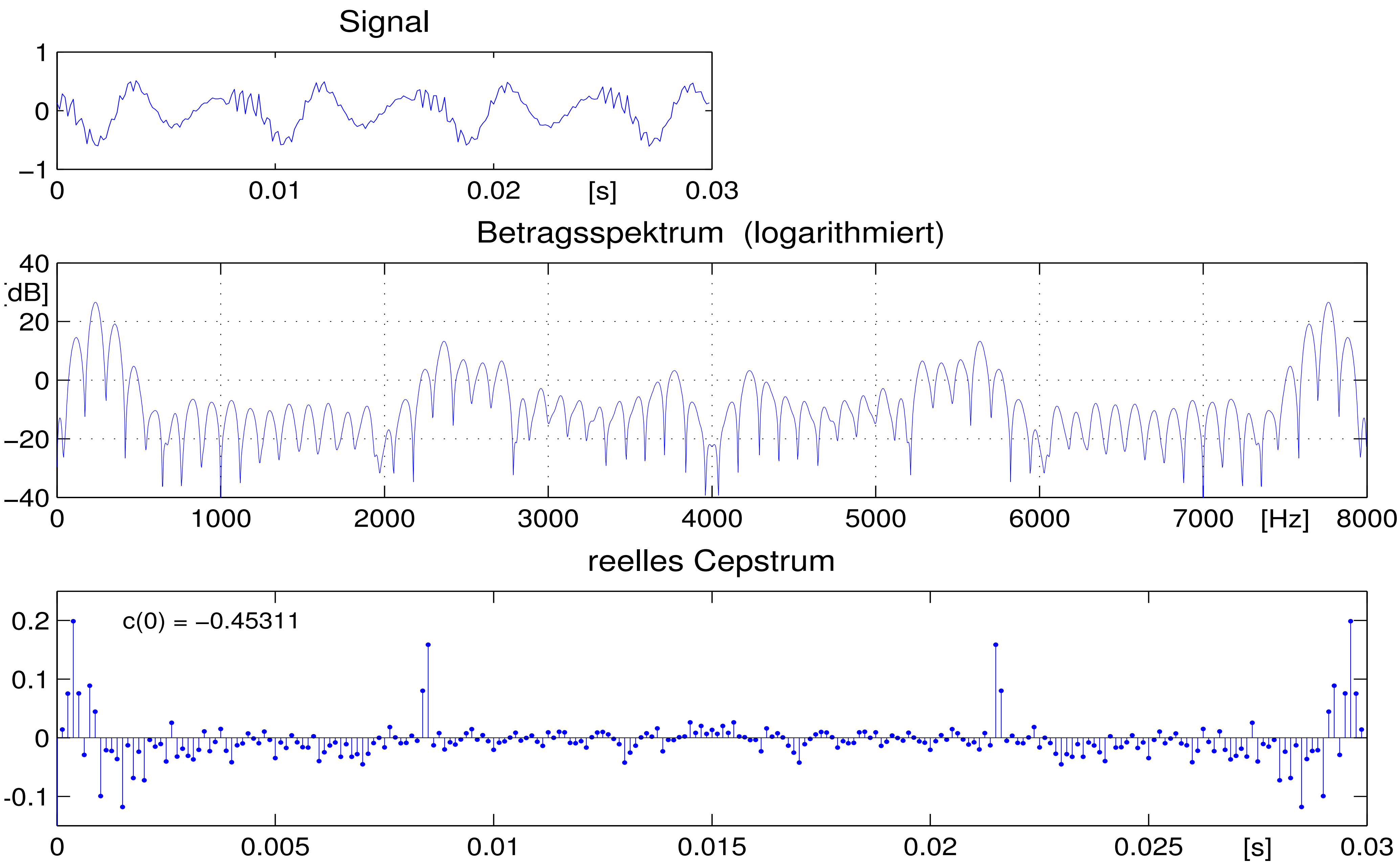
Ist nur für harmonische Signale sinnvoll !

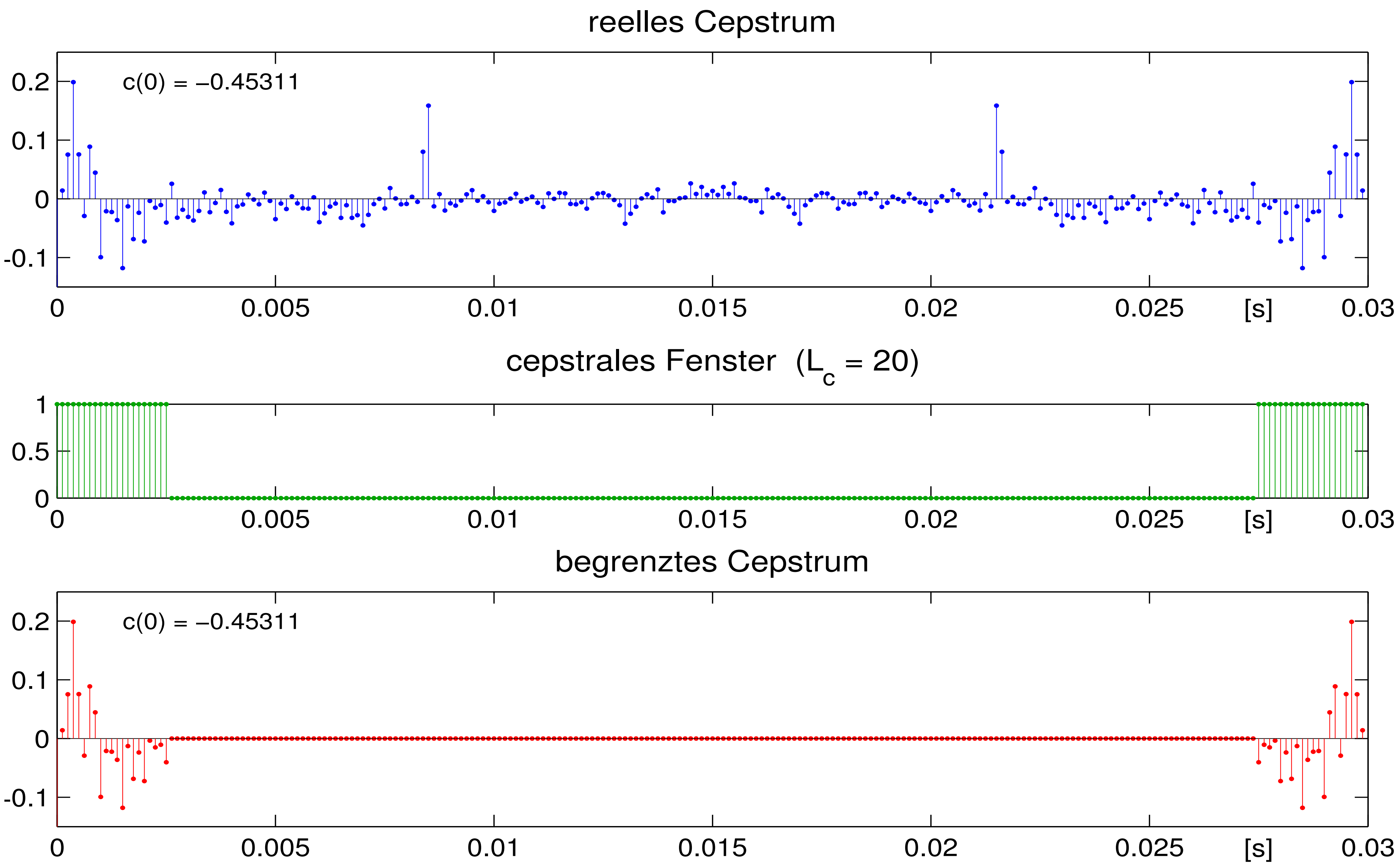
Verwendung des DFT-Cepstrums

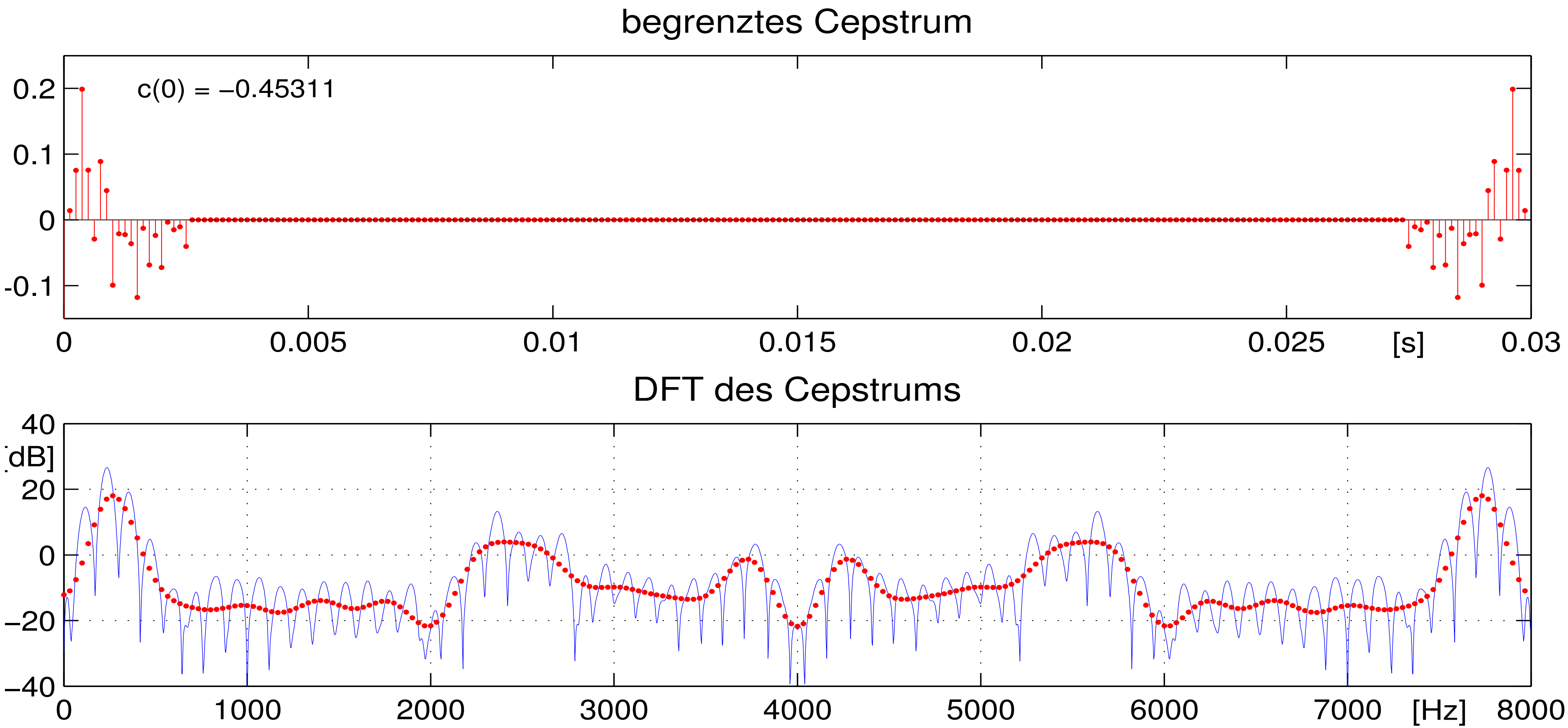
- zum Glätten von Spektren bzw. Spektrogrammen
- zum Schätzen der Periode von Sprachsignalen
- zur Kompensation des Übertragungskanals
- zum Ermitteln des Unterschiedes zwischen Spektren
- als Merkmal in der Sprach- und Sprechererkennung
- etc.

Verwendung des DFT-Cepstrums

- zum Glätten von Spektren bzw. Spektrogrammen
- zum Schätzen der Periode von Sprachsignalen
- zur Kompensation des Übertragungskanals
- zum Ermitteln des Unterschiedes zwischen Spektren
- als Merkmal in der Sprach- und Sprechererkennung
- etc.

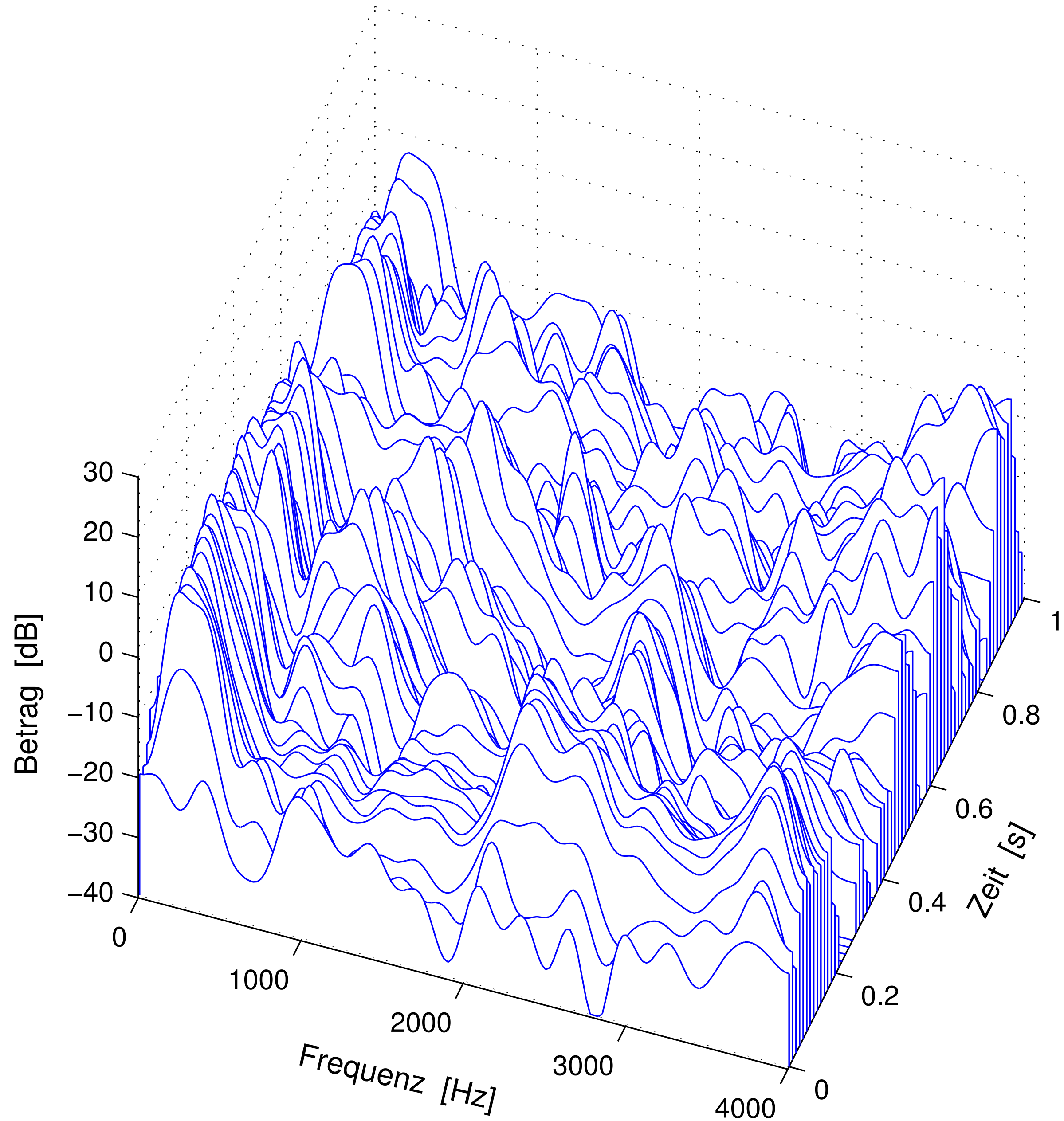

Signal
1
0
–1
0
0.01
0.02
[s]
0.03
Betragsspektrum (logarithmiert)
40
[dB]
20
0
–20
–40
0
1000
2000
3000
4000
5000
6000
7000
[Hz]
8000
reelles Cepstrum
0.2
c(0) = –0.45311
0.1
0
-0.1
0
0.005
0.01
0.015
0.02
0.025
[s]
0.03


reelles Cepstrum
c(0) = −0.45311
0.2
0.1
0
-0.1
0
0.005
0.01
0.015
0.02
0.025
[s]
0.03
cepstrales Fenster (L_c = 20)
1
0.5
0
begrenztes Cepstrum
c(0) = −0.45311


begrenztes Cepstrum
c(0) = −0.45311
0.2
0.1
0
-0.1
0
0.005
0.01
0.015
0.02
0.025
[s]
0.03
DFT des Cepstrums
40
[dB]
20
0
−20
−40
0
1000
2000
3000
4000
5000
6000
7000
[Hz]
8000

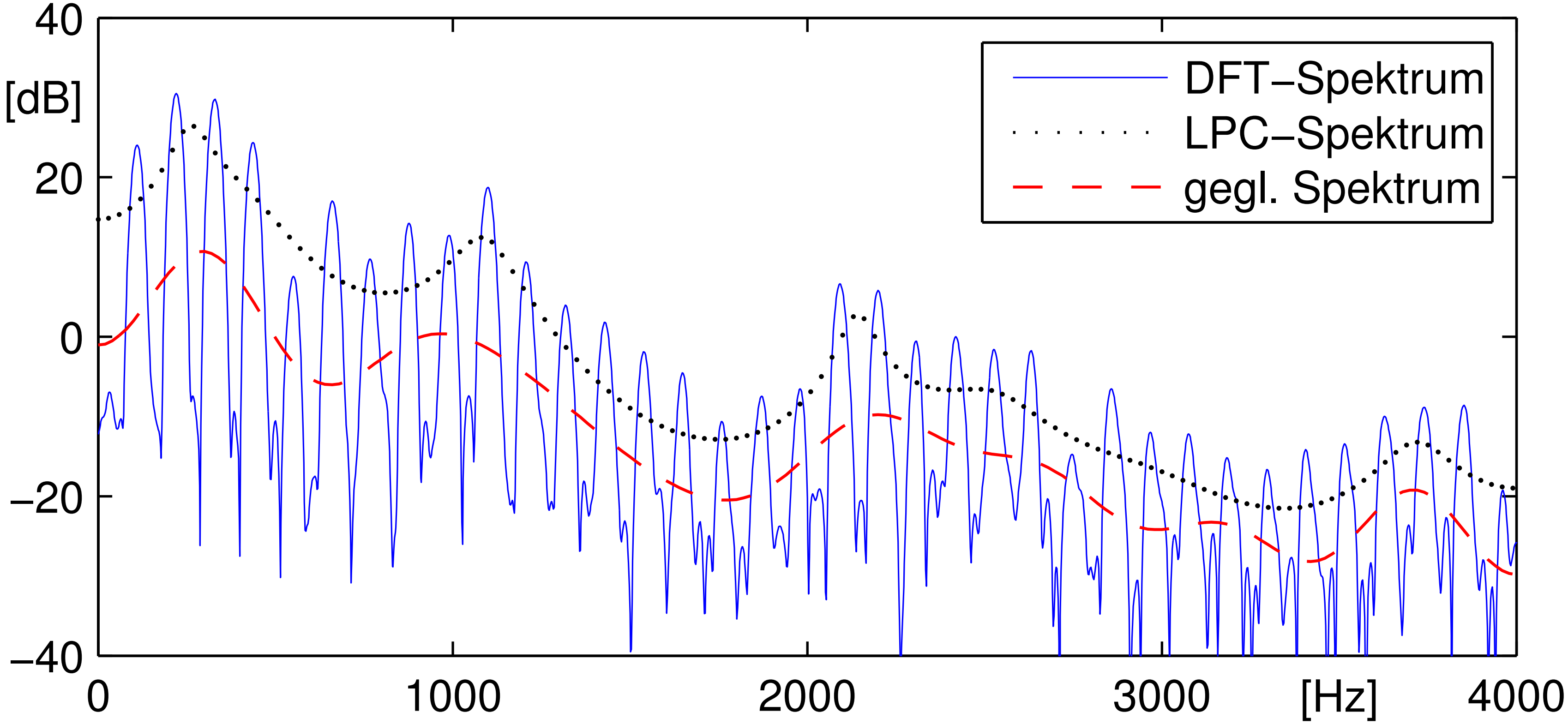
Geglättetes Kurzzeitspektrum

3-dimensionale Darstellung



>>>

Vergleich geglätteter Spektren

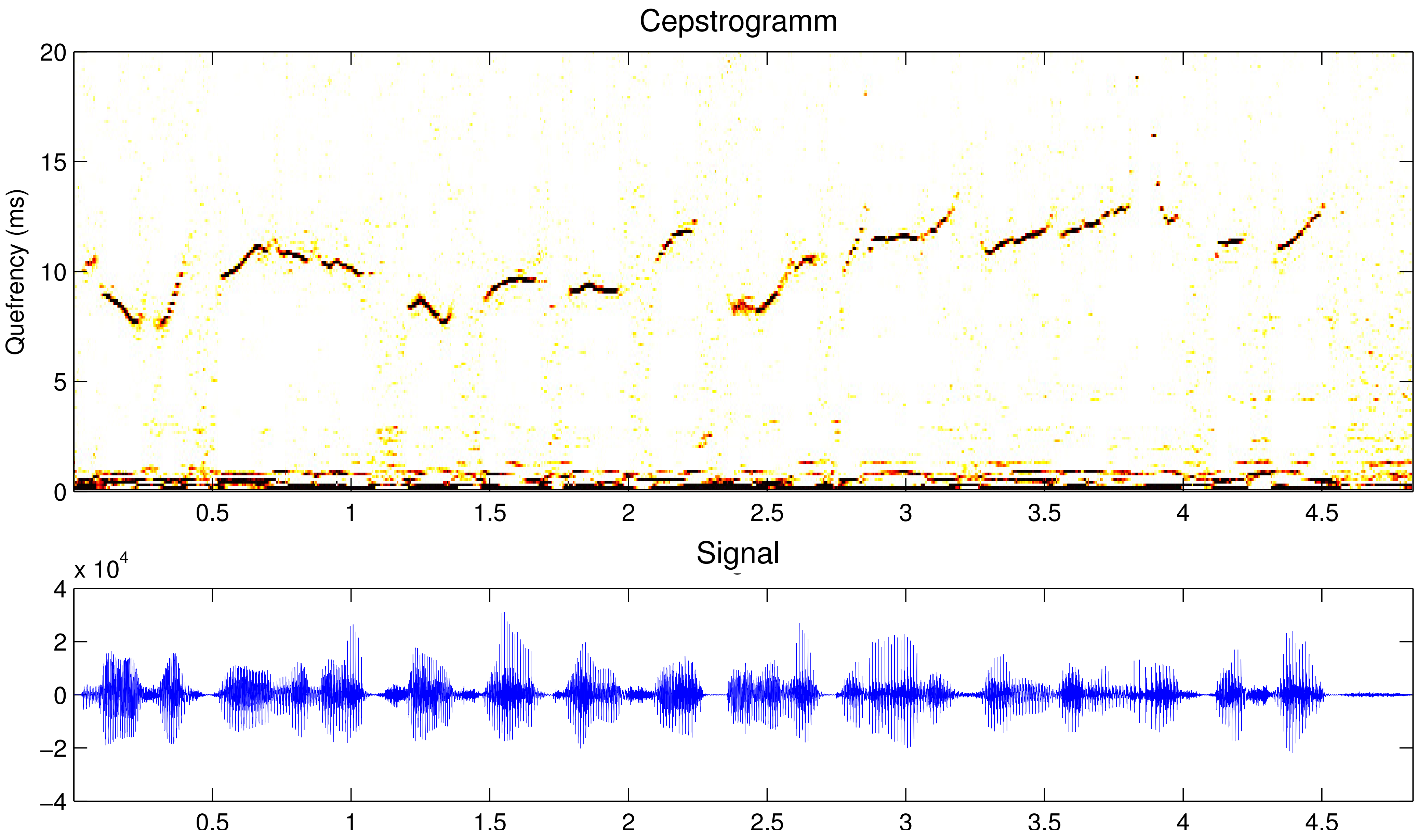


logarithmische Darstellung der Spektren

>>>

Verwendung des DFT-Cepstrums

- zum Glätten von Spektren bzw. Spektrogrammen
- zum Schätzen der Periode von Sprachsignalen
- zur Kompensation des Übertragungskanals
- zum Ermitteln des Unterschiedes zwischen Spektren
- als Merkmal in der Sprach- und Sprechererkennung
- etc.


Cepstrogramm
Quefrency (ms)
20
15
10
5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Signal
x 10^4
4
2
0
−2
−4
0.5
1
1.5
2
2.5
3
3.5
4
4.5

Verwendung des DFT-Cepstrums

- zum Glätten von Spektren bzw. Spektrogrammen
- zum Schätzen der Periode von Sprachsignalen
- zur Kompensation des Übertragungskanals
- zum Ermitteln des Unterschiedes zwischen Spektren
- als Merkmal in der Sprach- und Sprechererkennung
- etc.

Kompensation der Kanalcharakteristik

Gegeben: Kanal mit Übertragungsfunktion $g(n)$ >>>

Problem: Kanal $g(n)$ beeinflusst Signal >>>

$s_2(n) = s_1(n) * g(n)$

Frage: Einfluss auf das Cepstrum?

Zeitbereich: $*$ Cepstrum: $+$

$c_{2j}(m) = c_{1j}(m) + c_G(m)$ >>>

Kompensation: Subtraktion von $c_G(m)$

Nicht möglich, da $g(n)$ und damit auch $c_G(m)$ i.a. unbekannt !

Lösung: Mittelwertfreies Cepstrum >>>

Verwendung des DFT-Cepstrums

- zum Glätten von Spektren bzw. Spektrogrammen
- zum Schätzen der Periode von Sprachsignalen
- zur Kompensation des Übertragungskanals
- zum Ermitteln des Unterschiedes zwischen Spektren
- als Merkmal in der Sprach- und Sprechererkennung
- etc.

Cepstrale Distanz

Bestimmung der RMS-Distanz zwischen zwei logarithmierten Betragsspektren:

$$d_c = \frac{1}{N}\sqrt{\sum_{k=0}^{N-1}\left[X_1^+(k) - X_2^+(k)\right]^2} \qquad (1)$$

$$X^+(k) = \sum_{m=0}^{N-1} c(m)e^{-j(2\pi/N)km} \qquad (2)$$

Einsetzen von (2) in (1) und umformen ergibt:

$$d_c = \sqrt{\sum_{m=0}^{N-1}\left[c_1(m) - c_2(m)\right]^2}$$

cepstrale Distanz

Verwendung des DFT-Cepstrums

- zum Glätten von Spektren bzw. Spektrogrammen
- zum Schätzen der Periode von Sprachsignalen
- zur Kompensation des Übertragungskanals
- zum Ermitteln des Unterschiedes zwischen Spektren
- als Merkmal in der Sprach- und Sprechererkennung
- etc.

Cepstrum in der Spracherkennung

Ein Sprachsignalabschnitt lässt sich beschreiben mit:

$$c(1) \;\ldots\; c(m)$$

- unabhängig von der Signalleistung
- unabhängig von der Grundfrequenz (falls $m < T_0/T_s = F_s/F_0$)

Variante zur Ermittlung des Cepstrums

Unterschiedlicher erster Systemteil (Signal $\longrightarrow$ Betragsspektrum):

a) DFT-Cepstrum: DFT + Betragsbildung $x(n) \longrightarrow |X(k)|$

b) Mel-Cepstrum: DFT + Betragsbildung $x(n) \longrightarrow |X(k)|$

Hz-Mel-Transformation $|X(k)| \longrightarrow \bar{S}_j$ >>>

Anmerkung: Diese Varianten sind nicht gleichwertig!

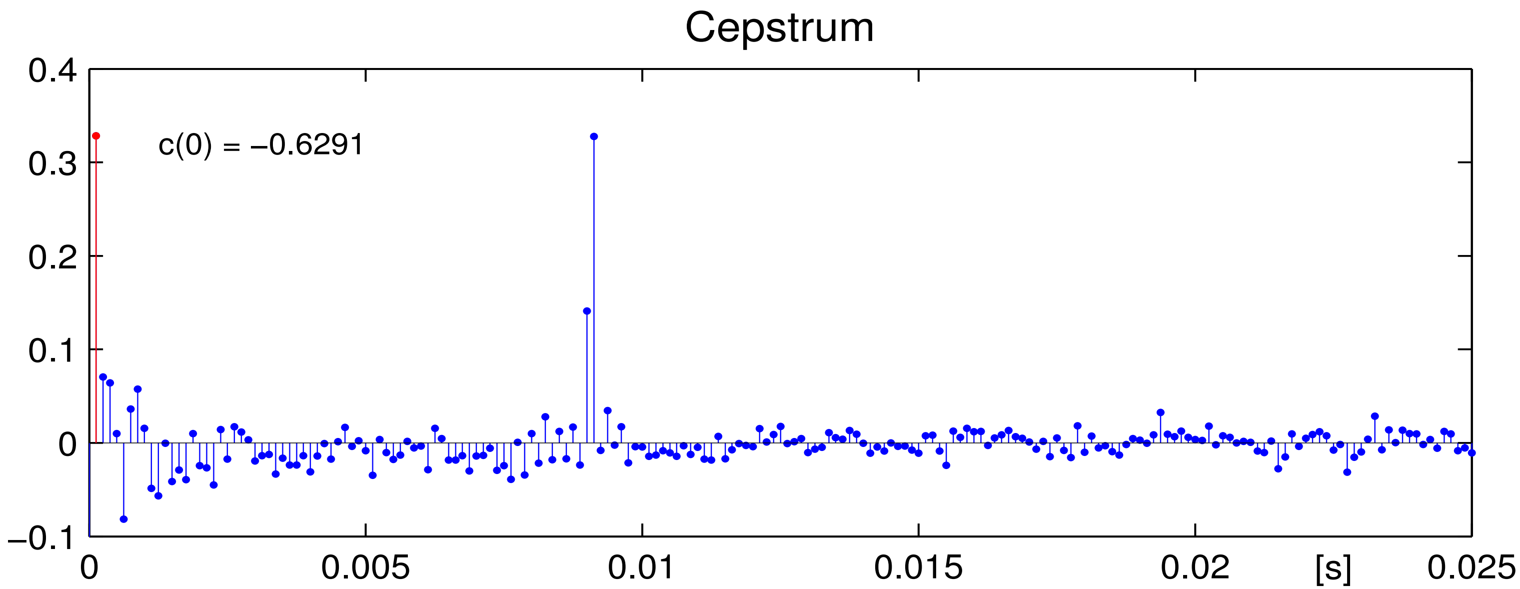
Verwendung des DFT-Cepstrums

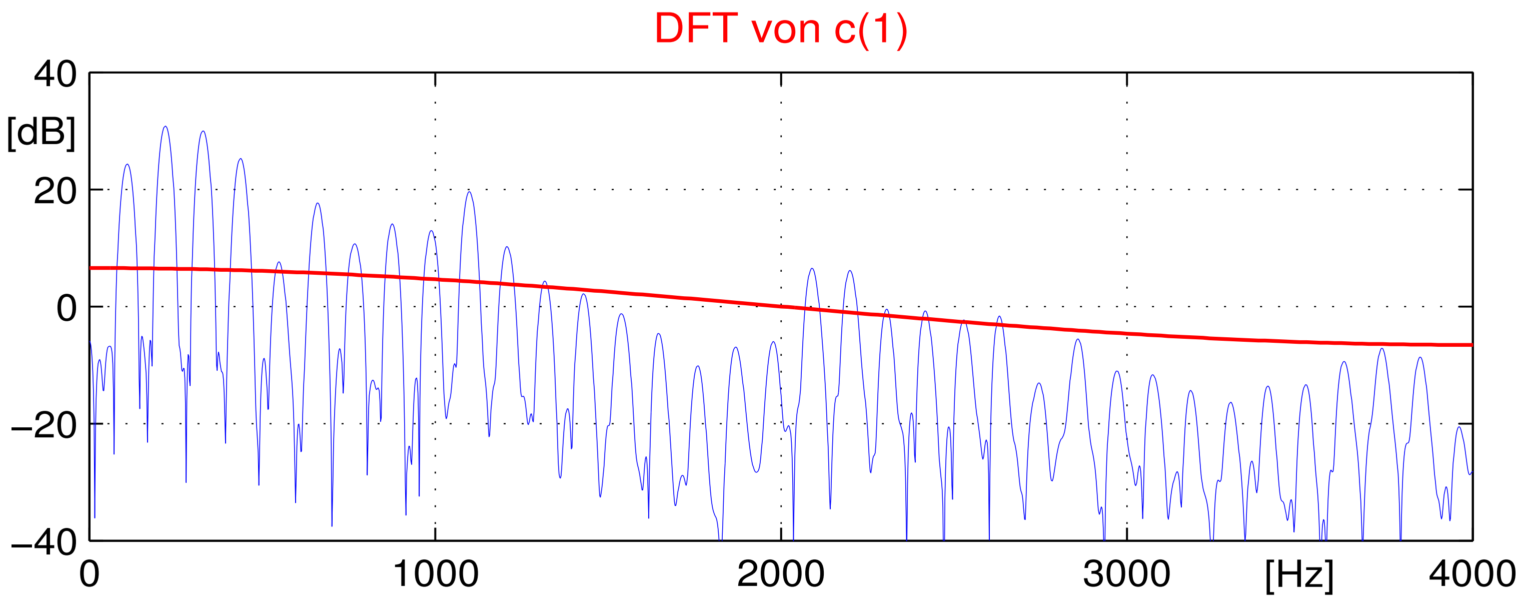
- zum Glätten von Spektren bzw. Spektrogrammen
- zum Schätzen der Periode von Sprachsignalen
- zur Kompensation des Übertragungskanals
- zum Ermitteln des Unterschiedes zwischen Spektren
- als Merkmal in der Sprach- und Sprechererkennung
- etc.

Thema der nächsten Lektion:

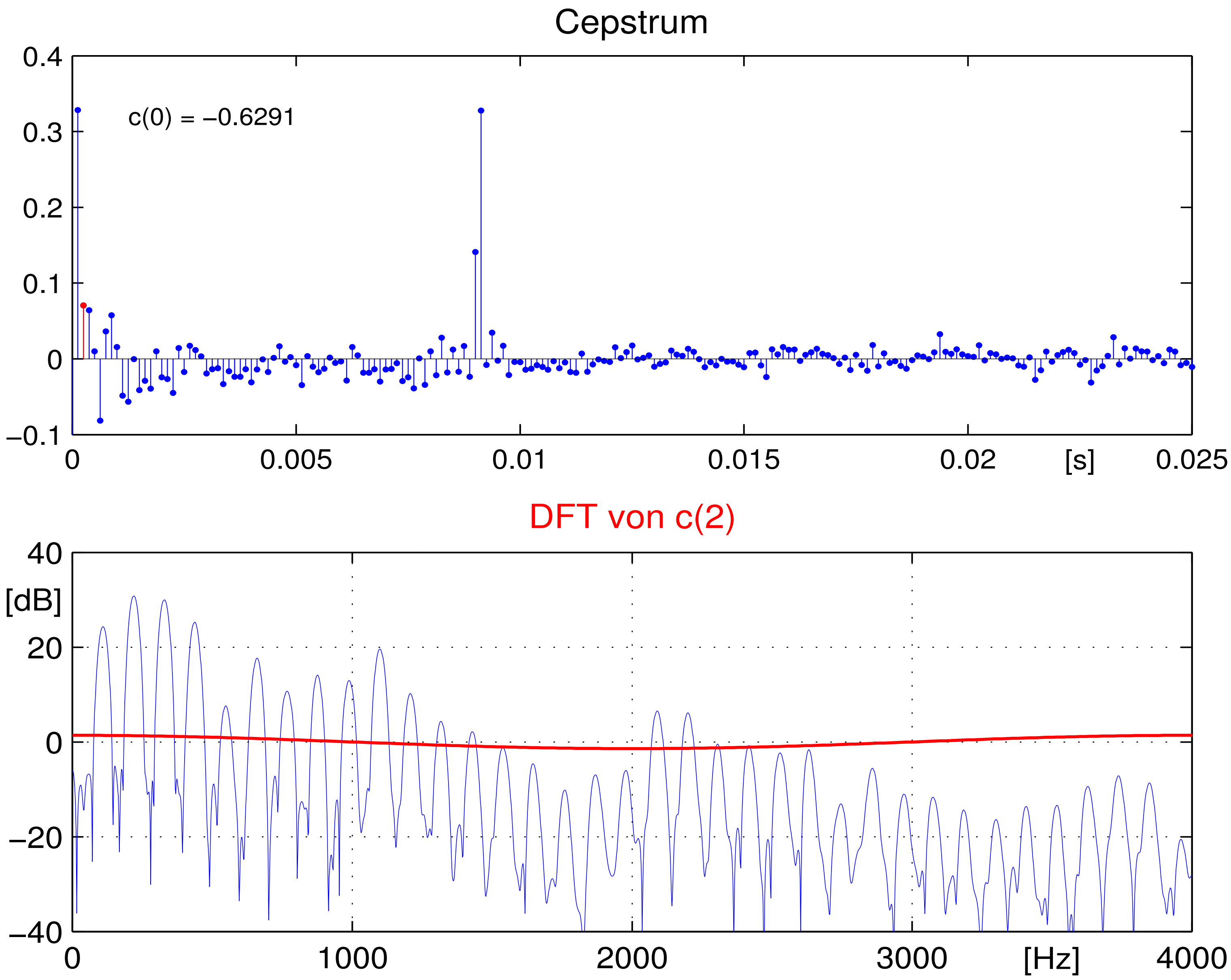
- Vektorquantisierung
- Einführung in die Sprachsynthese

Zur Übersicht der Vorlesung *Sprachverarbeitung I* >>>

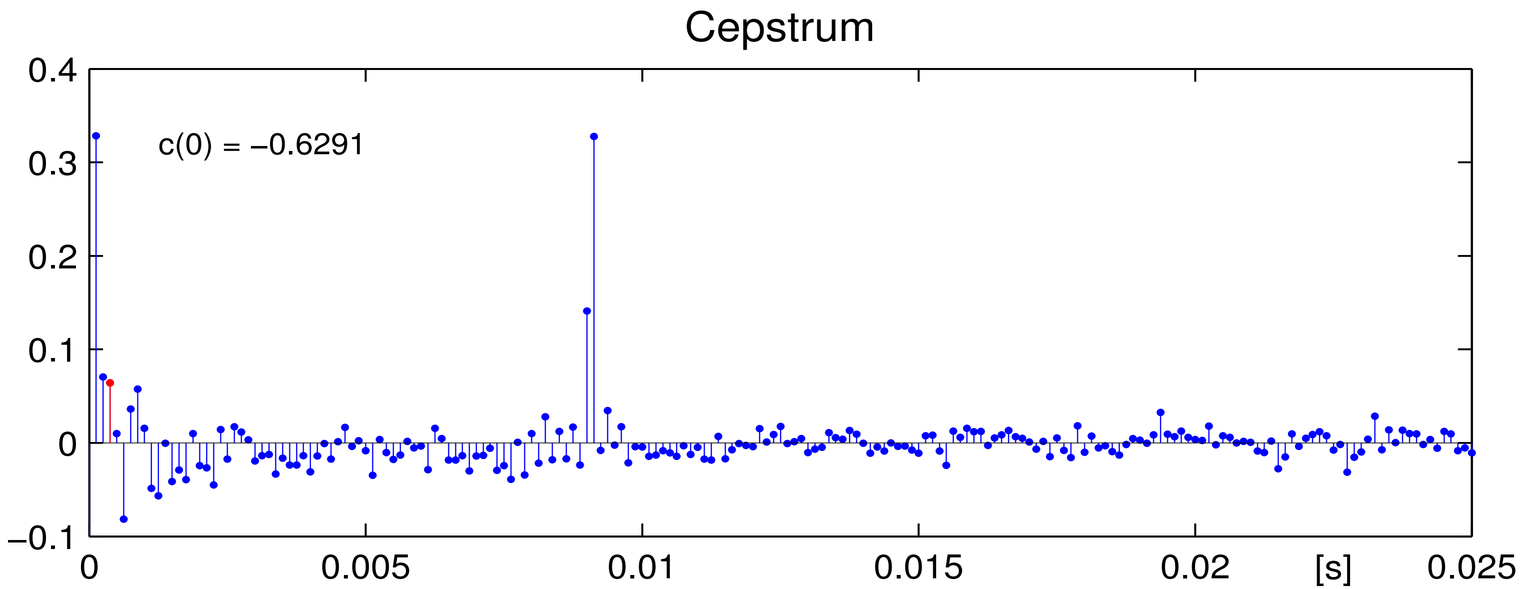

Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025

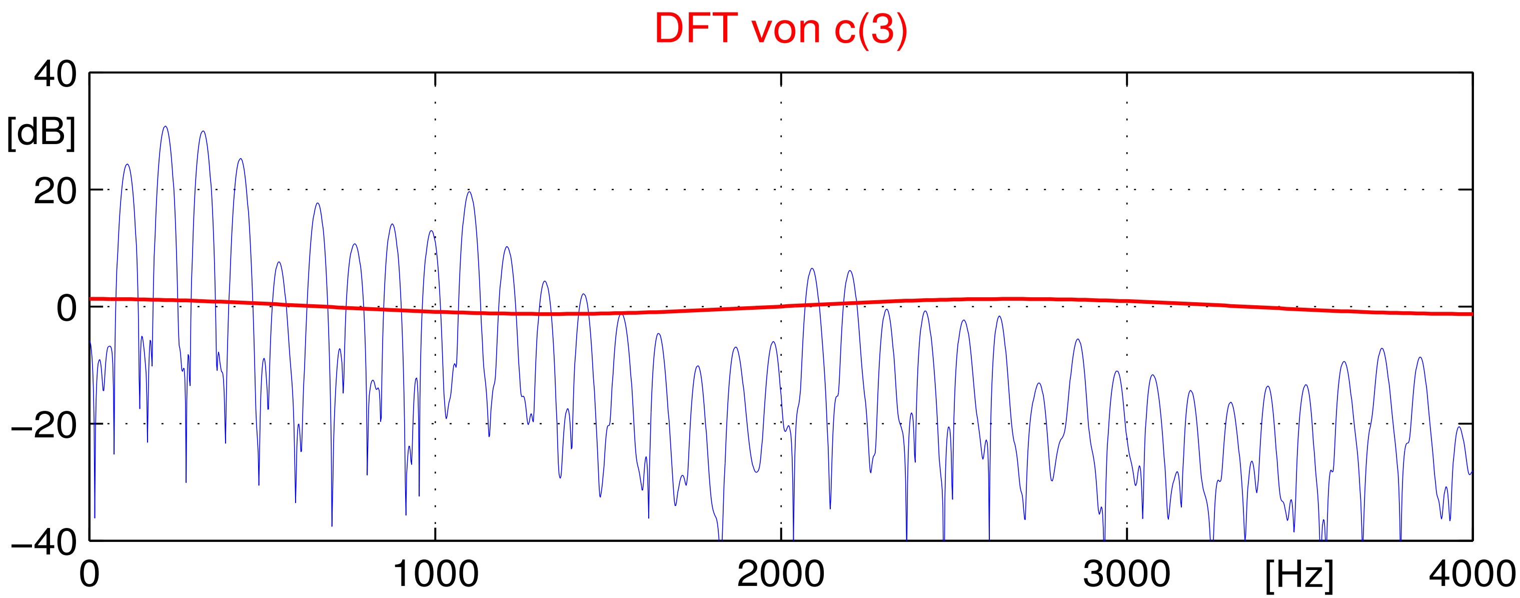

DFT von c(1)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

<<<

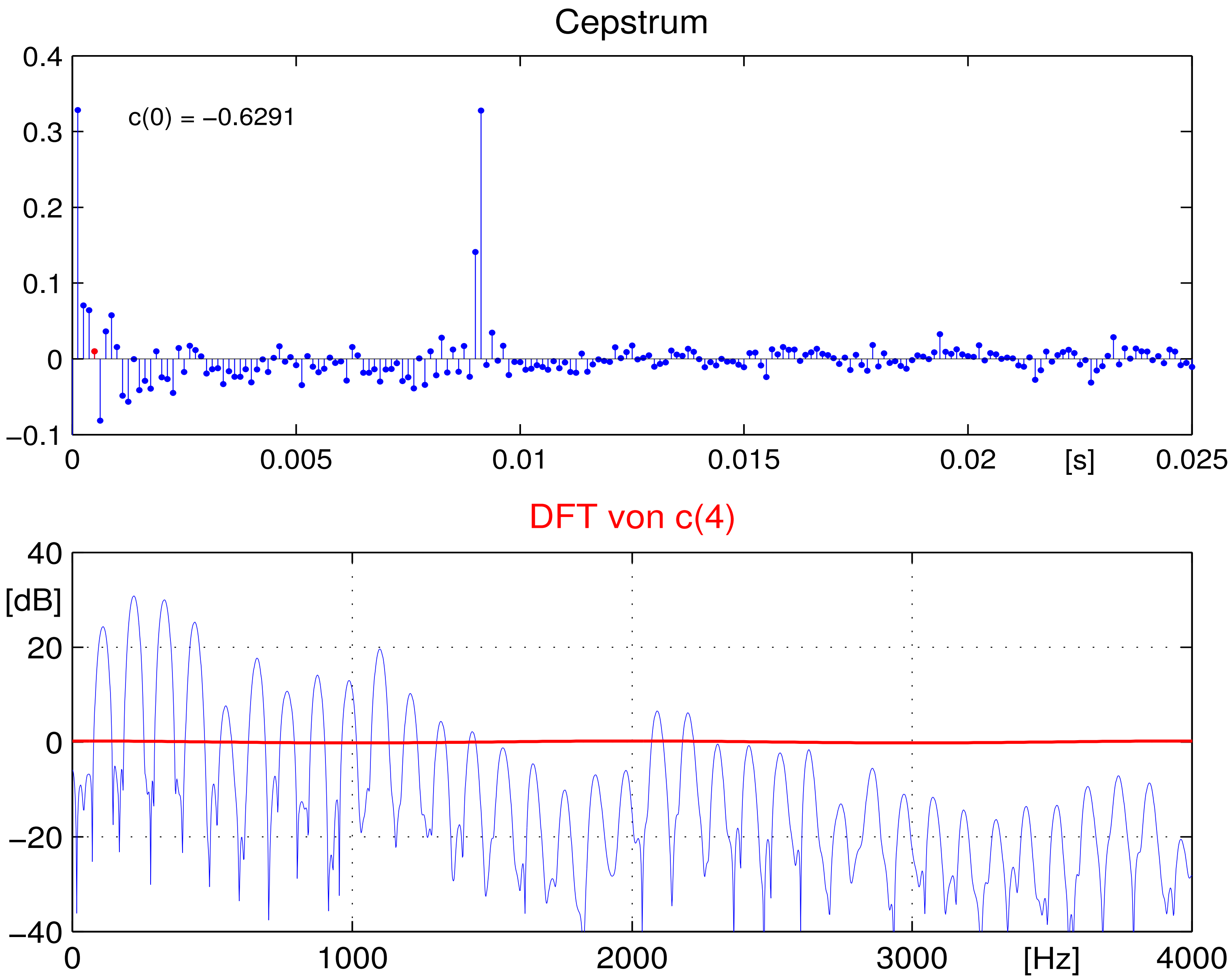

Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025
DFT von c(2)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

<<<

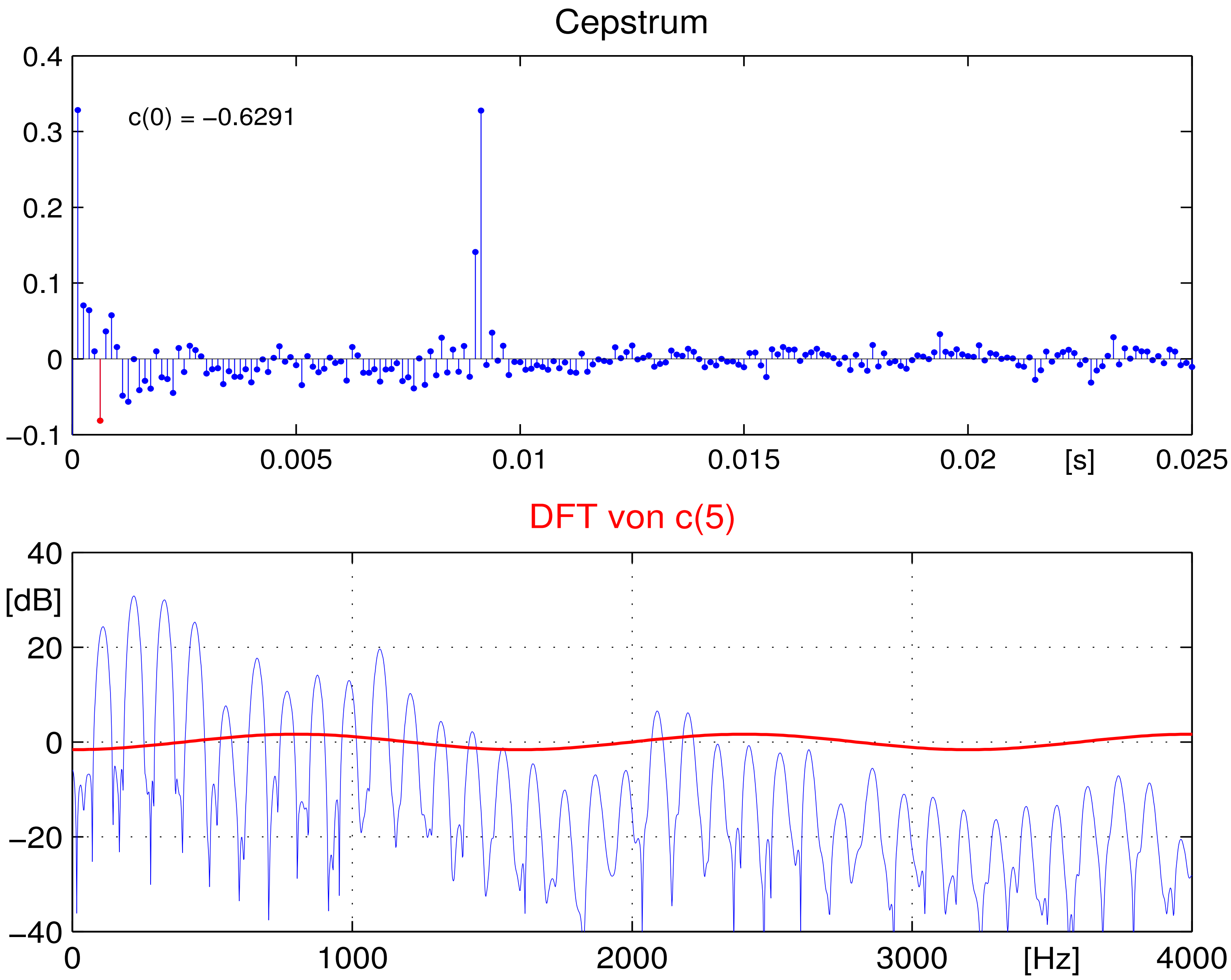

Cepstrum
0.4
0.3
0.2
0.1
0
–0.1
c(0) = –0.6291
0
0.005
0.01
0.015
0.02
[s]
0.025


DFT von c(3)
40
[dB]
20
0
–20
–40
0
1000
2000
3000
[Hz]
4000

<<<

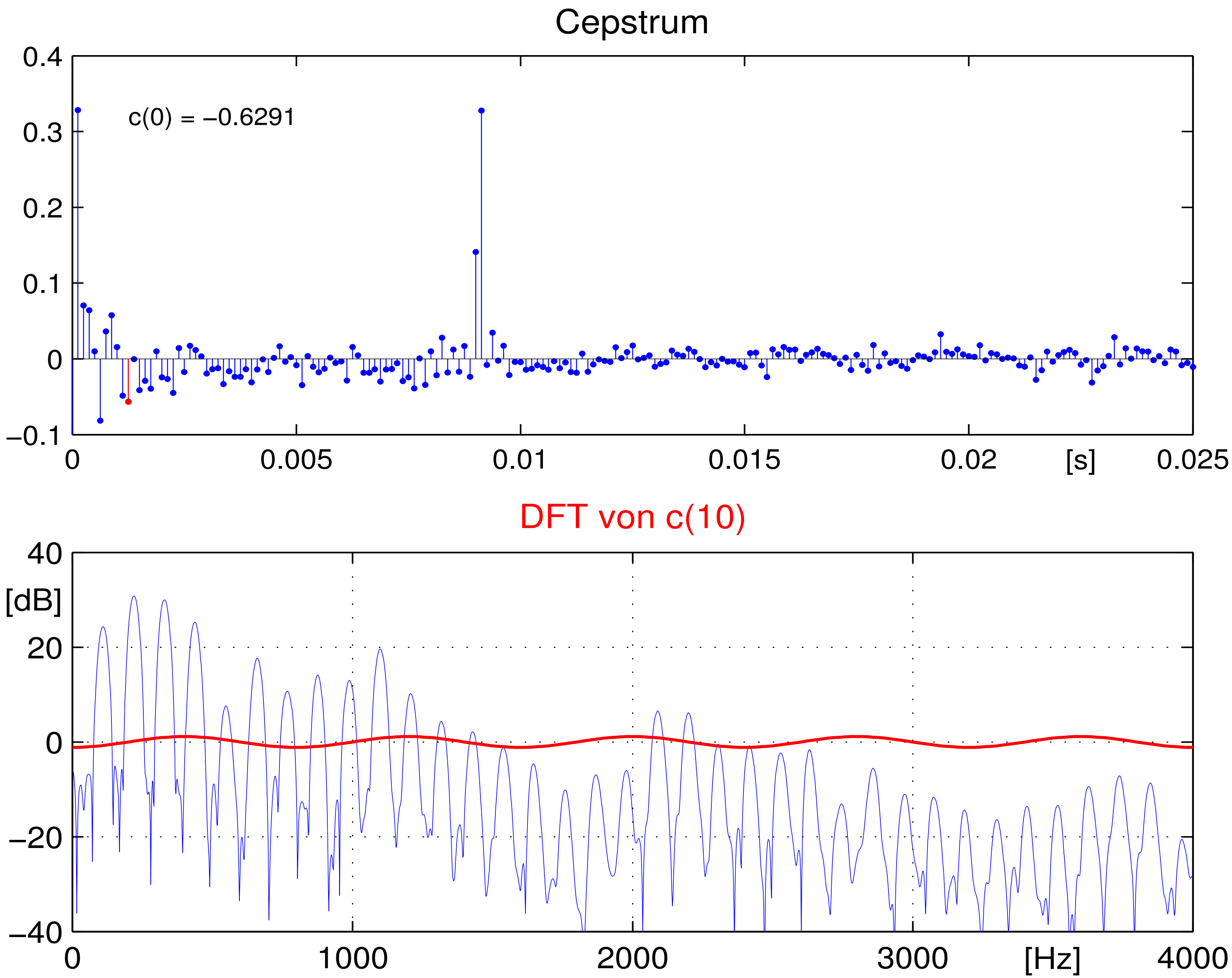

Cepstrum
0.4
0.3
0.2
0.1
0
–0.1
c(0) = –0.6291
0
0.005
0.01
0.015
0.02
[s]
0.025
DFT von c(4)
40
[dB]
20
0
–20
–40
0
1000
2000
3000
[Hz]
4000

<<<

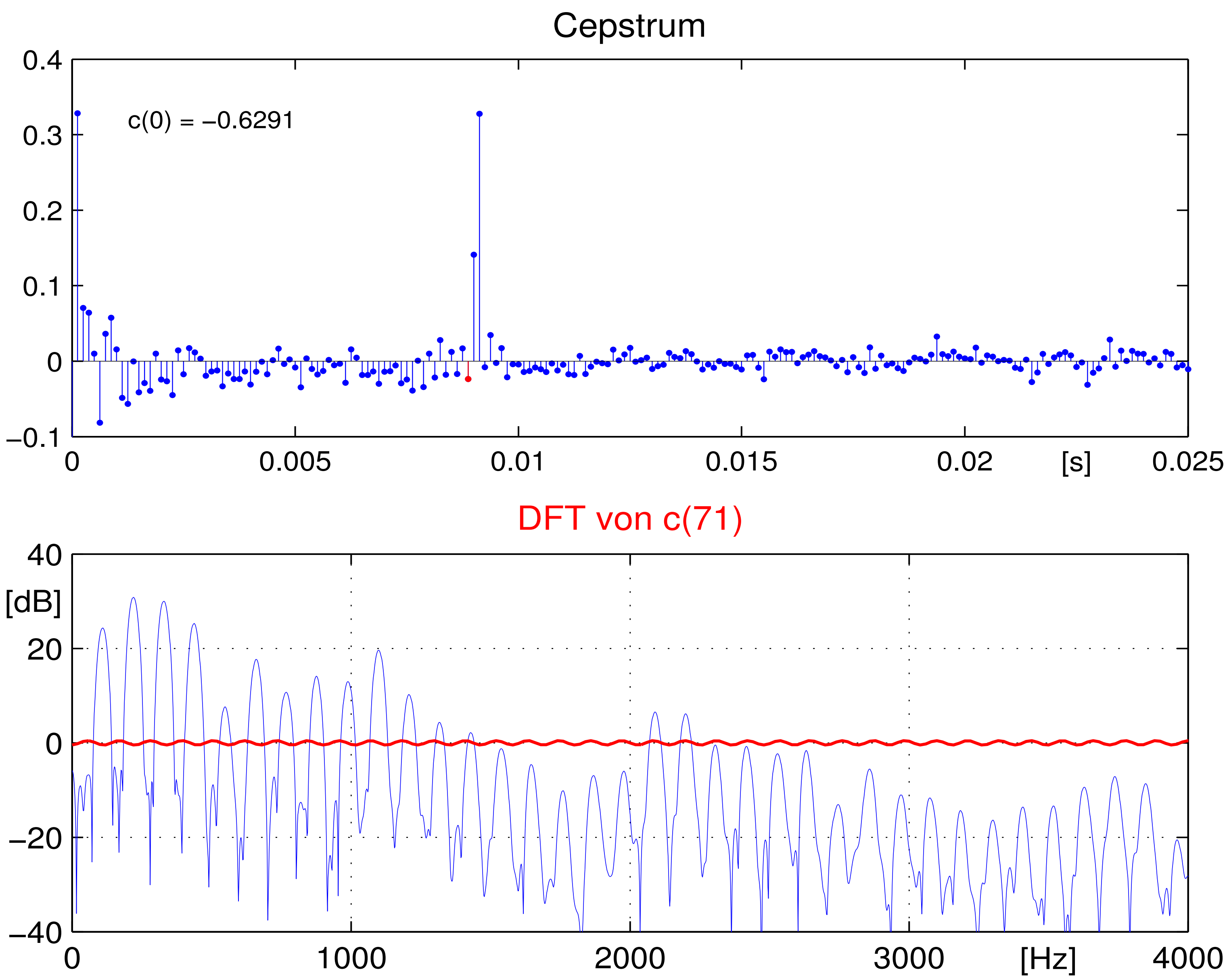


Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025
DFT von c(5)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

<<<

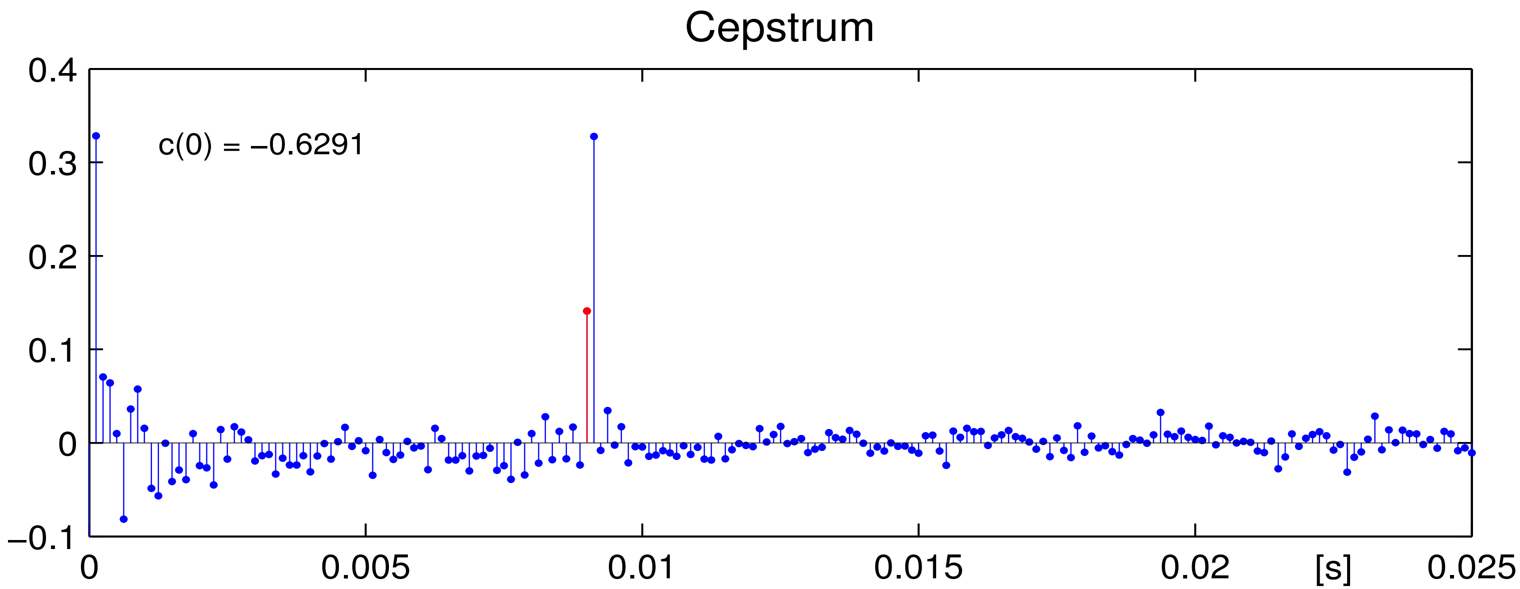

Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025
DFT von c(10)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

<<<

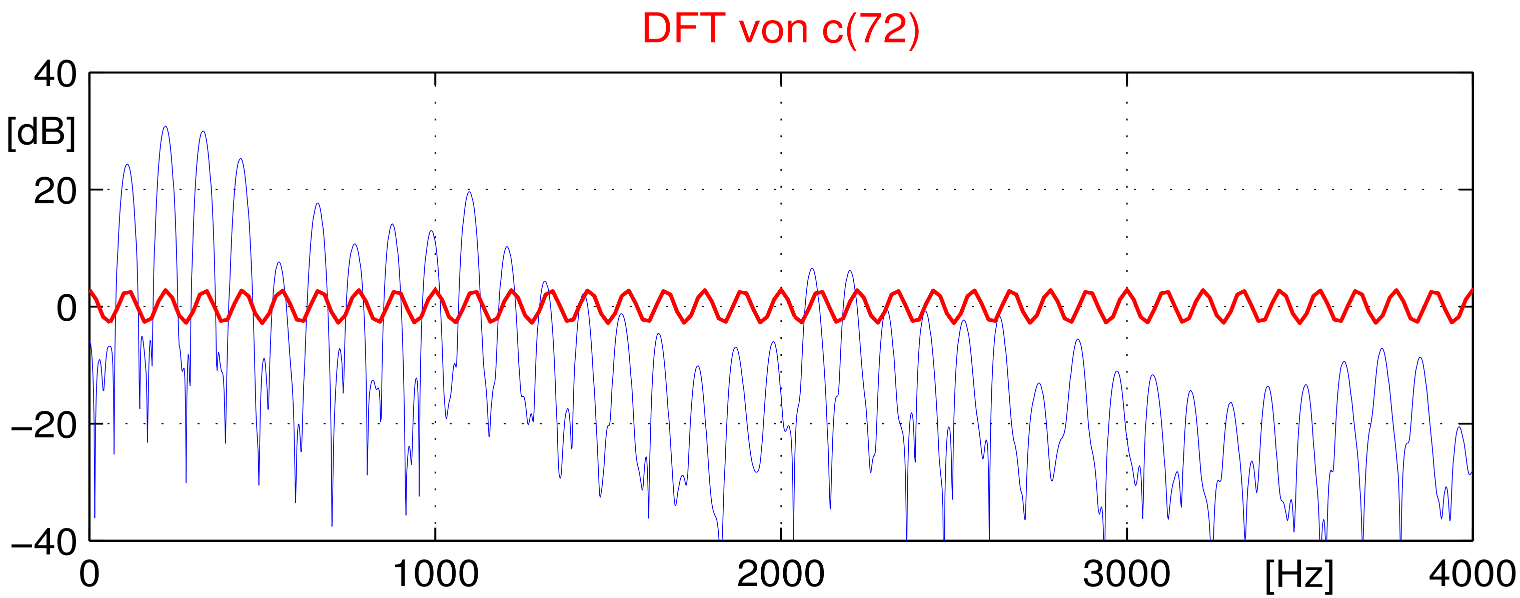


Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025
DFT von c(71)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

<<<

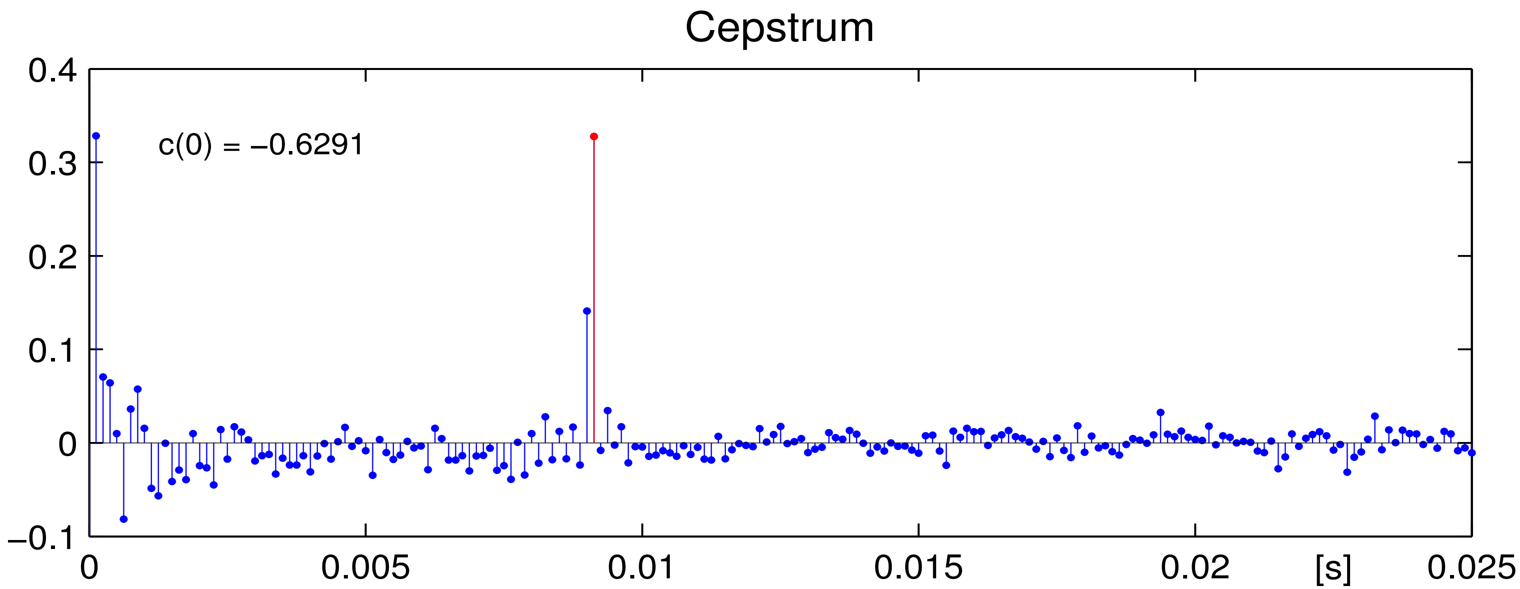


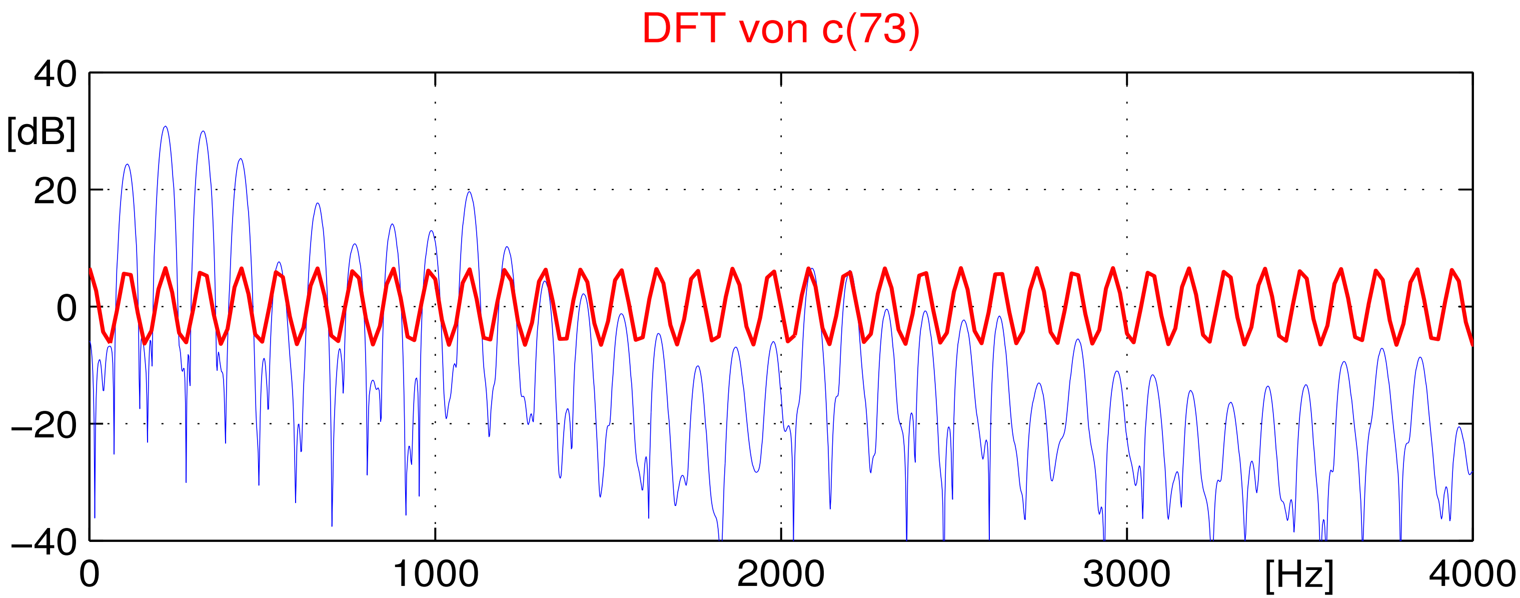
Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025



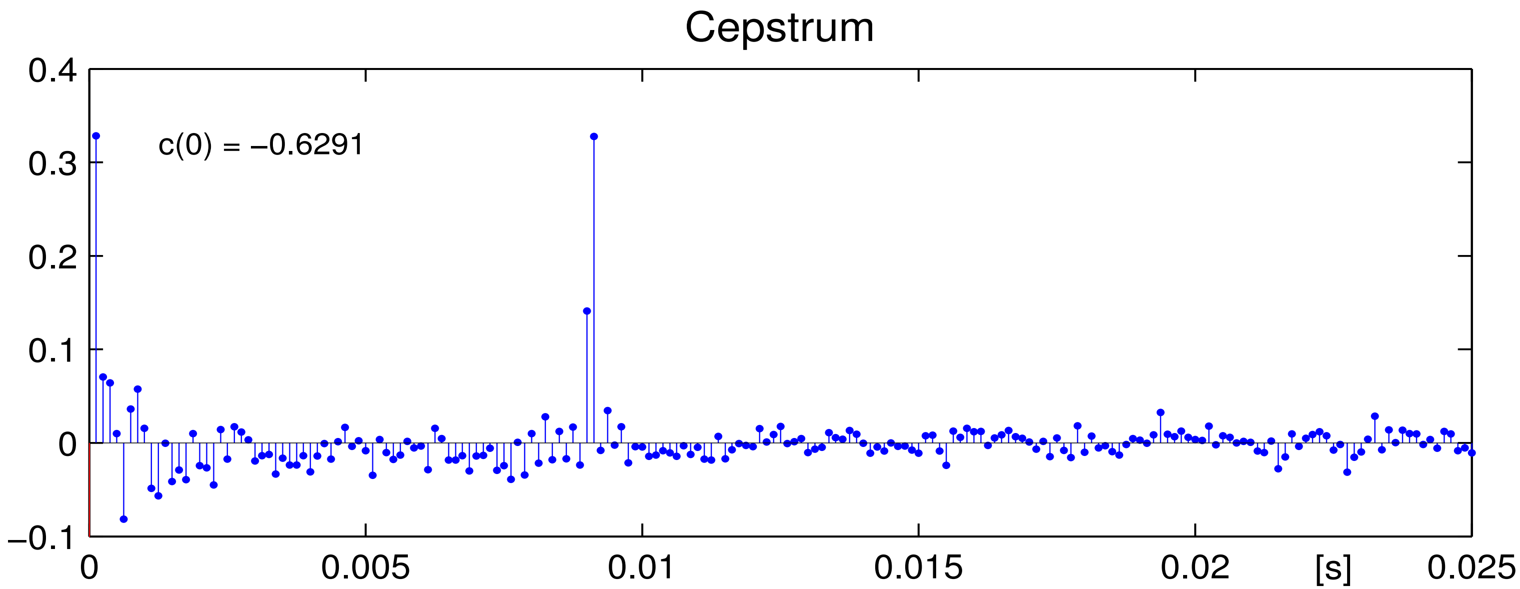
DFT von c(72)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

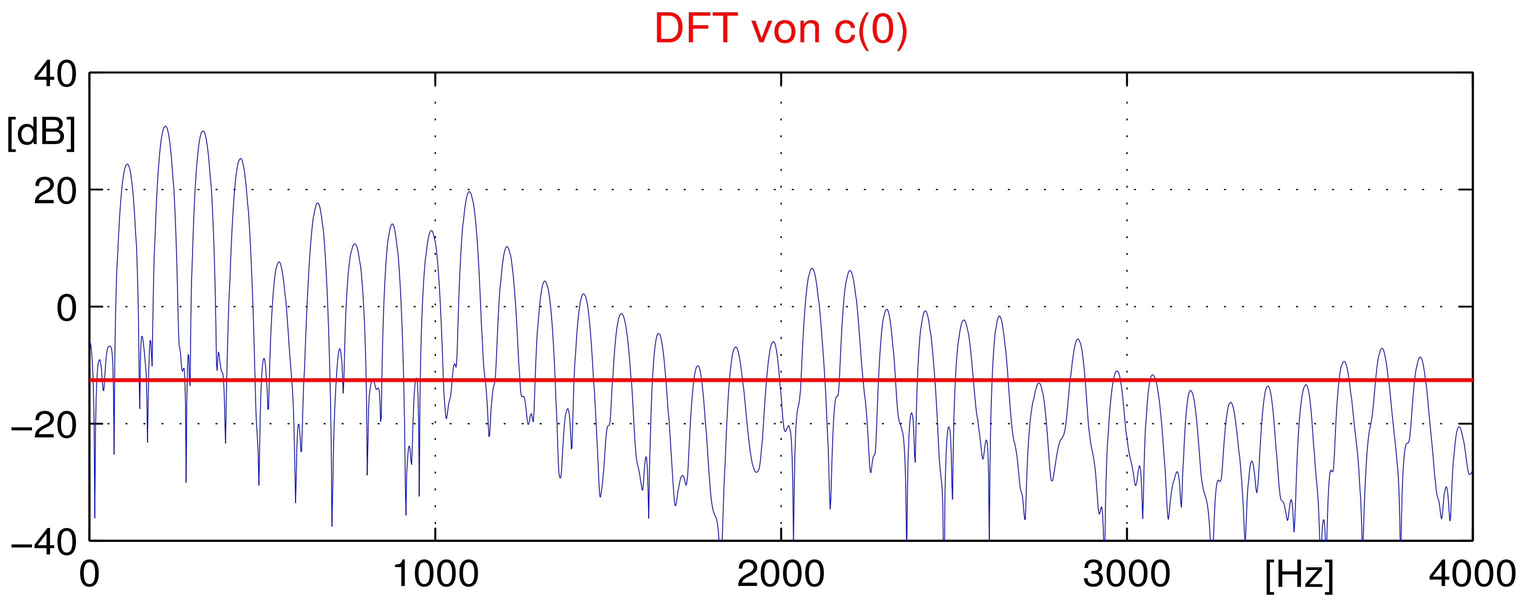
<<<


Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025

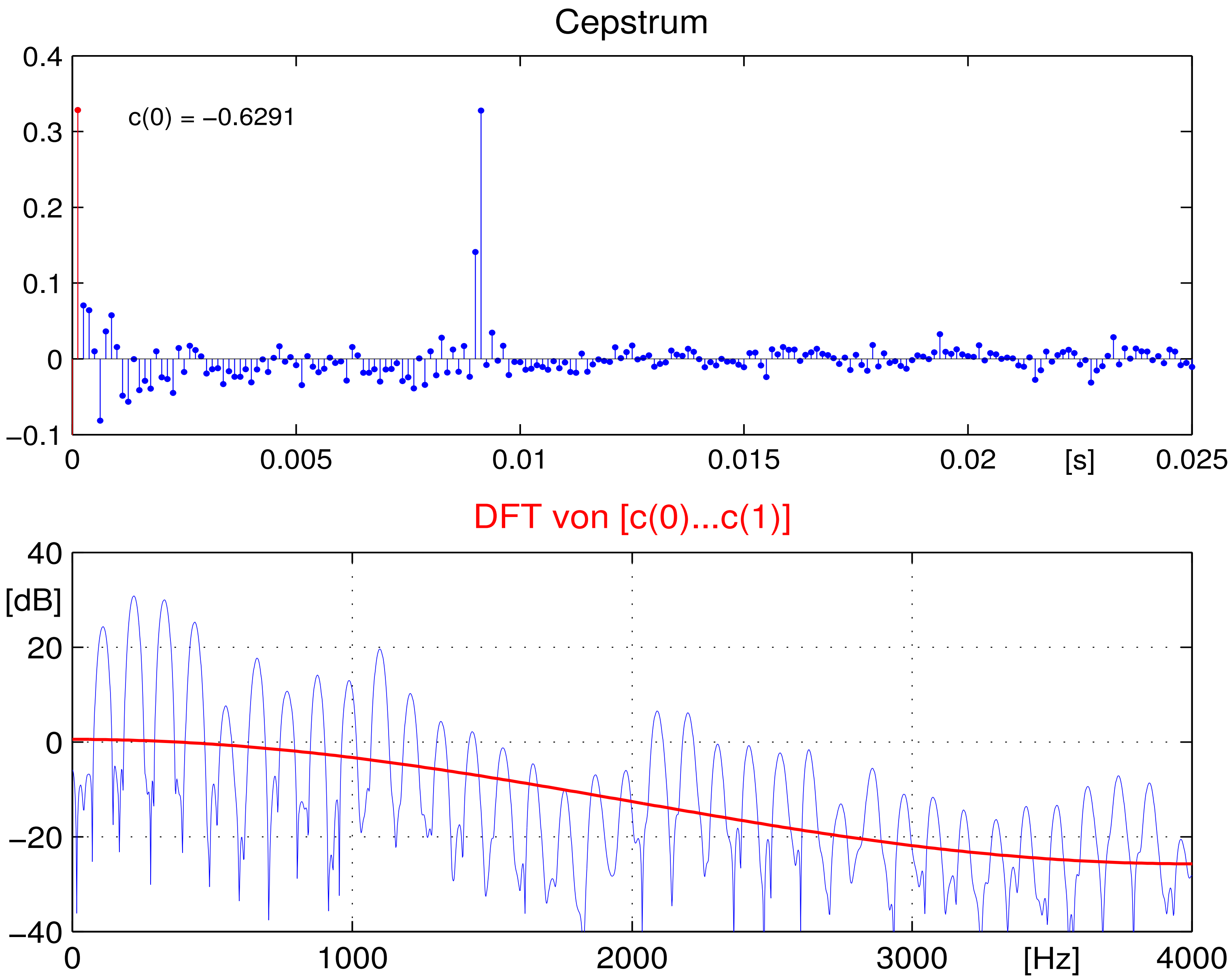

DFT von c(73)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

<<<


Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025

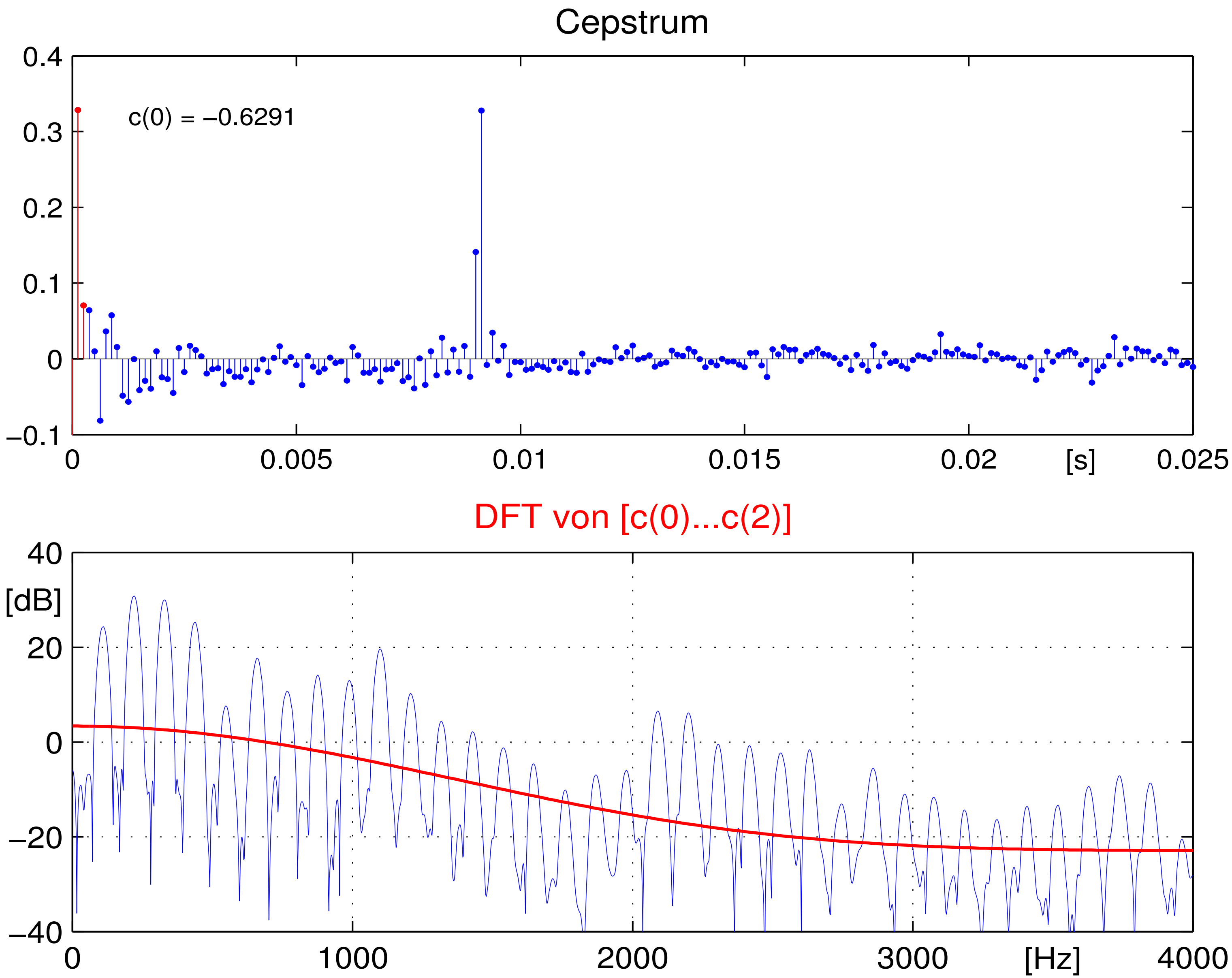

DFT von c(0)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

<<<

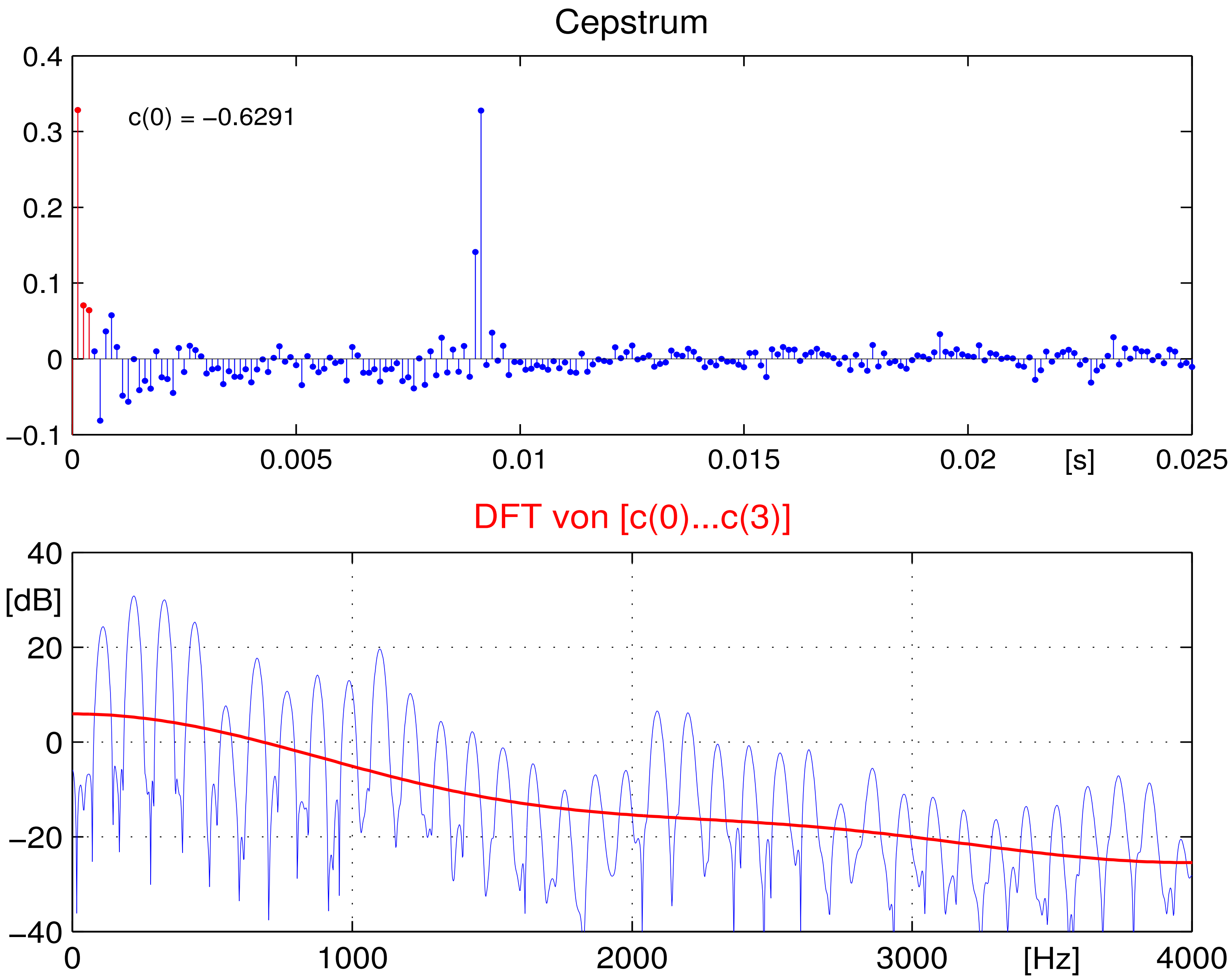


Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025
DFT von [c(0)...c(1)]
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

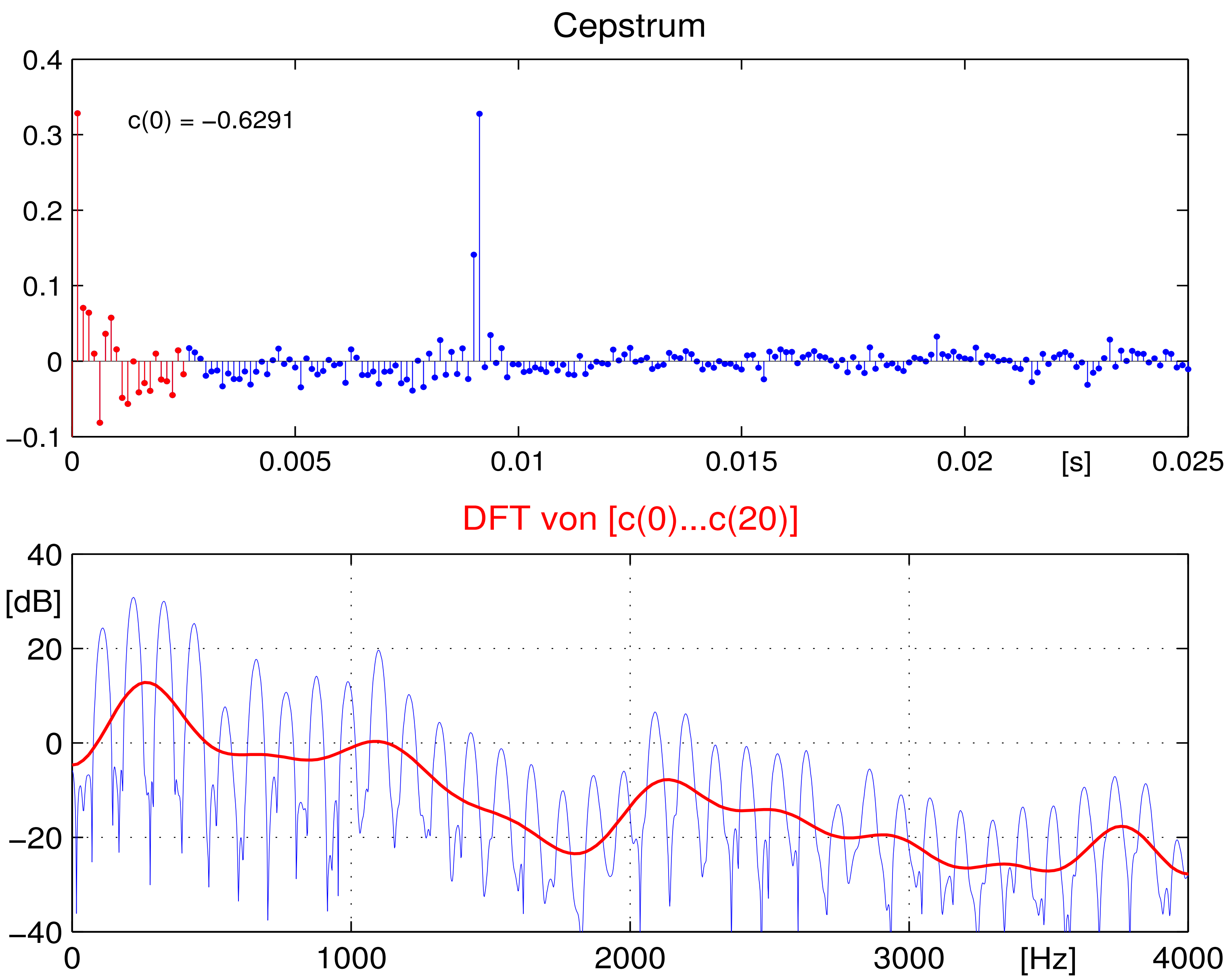
<<<


Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025
DFT von [c(0)...c(2)]
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

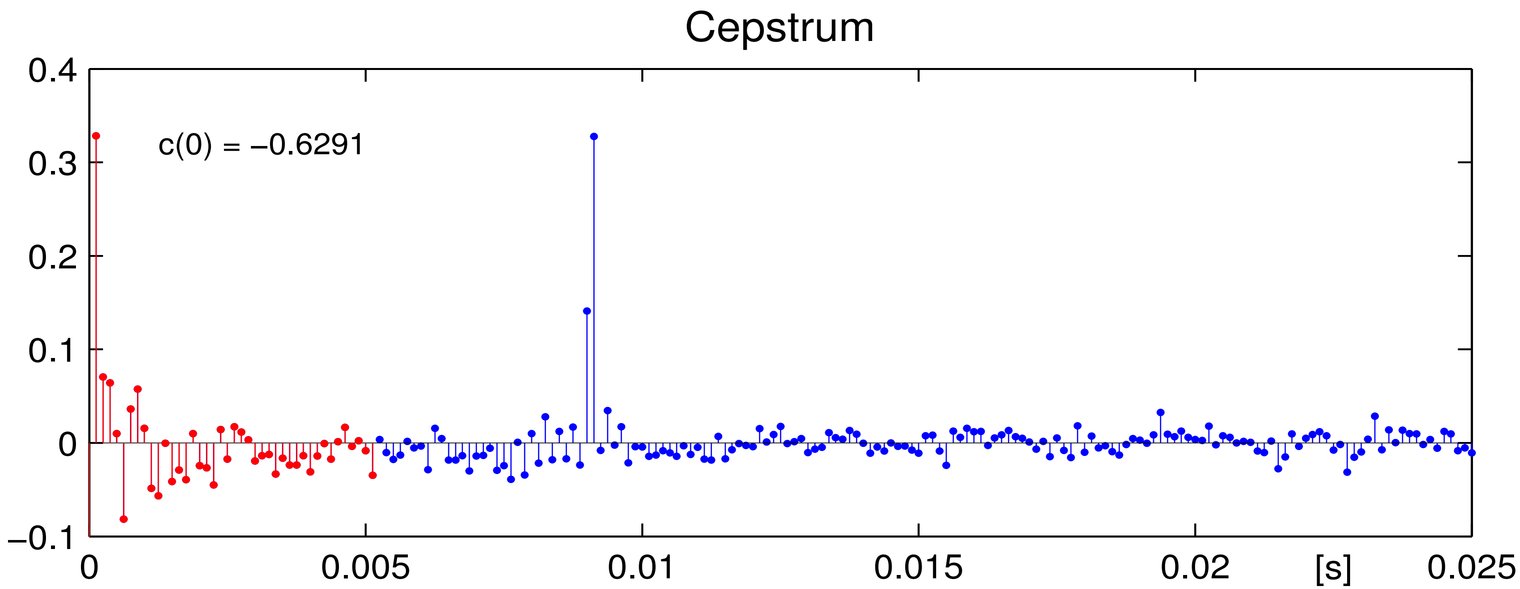
<<<

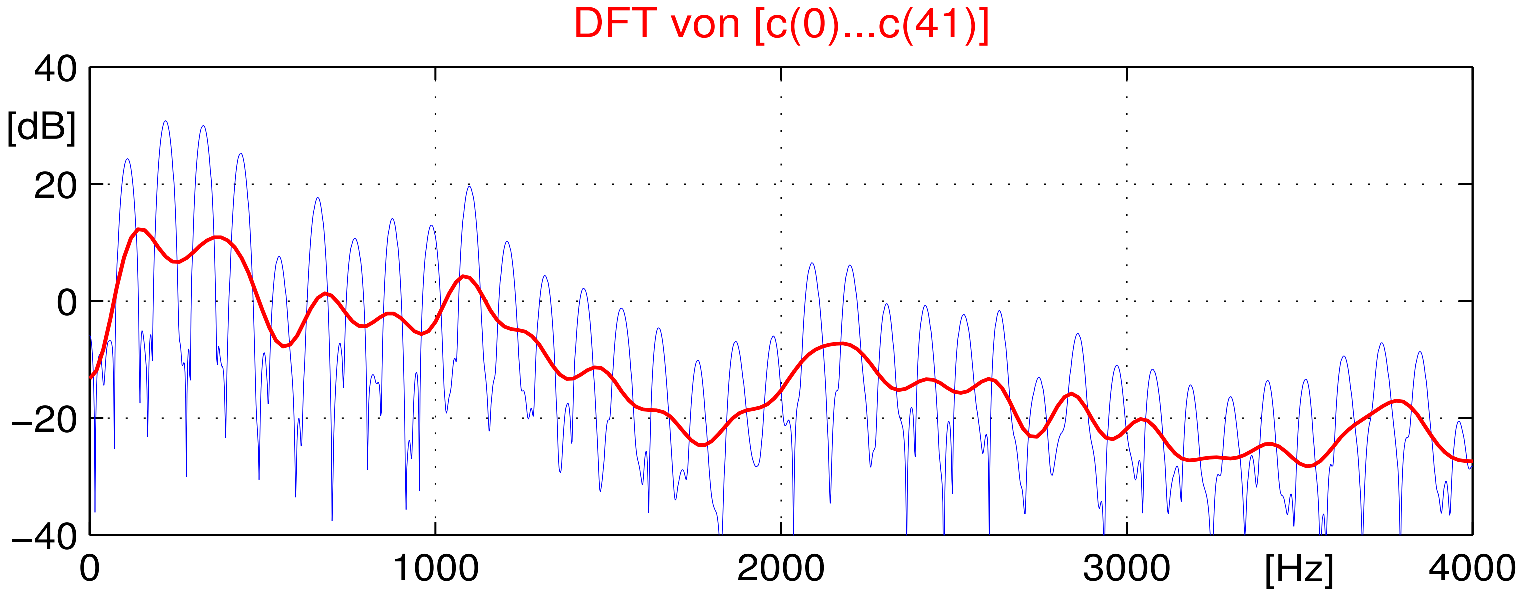

Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025
DFT von [c(0)...c(3)]
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

<<<

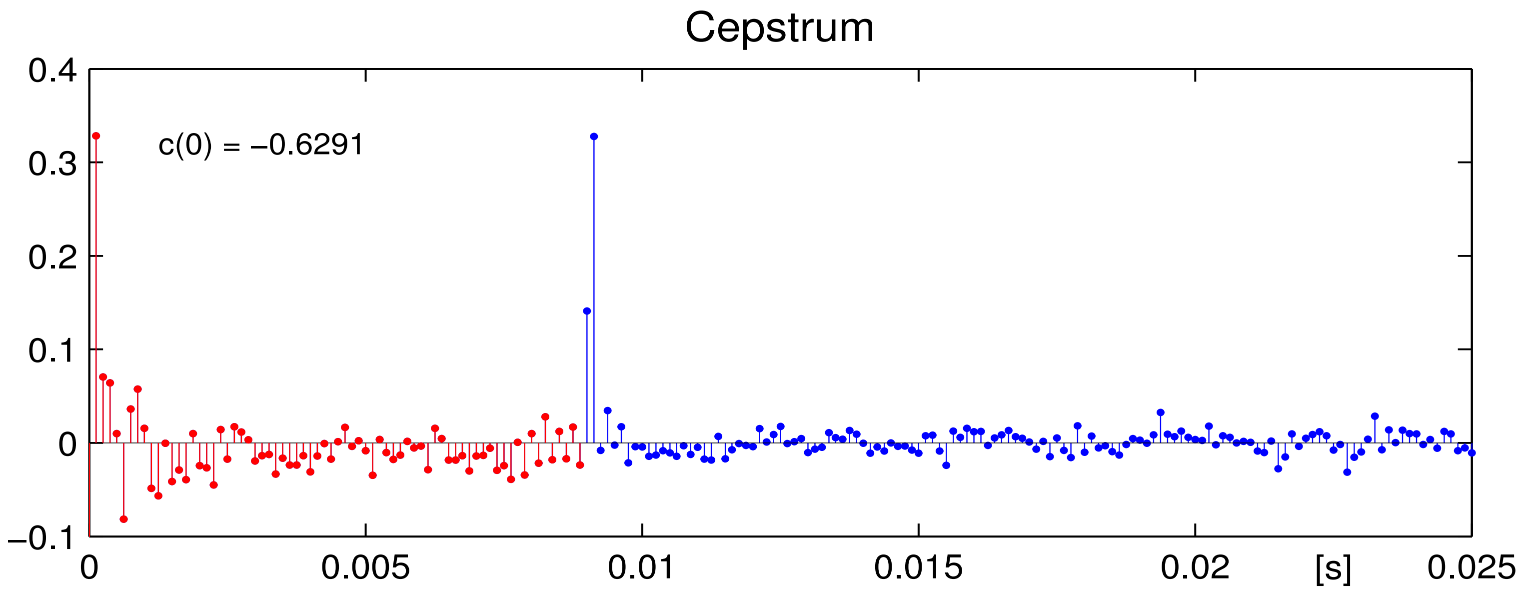

Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025
DFT von [c(0)...c(20)]
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

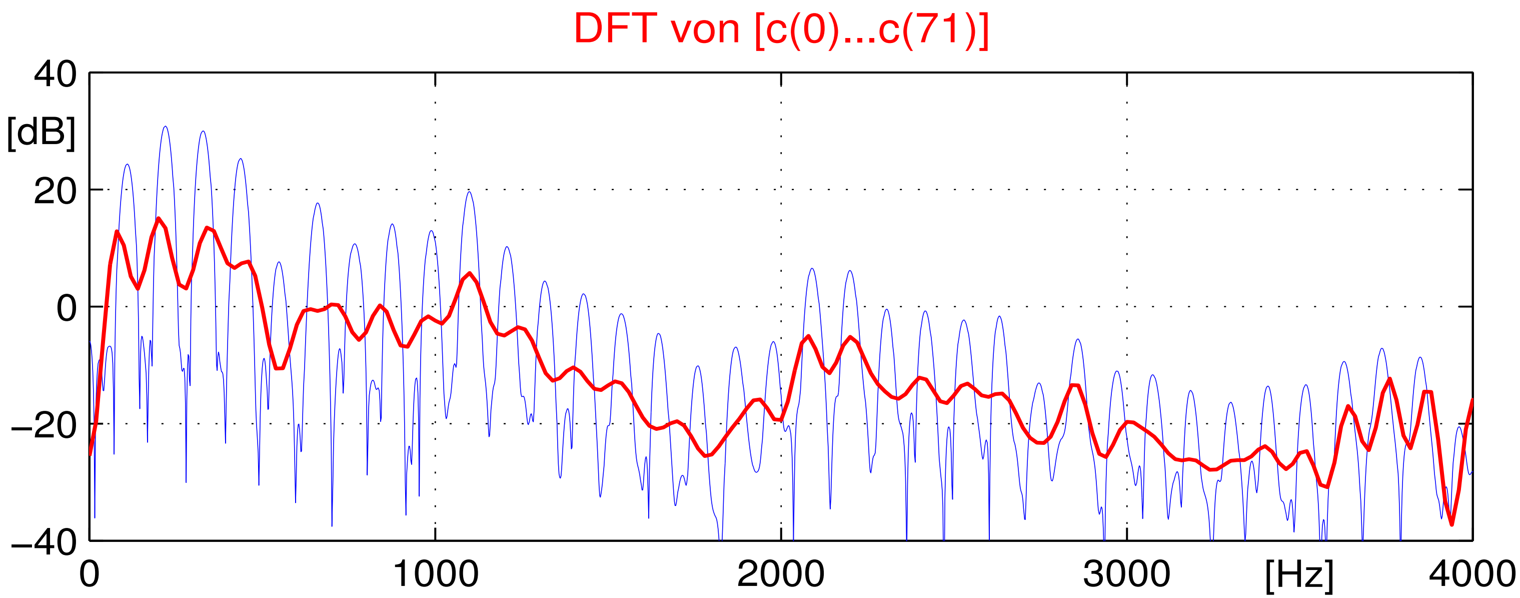
<<<


Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025

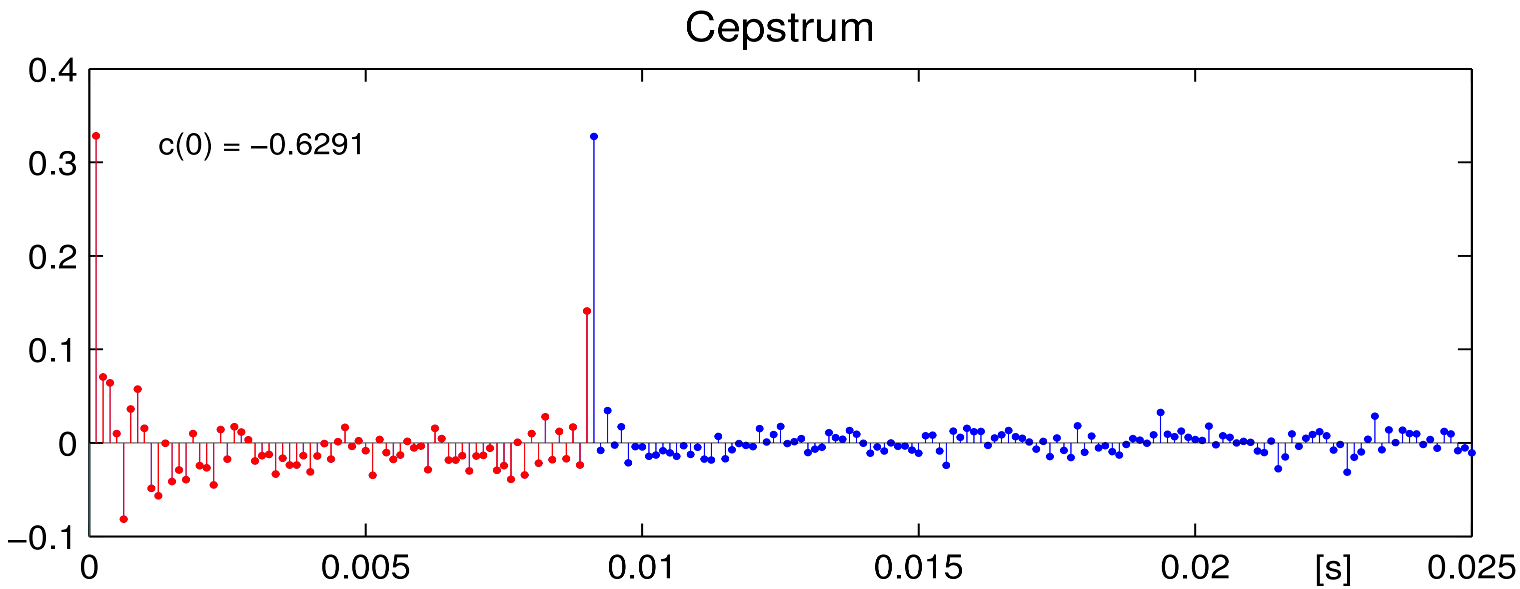

DFT von [c(0)...c(41)]
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

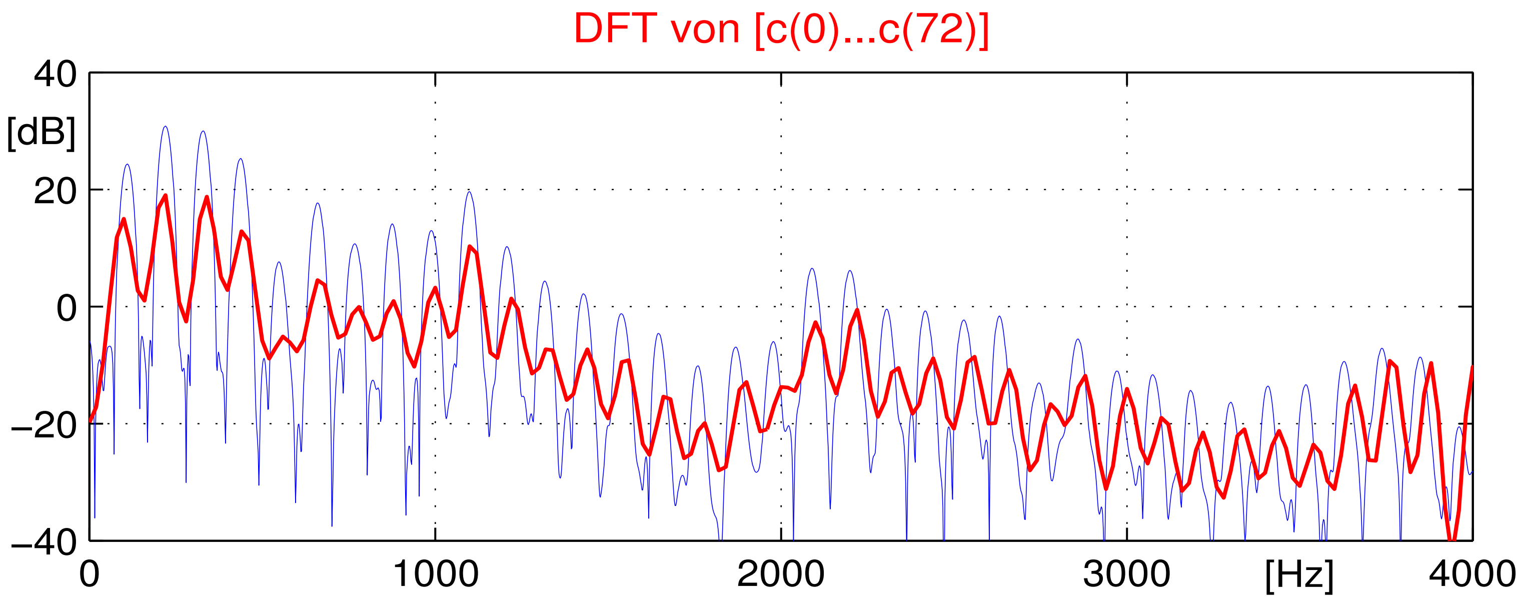
<<<


Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025

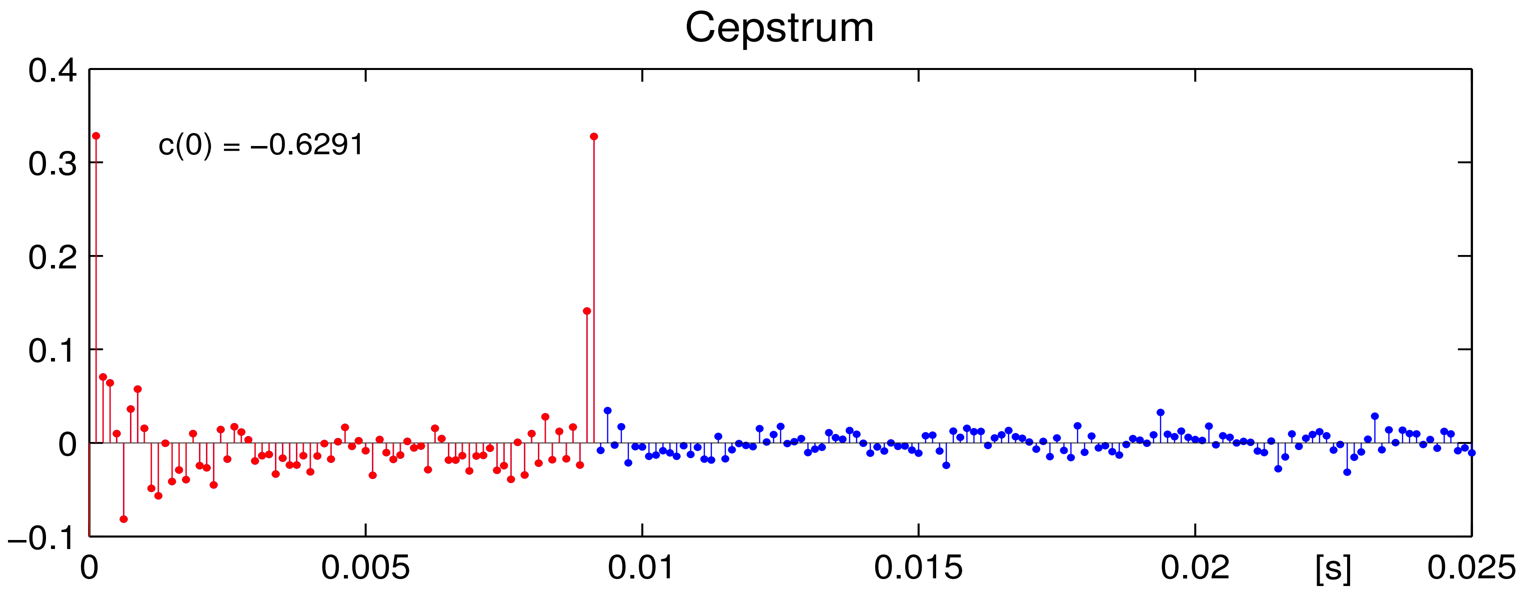

DFT von [c(0)...c(71)]
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

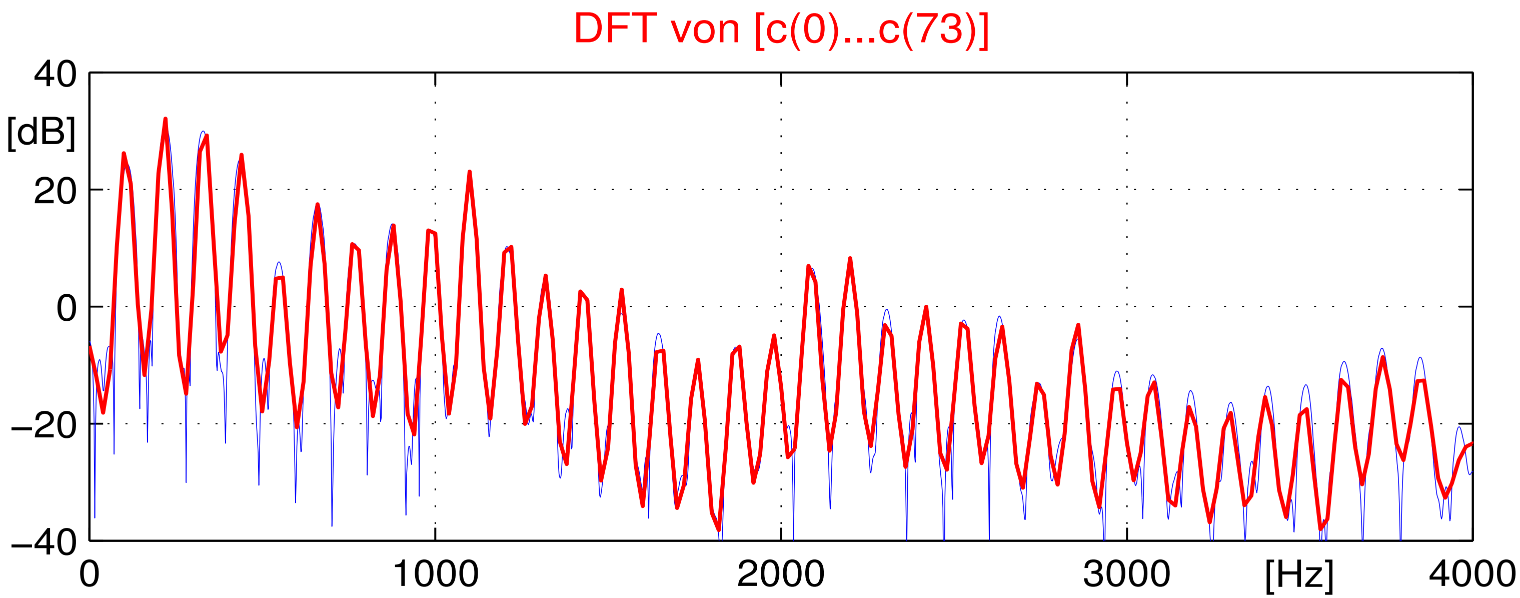
<<<


Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025


DFT von [c(0)...c(72)]
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

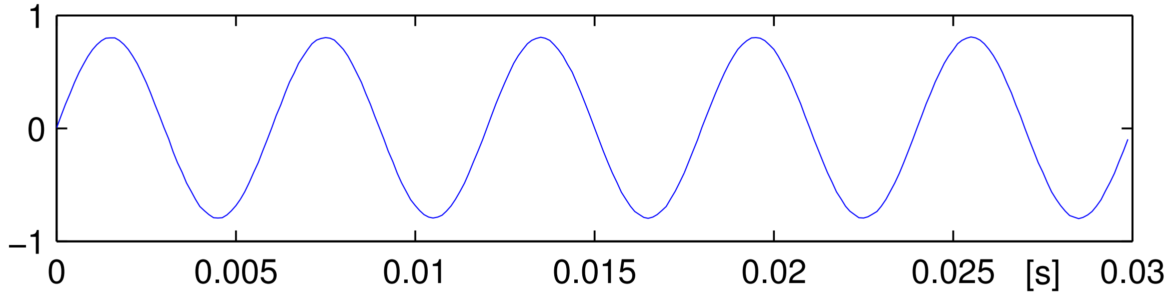
<<<


Cepstrum
c(0) = −0.6291
0.4
0.3
0.2
0.1
0
−0.1
0
0.005
0.01
0.015
0.02
[s]
0.025


DFT von [c(0)...c(73)]
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000

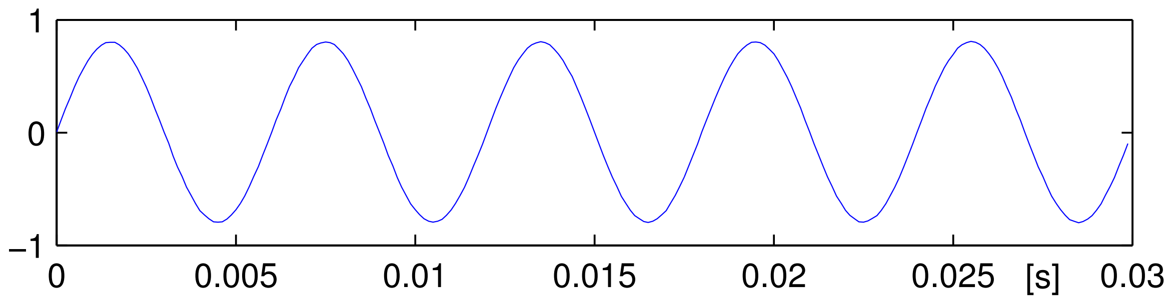
<<<

Gegeben: Sinussignal $x(n)$

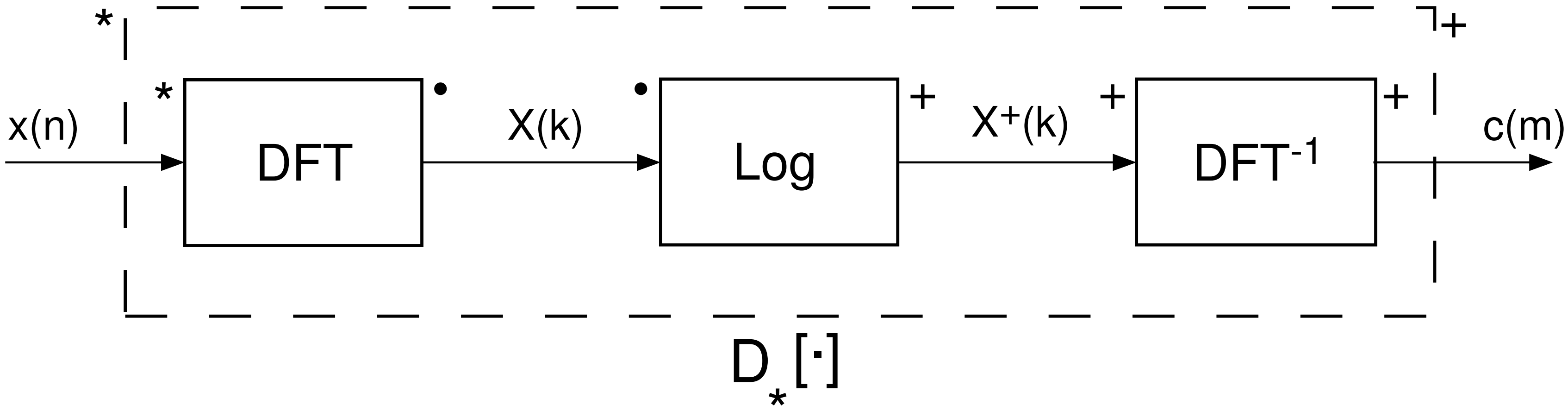


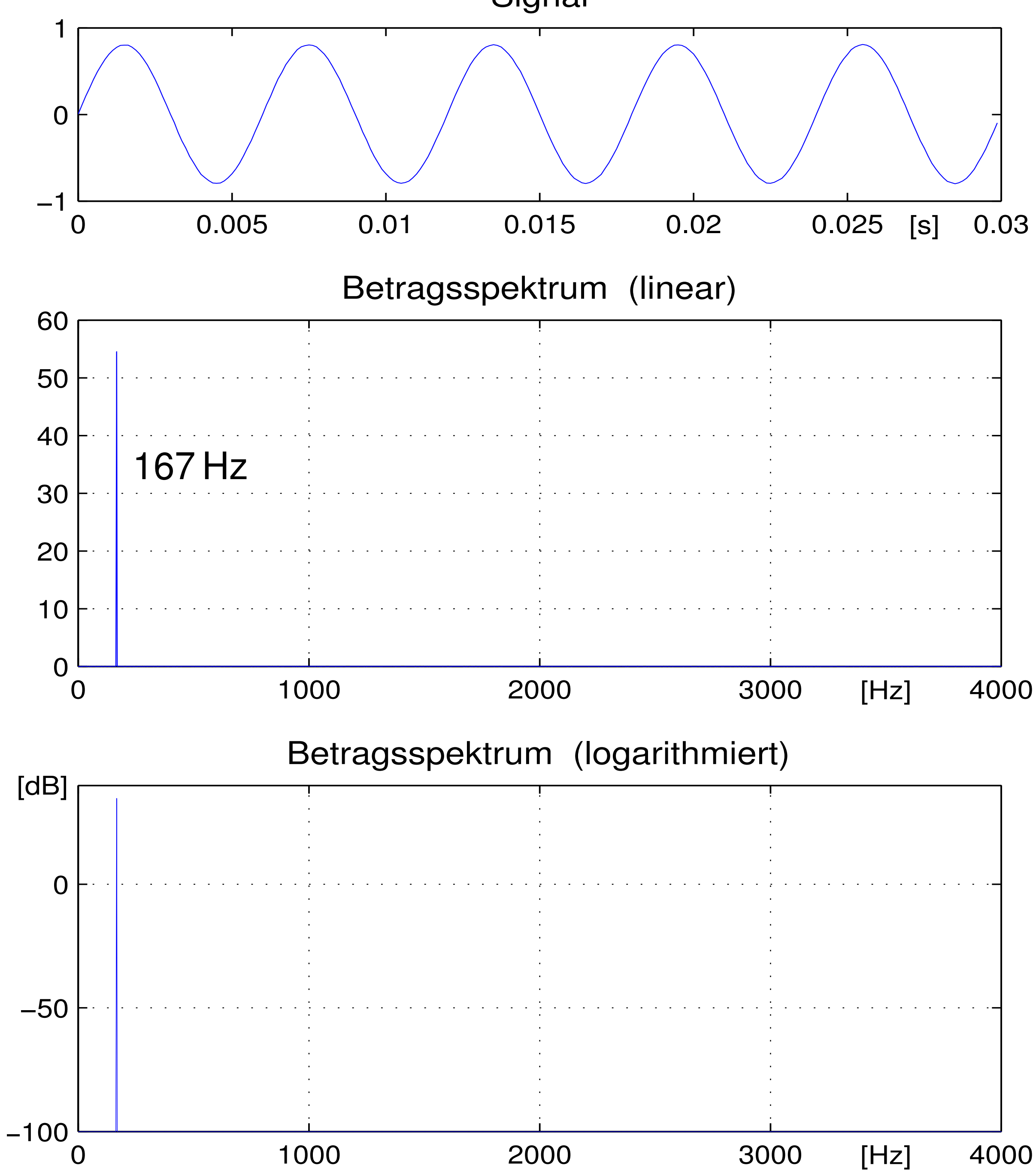
Frage: Wie sieht das Cepstrum $c(m)$ aus?

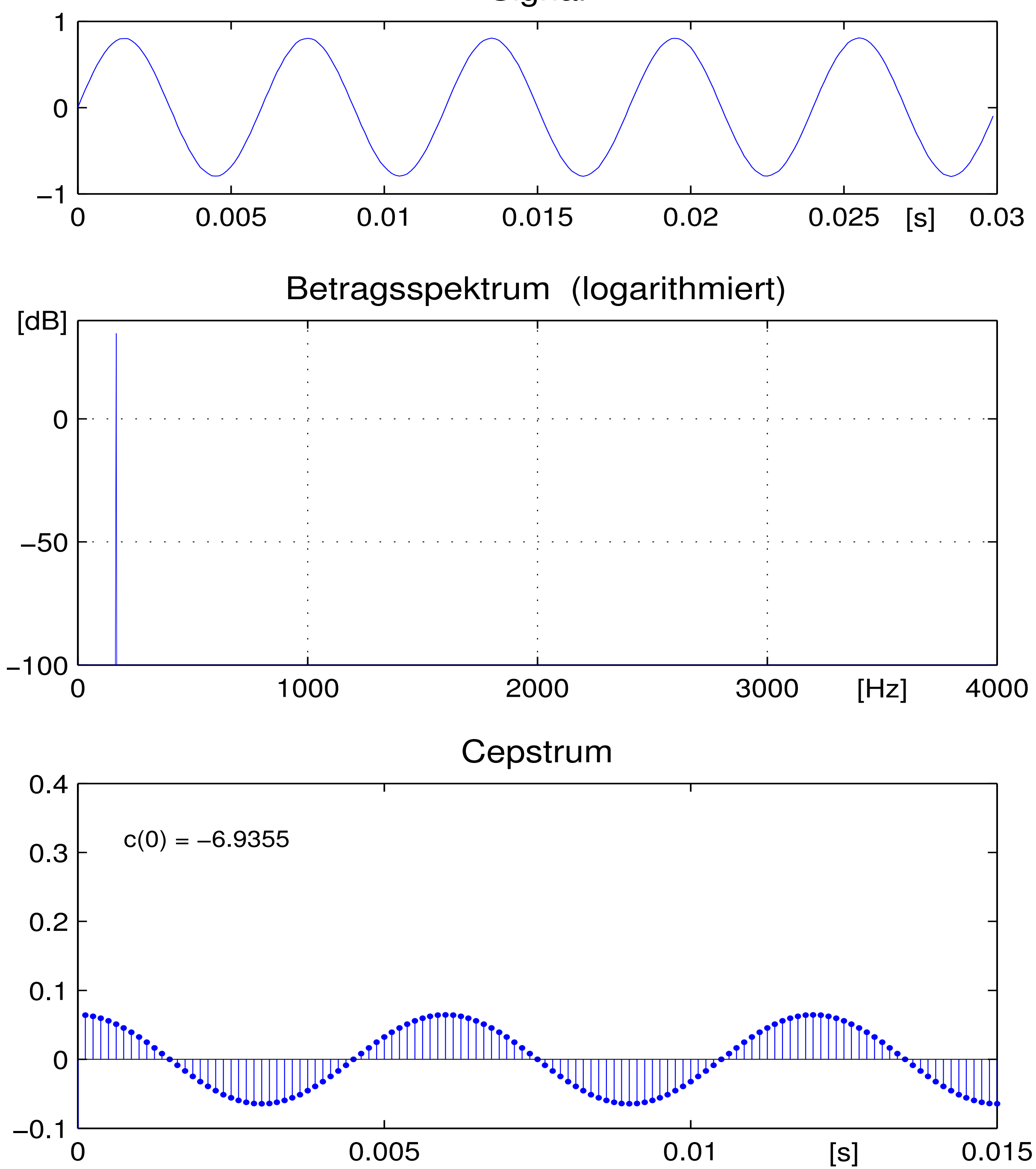
Gegeben: Sinussignal $x(n)$

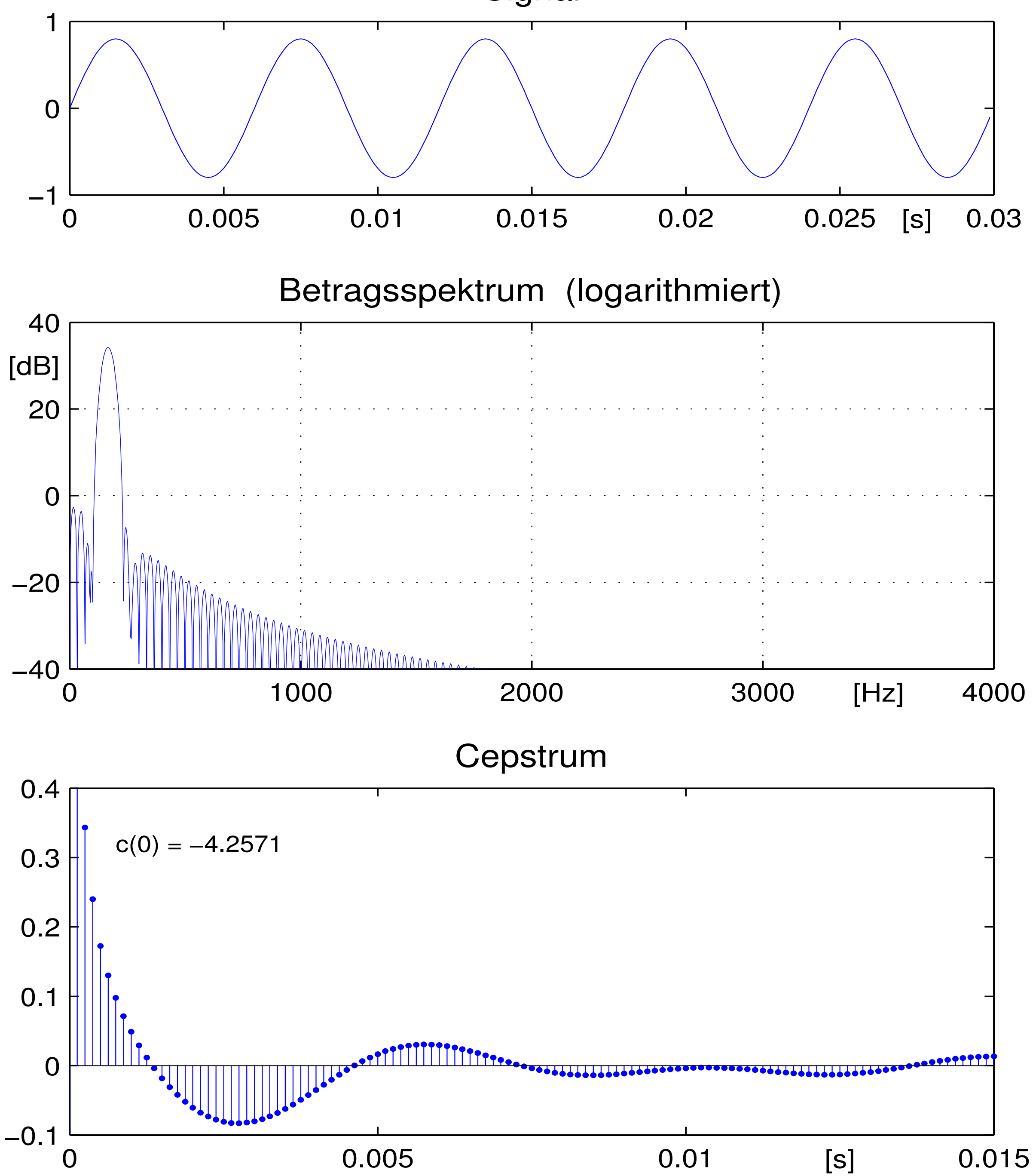


Frage: Wie sieht das Cepstrum $c(m)$ aus?

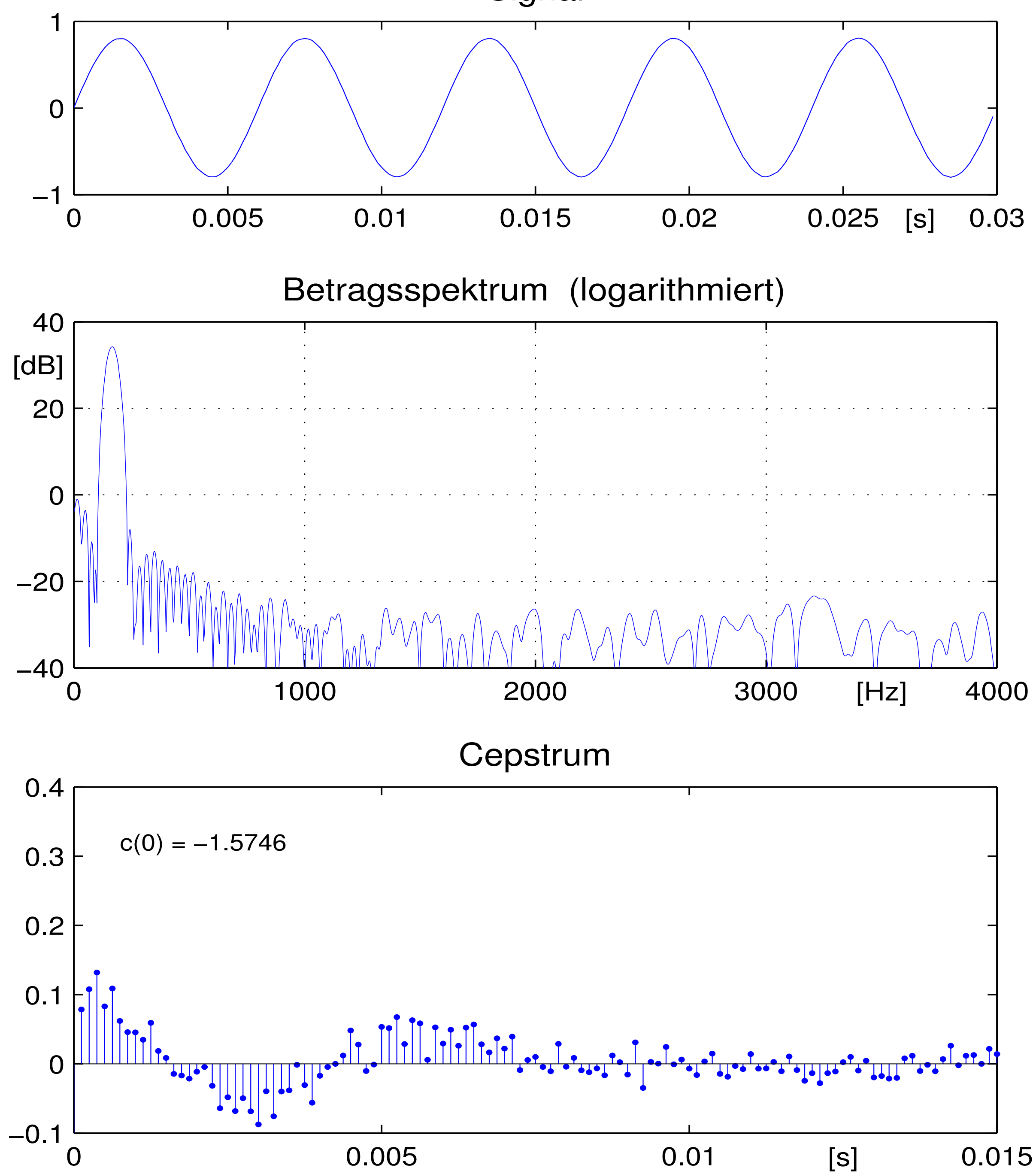



Signal
1
0
−1
0
0.005
0.01
0.015
0.02
0.025
[s]
0.03
Betragsspektrum (linear)
60
50
40
30
20
10
0
167 Hz
1000
2000
3000
[Hz]
4000
Betragsspektrum (logarithmiert)
[dB]
0
−50
−100
0
1000
2000
3000
[Hz]
4000


1
0
−1
0
0.005
0.01
0.015
0.02
0.025
[s]
0.03
Betragsspektrum (logarithmiert)
[dB]
0
−50
−100
0
1000
2000
3000
[Hz]
4000
Cepstrum
0.4
0.3
0.2
0.1
0
−0.1
c(0) = −6.9355
0
0.005
0.01
[s]
0.015


1
0
−1
0
0.005
0.01
0.015
0.02
0.025
[s]
0.03
Betragsspektrum (logarithmiert)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000
Cepstrum
0.4
0.3
0.2
0.1
0
−0.1
c(0) = −4.2571
0
0.005
0.01
[s]
0.015

SNR ≈ 60 dB

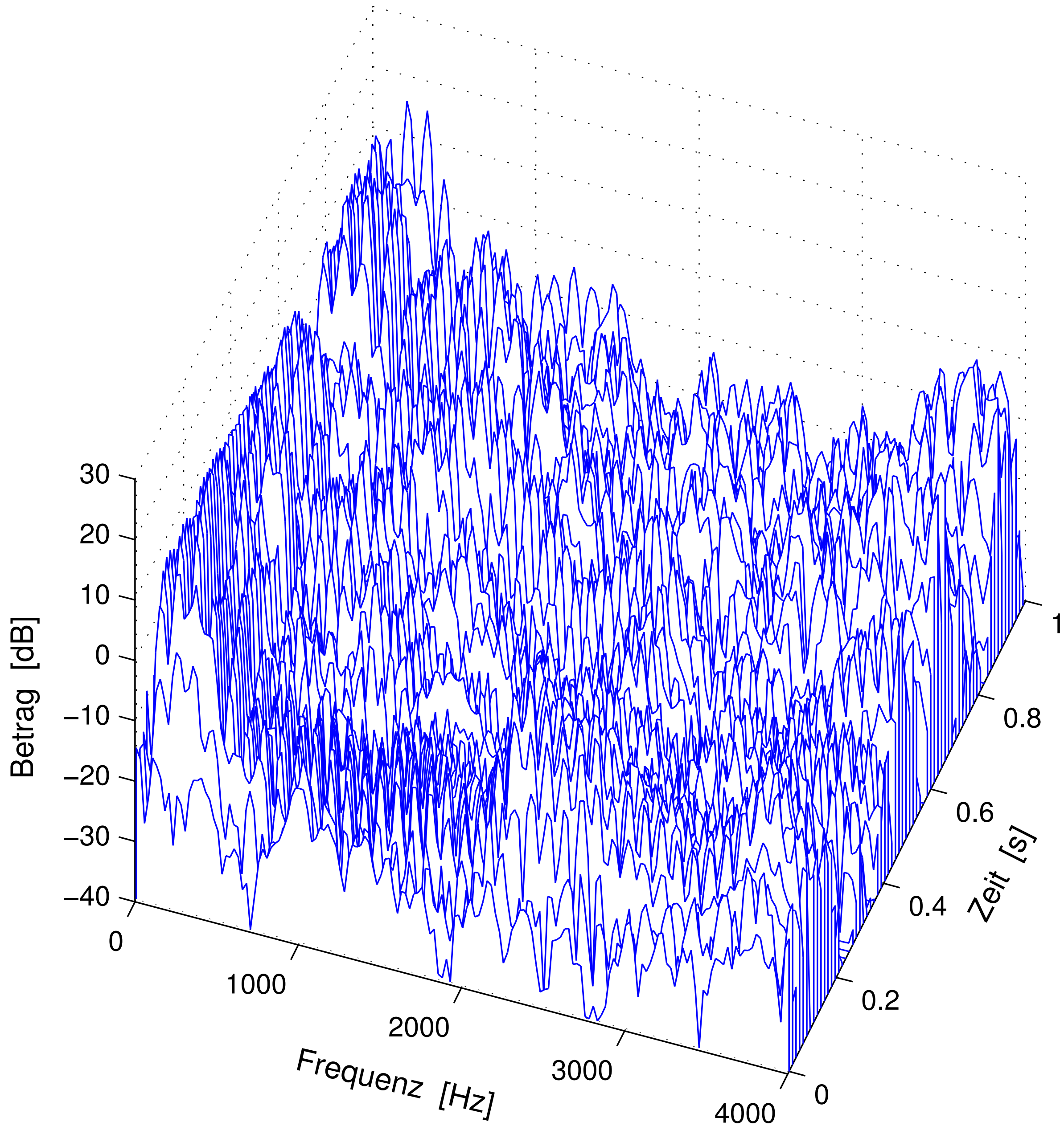


Signal
1
0
−1
0
0.005
0.01
0.015
0.02
0.025
[s]
0.03
Betragsspektrum (logarithmiert)
40
[dB]
20
0
−20
−40
0
1000
2000
3000
[Hz]
4000
Cepstrum
0.4
0.3
0.2
0.1
0
−0.1
c(0) = −1.5746
0
0.005
0.01
[s]
0.015

<<<

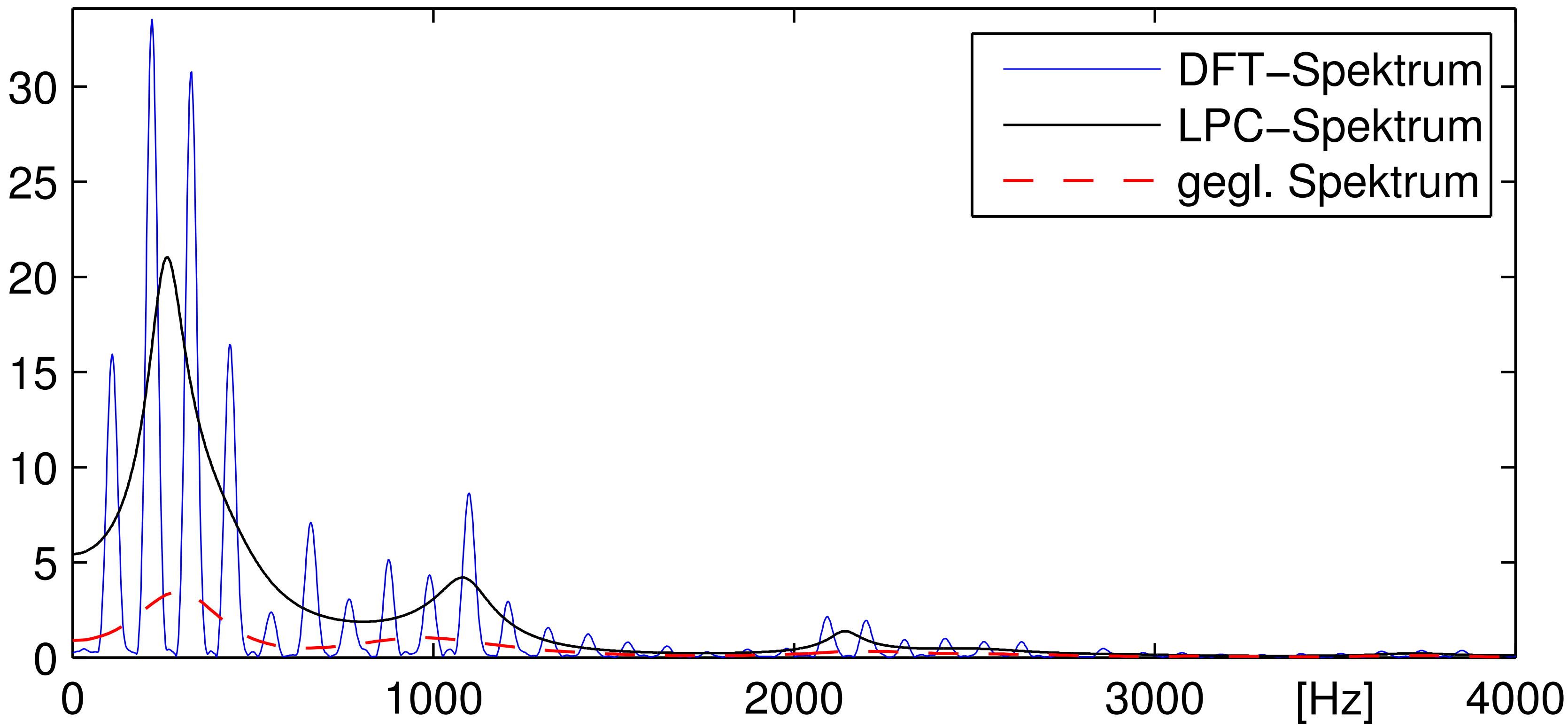
Kurzzeitspektrum

3-dimensionale Darstellung



<<<

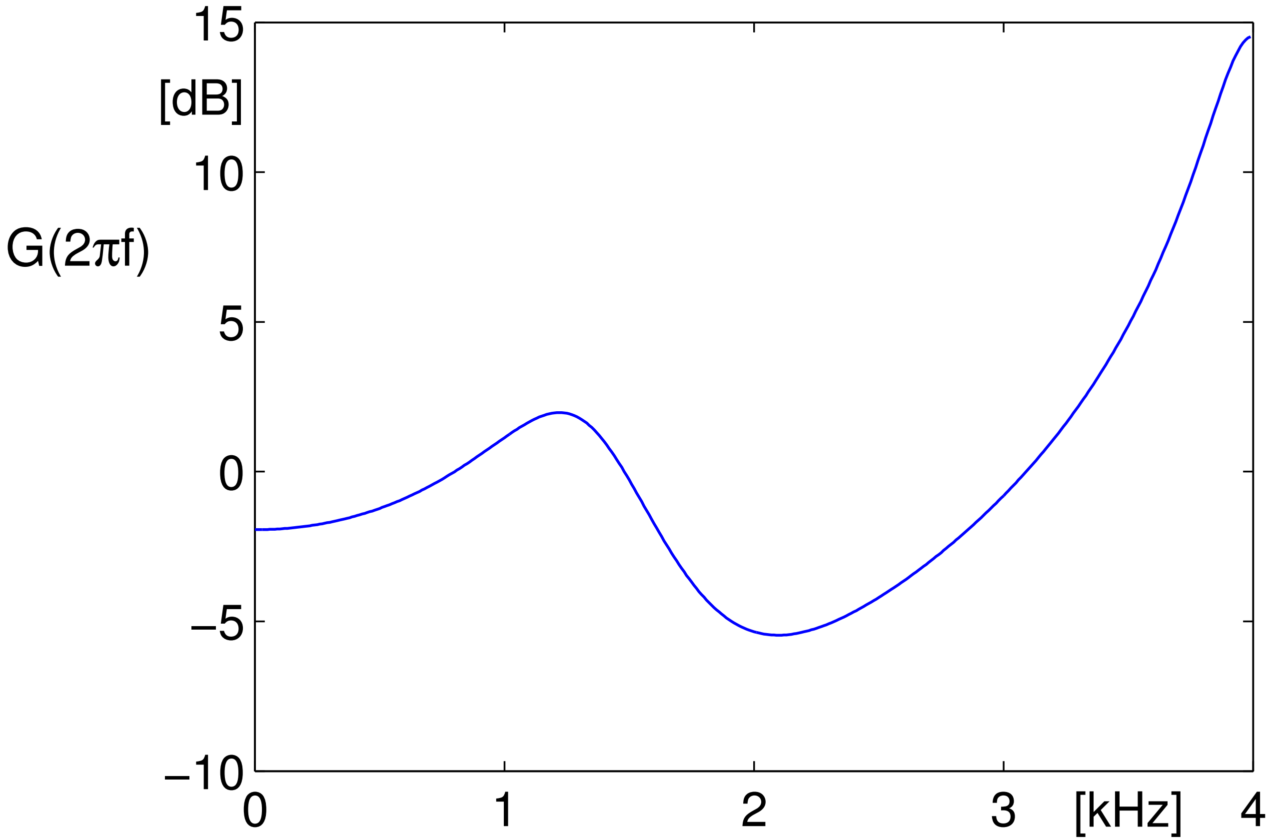
Vergleich geglätteter Spektren



lineare Darstellung der Spektren

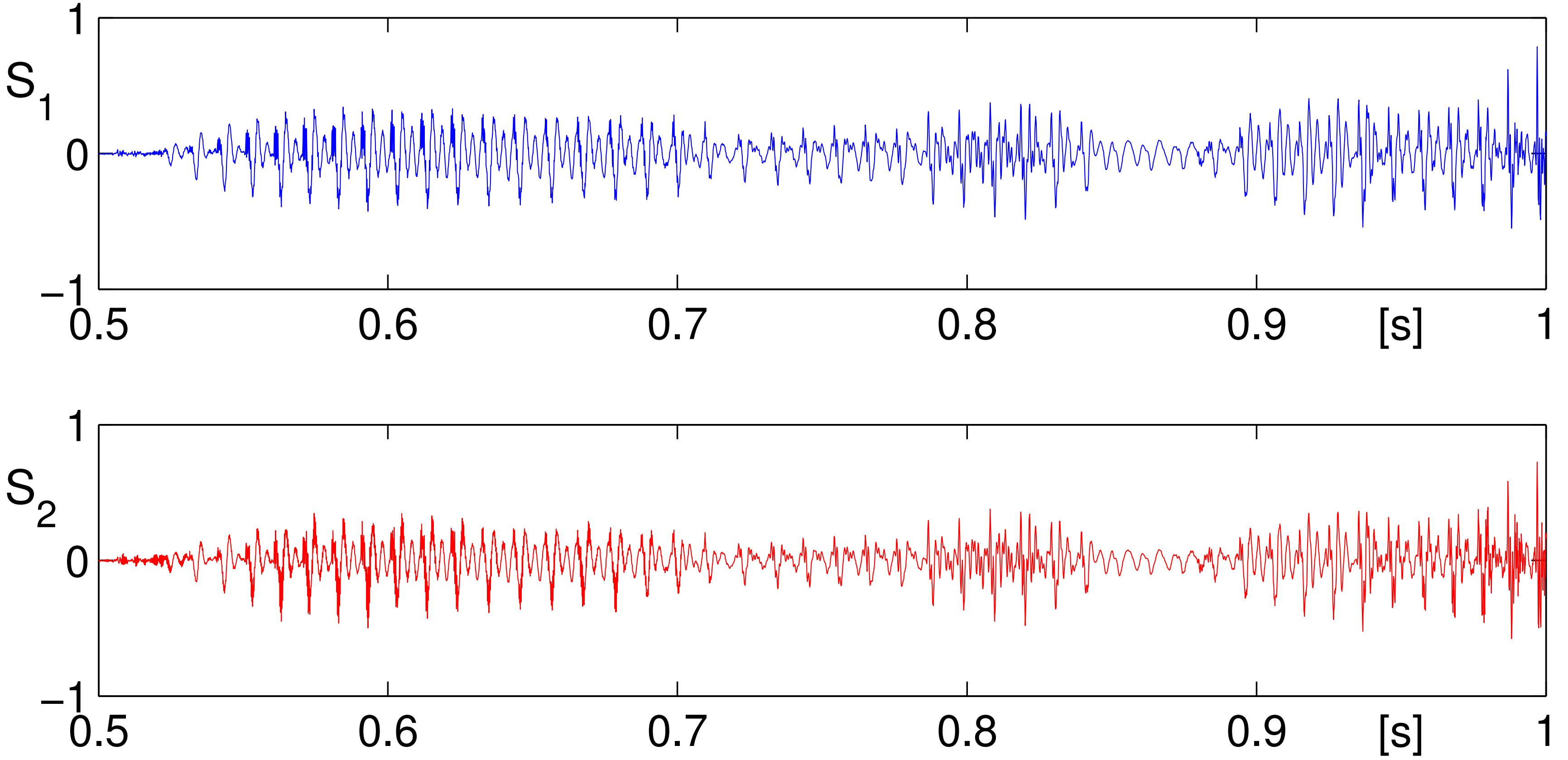
<<<

Charakteristik des Übertragungskanals



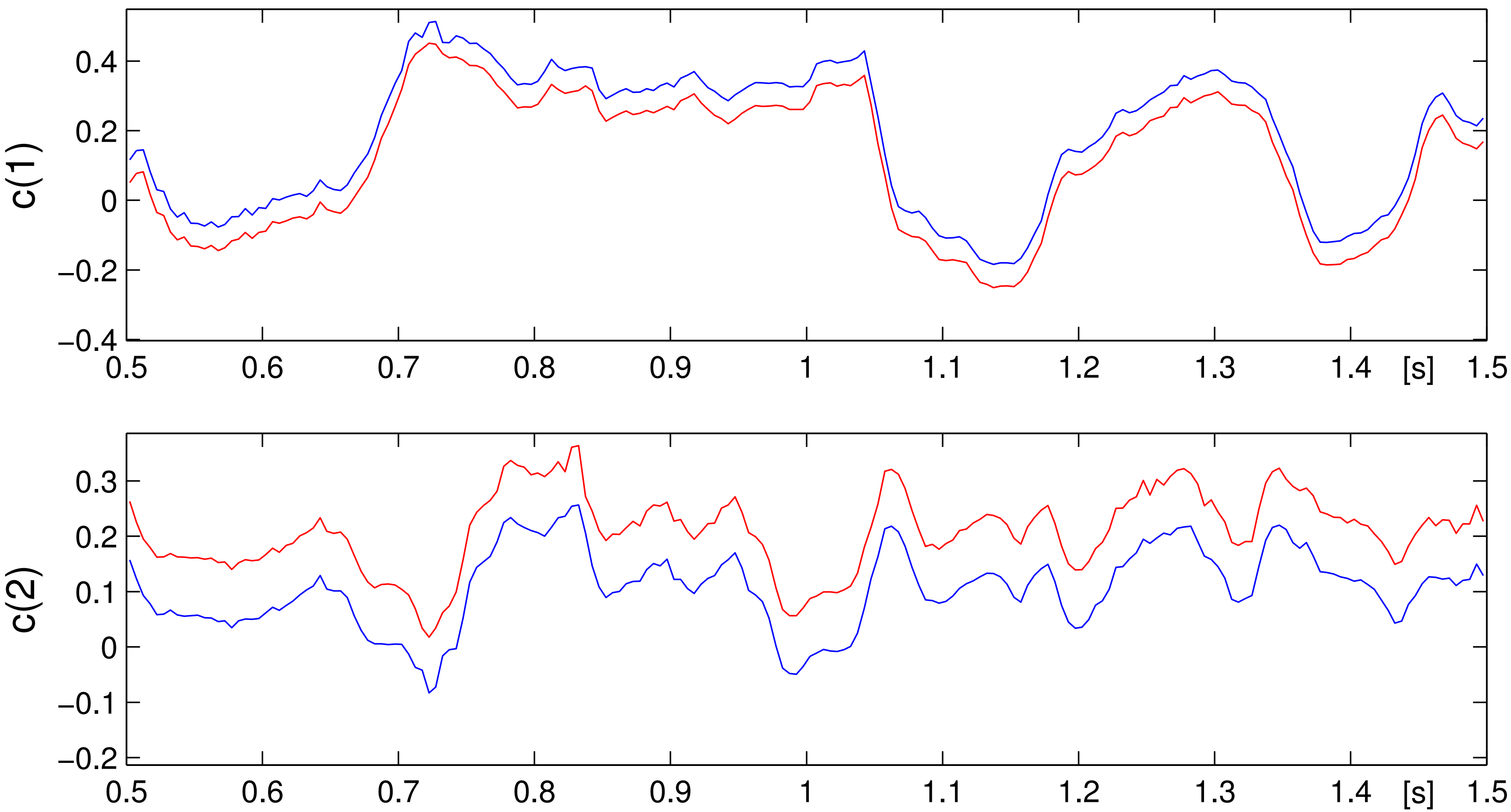
<<<

Signale am Ein- und Ausgang des Kanals



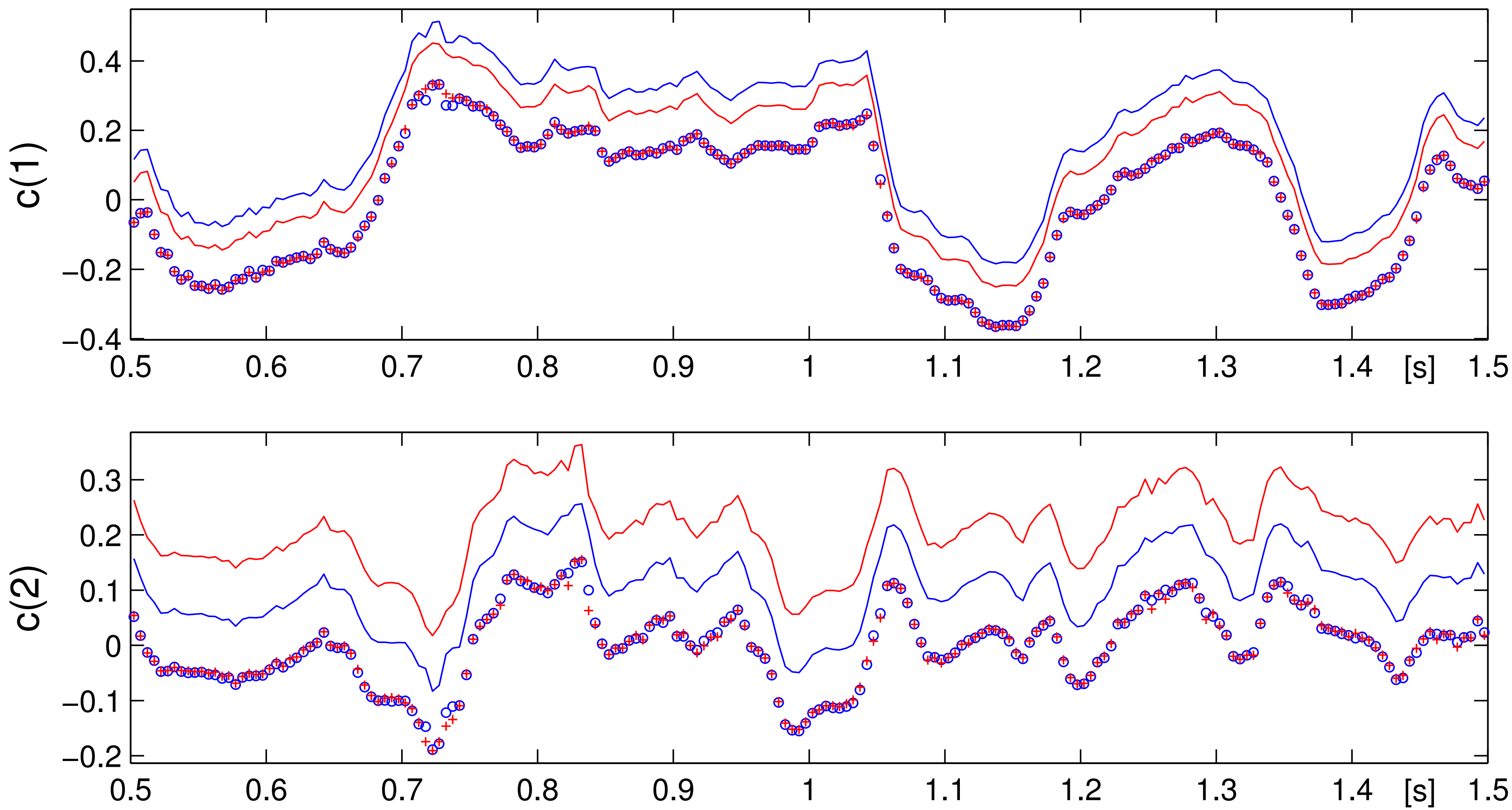
<<<

Wirkung des Übertragungskanals auf Cepstren von $s_1(n)$ und $s_2(n)$



<<<

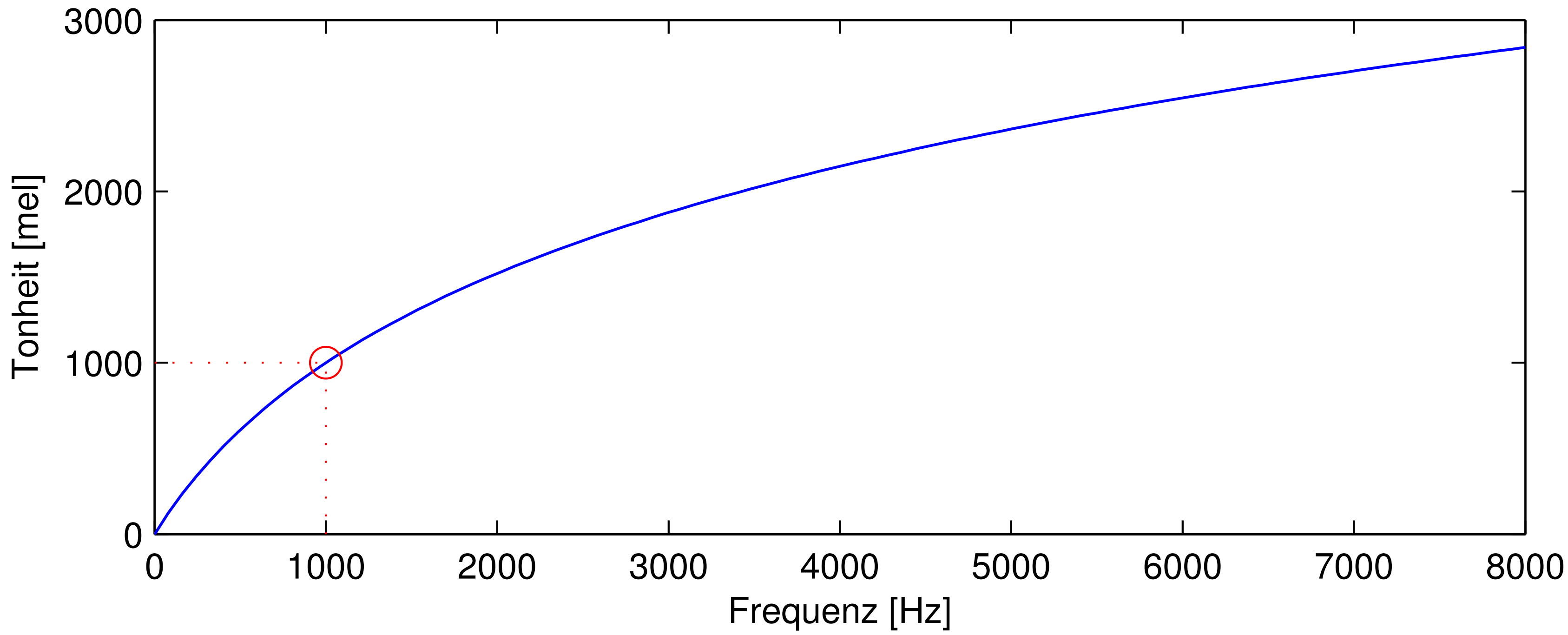
Übertragungskanal hat auf mittelwertfreies Cepstrum keine Auswirkung !



<<<

Mel-Skala

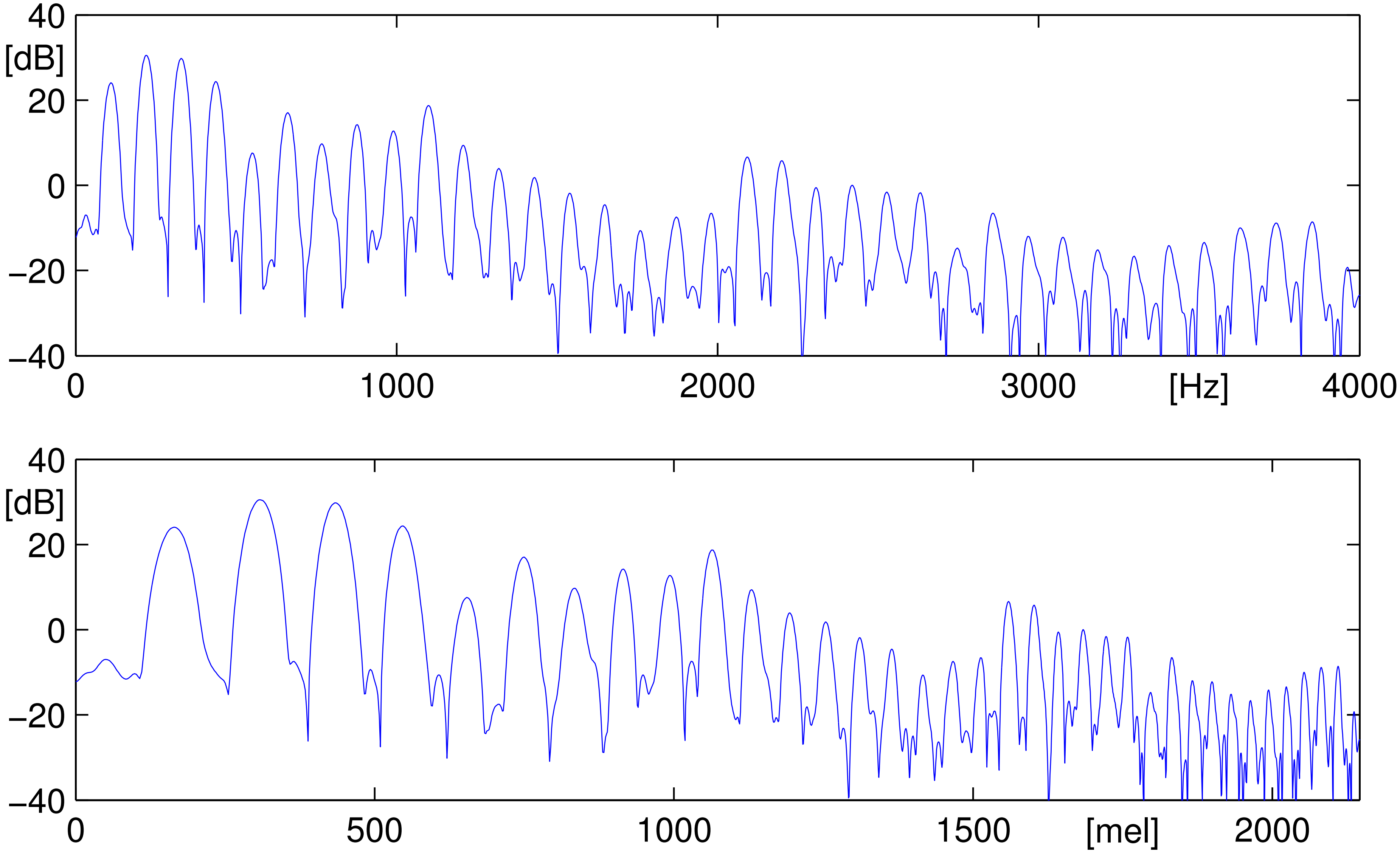
Zusammenhang zwischen Frequenz und Tonheit



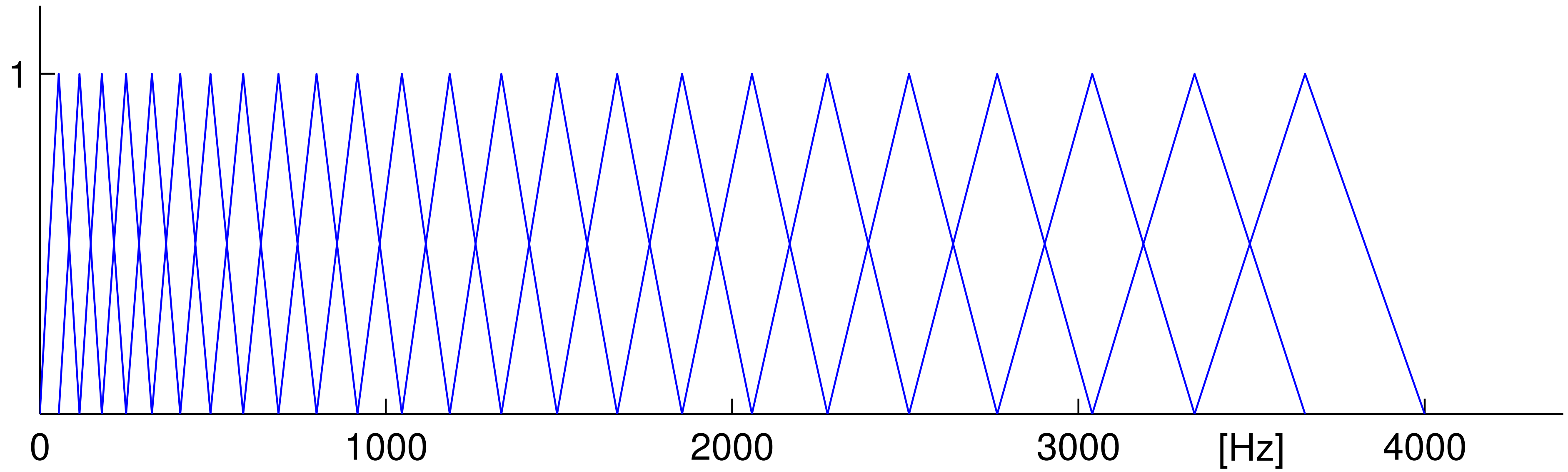
Psychoakustik:

- Wahrnehmung der logarithmischen Frequenz als Tonhöhe
- Frequenzgruppenbreite mit zunehmender Frequenz grösser

Spektrum mit Hz-Skala vs. Mel-Skala

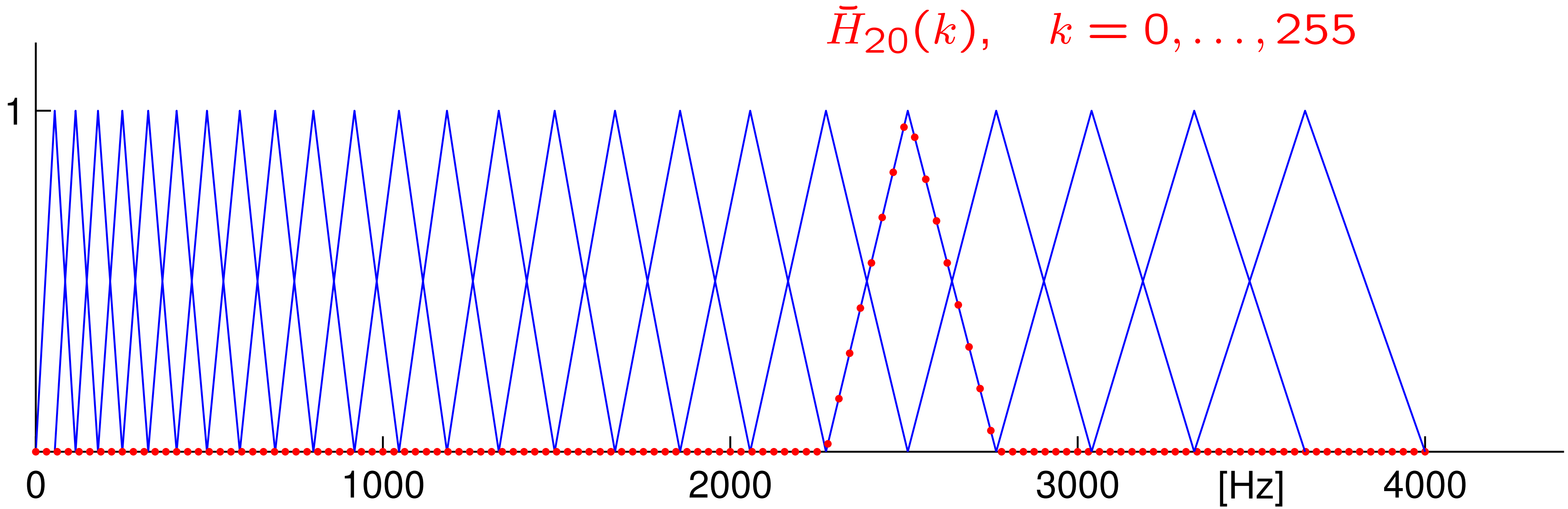


Mel-Filterbank   nicht-uniform, 24 Kanäle



Filterung: Multiplikation der Fourier-Spektren (im Frequenzbereich)

Mel-Filterbank nicht-uniform, 24 Kanäle



Filterung: $\bar{S}_j = \sum_k X(k) \cdot \breve{H}_j(k) \;, \quad 1 \leq j \leq J$

<<<