Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Sprachverarbeitung I / 9 HS 2016

Sprachsynthese: Fourier-Analyse-Synthese

Buch: Kapitel 9.2.5

Beat Pfister

Sprachverarbeitung I / 9

Vorlesung: **Sprachsynthese** (Teil I.4)

Dauer- und Grundfrequenzveränderung von Sprachsignalen mittels Fourier-Analyse-Synthese

Dieses Thema wird im Buch Ausgabe 2017 nur noch rudimentär behandelt.
Eine ausführliche Behandlung ist im Anhang B der Ausgabe 2008 zu finden.
(Dieser Anhang ist in ZIP-Archiv Folien_SPV1.zip enthalten.)

Übung: Grundfrequenzsteuerung

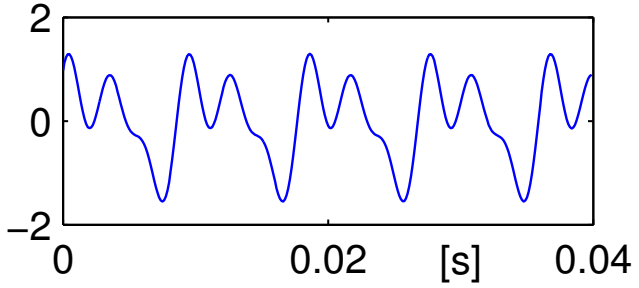
Sprachsynthese

Verkettung von Diphonen:

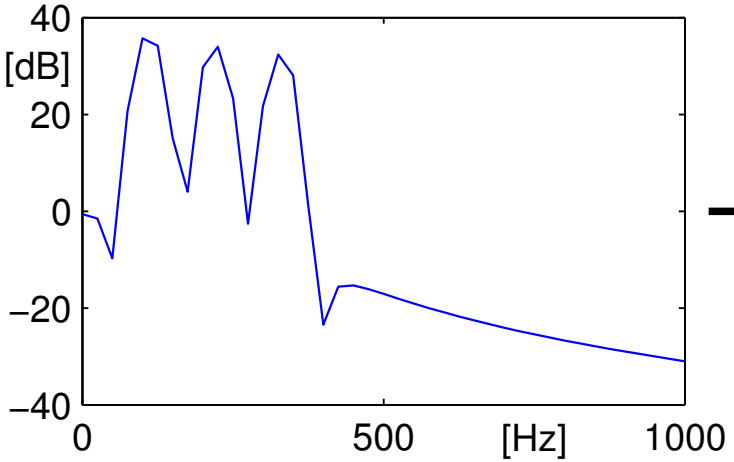
$\longrightarrow$ Dauer und Grundfrequenz müssen verändert werden
(einzeln, d.h. voneinander unabhängig)

Methoden:
- LPC-Analyse-Synthese
- PSOLA (pitch-synchronous overlap add)
- Fourier-Analyse-Synthese >>>

Dauerveränderung mittels Fourier-Analyse-Synthese

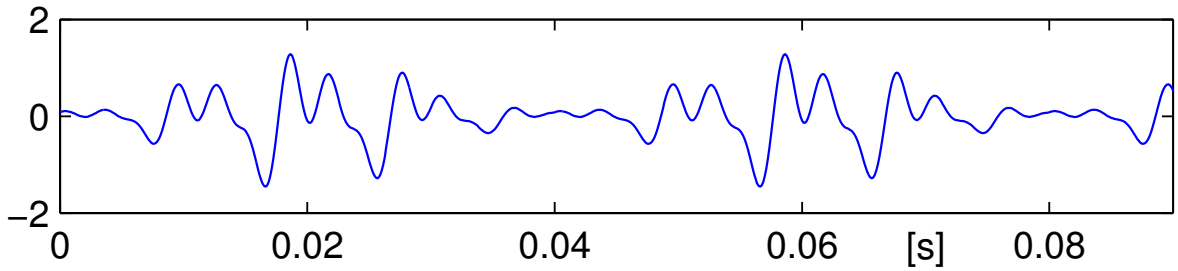


Signal mit 3 Komponenten
($f_s = 8000\,\text{Hz}$, $F_0 = 110\,\text{Hz}$)

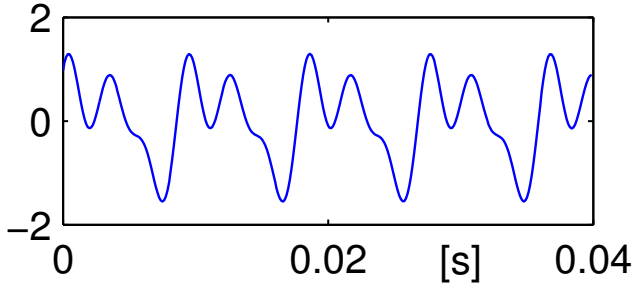


N-Punkt-DFT
($N = 0.04 \cdot f_s = 320$)

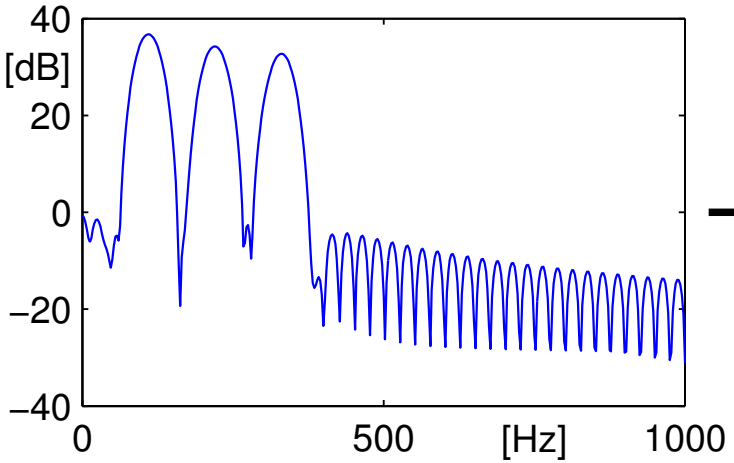
Signal verlängert auf 225 % Dauer



Dauerveränderung mittels Fourier-Analyse-Synthese

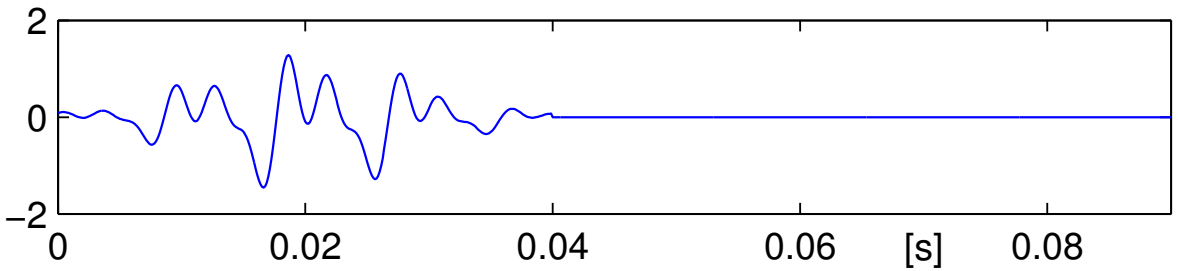


Signal mit 3 Komponenten
($f_s = 8000\,\text{Hz}$, $F_0 = 110\,\text{Hz}$)



hochauflösende DFT !

Signal verlängert auf 225 % Dauer



Fourier-Analyse-Synthese

Dauer und Grundfrequenz eines Sprachsignals verändern:

1. Zerlegung des Signals in Sinuskomponenten >>>

2. Veränderung der Frequenz der Sinuskomponenten

3. Rekonstruktion der Signalabschnitte mit gewünschter Dauer

Fourier-Analyse-Synthese

Dauer und Grundfrequenz eines Sprachsignals verändern:

1. Zerlegung des Signals in Sinuskomponenten
   $\longrightarrow$ Schätzen der spektralen Zusammensetzung >>>

2. Veränderung der Frequenz der Sinuskomponenten

3. Rekonstruktion der Signalabschnitte mit gewünschter Dauer

Fourier-Analyse-Synthese

Dauer und Grundfrequenz eines Sprachsignals verändern:

1. Zerlegung des Signals in Sinuskomponenten
   $\longrightarrow$ Komponenten schätzen aus Maxima der hochauflösenden FT
   $\longrightarrow$ Signal ist nicht stationär >>>
2. Veränderung der Frequenz der Sinuskomponenten
3. Rekonstruktion der Signalabschnitte mit gewünschter Dauer

Fourier-Analyse-Synthese

Dauer und Grundfrequenz eines Sprachsignals verändern:

1. Zerlegung des Signals in Sinuskomponenten
   $\longrightarrow$ Komponenten schätzen aus Maxima der hochauflösenden FT
   $\longrightarrow$ Abschnittweise Analyse (Kurzzeitanalyse)
2. Veränderung der Frequenz der Sinuskomponenten
3. Rekonstruktion der Signalabschnitte mit gewünschter Dauer

Fourier-Analyse-Synthese

Dauer und Grundfrequenz eines Sprachsignals verändern:

1. Zerlegung des Signals in Sinuskomponenten
   $\longrightarrow$ Komponenten schätzen aus Maxima der hochauflösenden FT
   $\longrightarrow$ Abschnittweise Analyse (Kurzzeitanalyse)
2. Veränderung der Frequenz der Sinuskomponenten
   (wird später behandelt)
3. Rekonstruktion der Signalabschnitte mit gewünschter Dauer

Fourier-Analyse-Synthese

Dauer und Grundfrequenz eines Sprachsignals verändern:

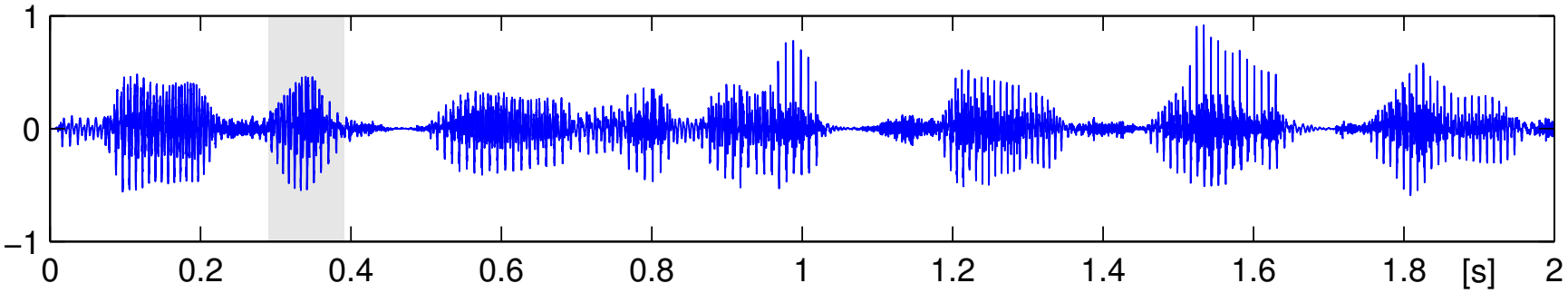
1. Zerlegung des Signals in Sinuskomponenten
   $\longrightarrow$ Komponenten schätzen aus Maxima der hochauflösenden FT
   $\longrightarrow$ Abschnittweise Analyse (Kurzzeitanalyse)
2. Veränderung der Frequenz der Sinuskomponenten
   (wird später behandelt)
3. Rekonstruktion der Signalabschnitte mit gewünschter Dauer >>>

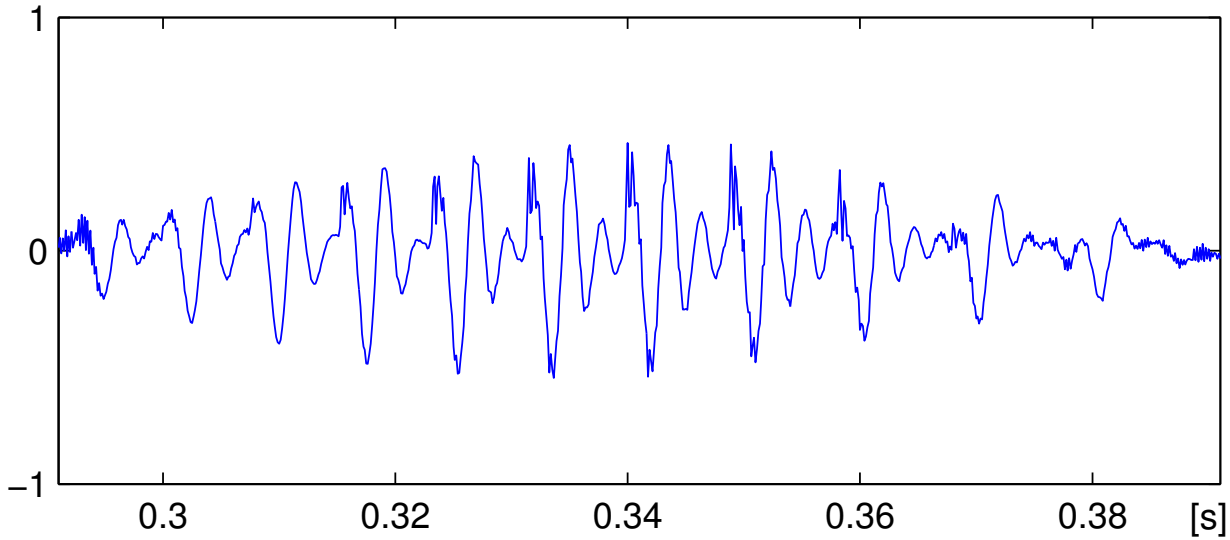
Fourier-Analyse-Synthese

Dauer und Grundfrequenz eines Sprachsignals verändern:

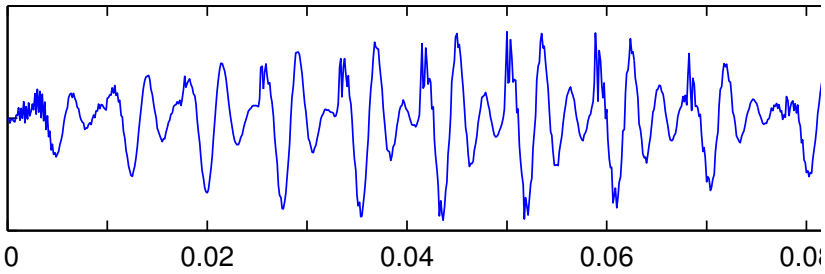
1. Zerlegung des Signals in Sinuskomponenten
   $\longrightarrow$ Komponenten schätzen aus Maxima der hochauflösenden FT
   $\longrightarrow$ Abschnittweise Analyse (Kurzzeitanalyse)
2. Veränderung der Frequenz der Sinuskomponenten
   (wird später behandelt)
3. Rekonstruktion der Signalabschnitte mit gewünschter Dauer

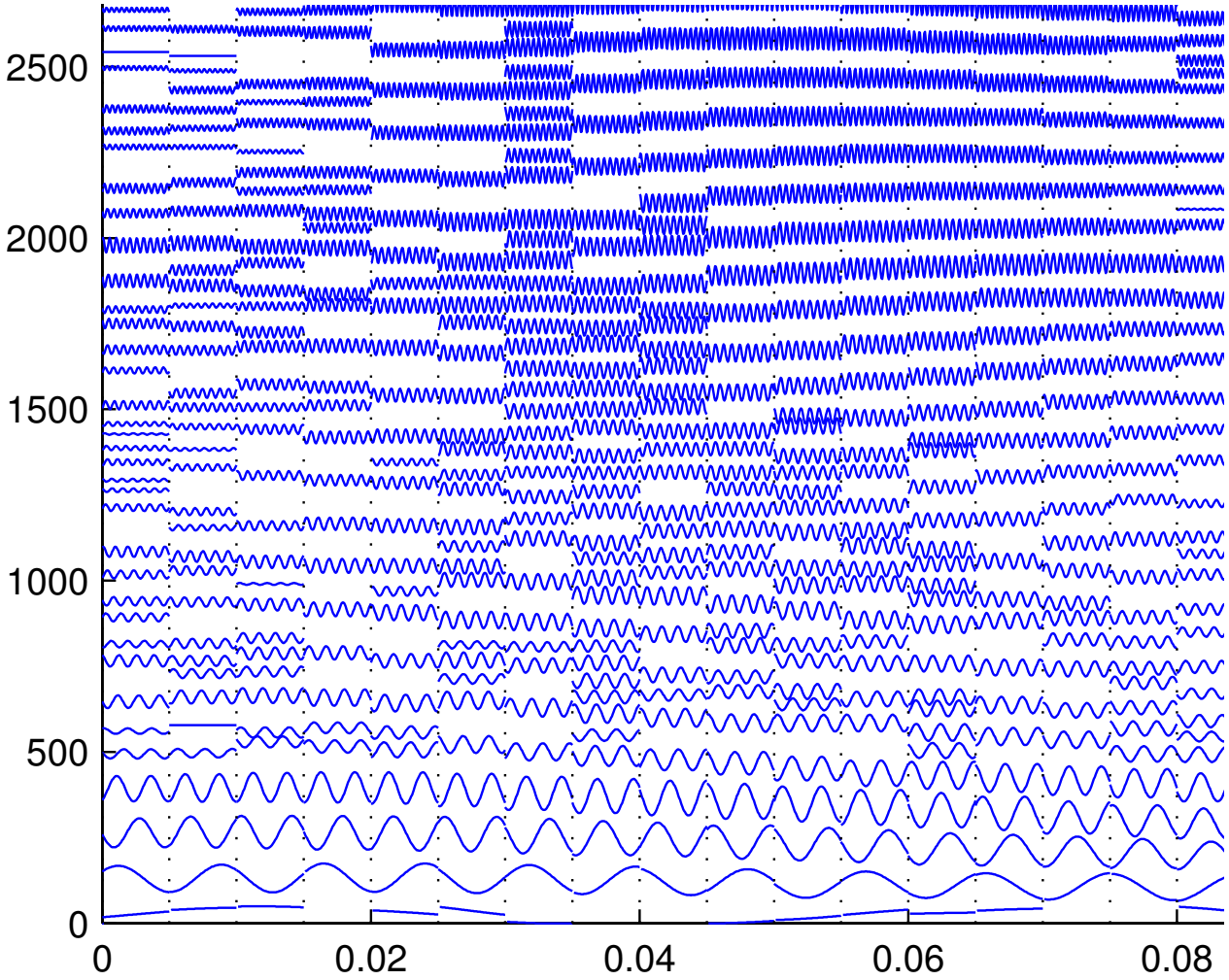
Anwendung auf einen Sprachsignalausschnitt





Spektrum eines Sprachsignals mit fallendem F_0



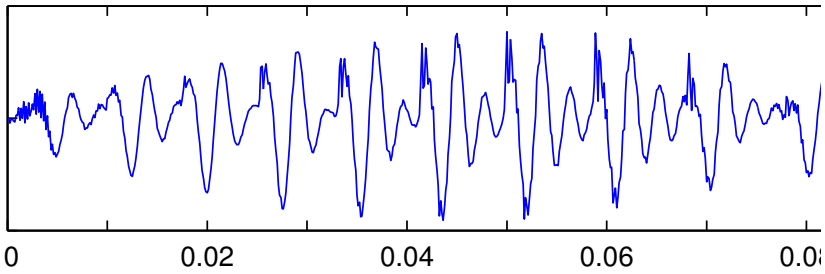


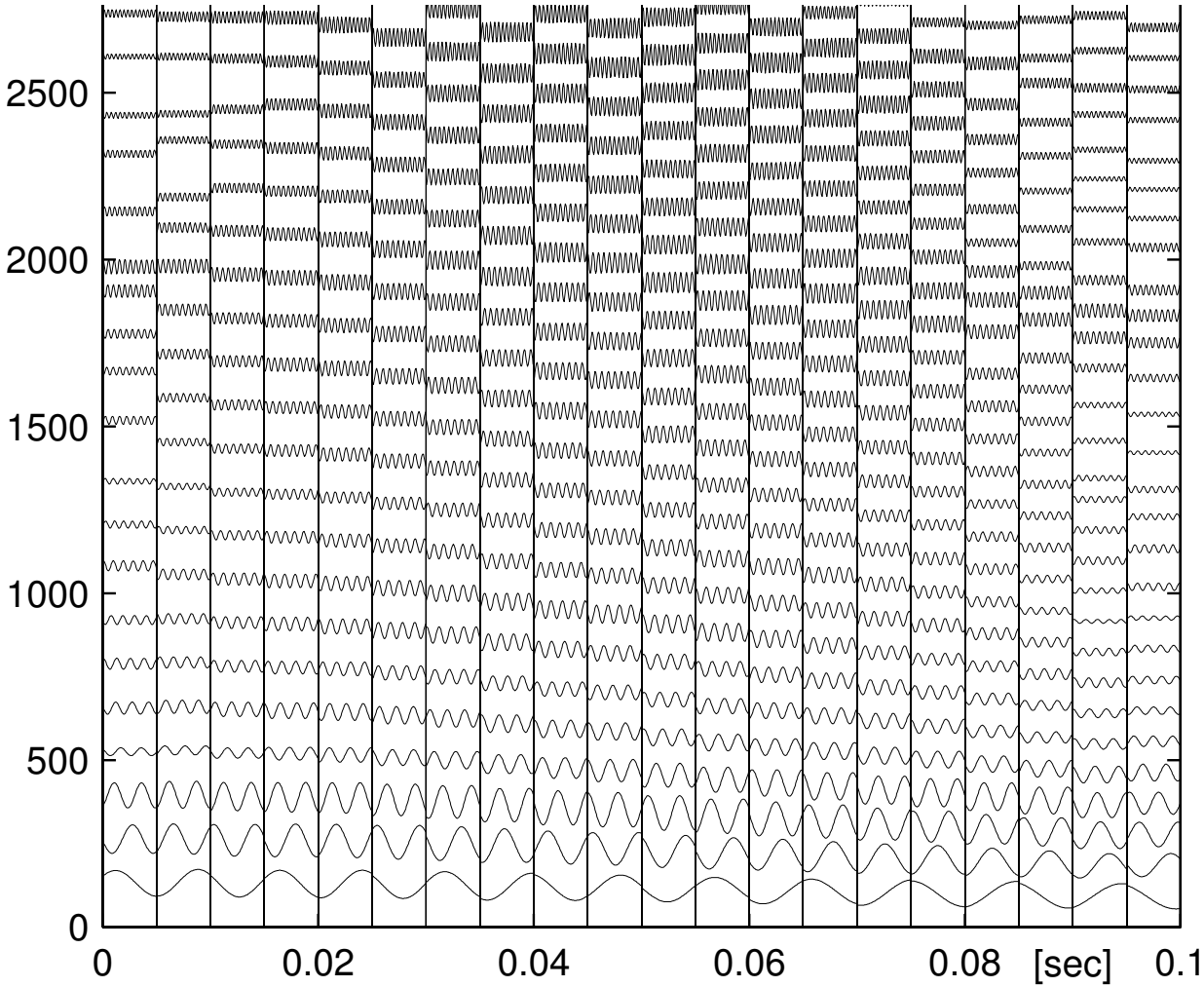
Beobachtung:

Spektrum des stimmhaften Signals ist nicht harmonisch!

Erwartes Spektrum: ?

Spektrum eines Sprachsignals mit fallendem F_0





Beobachtung:

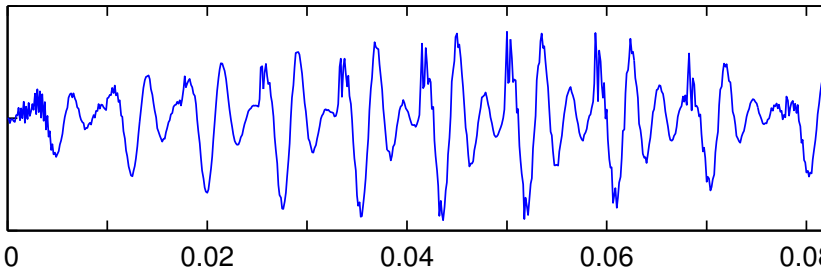
F_0 des Signals fällt schnell
($\approx$ 7 Oktaven/s)

Frage:

Ist das ein Problem ?

>>>

Spektrum eines Sprachsignals mit fallendem F_0



Sprachsignal:

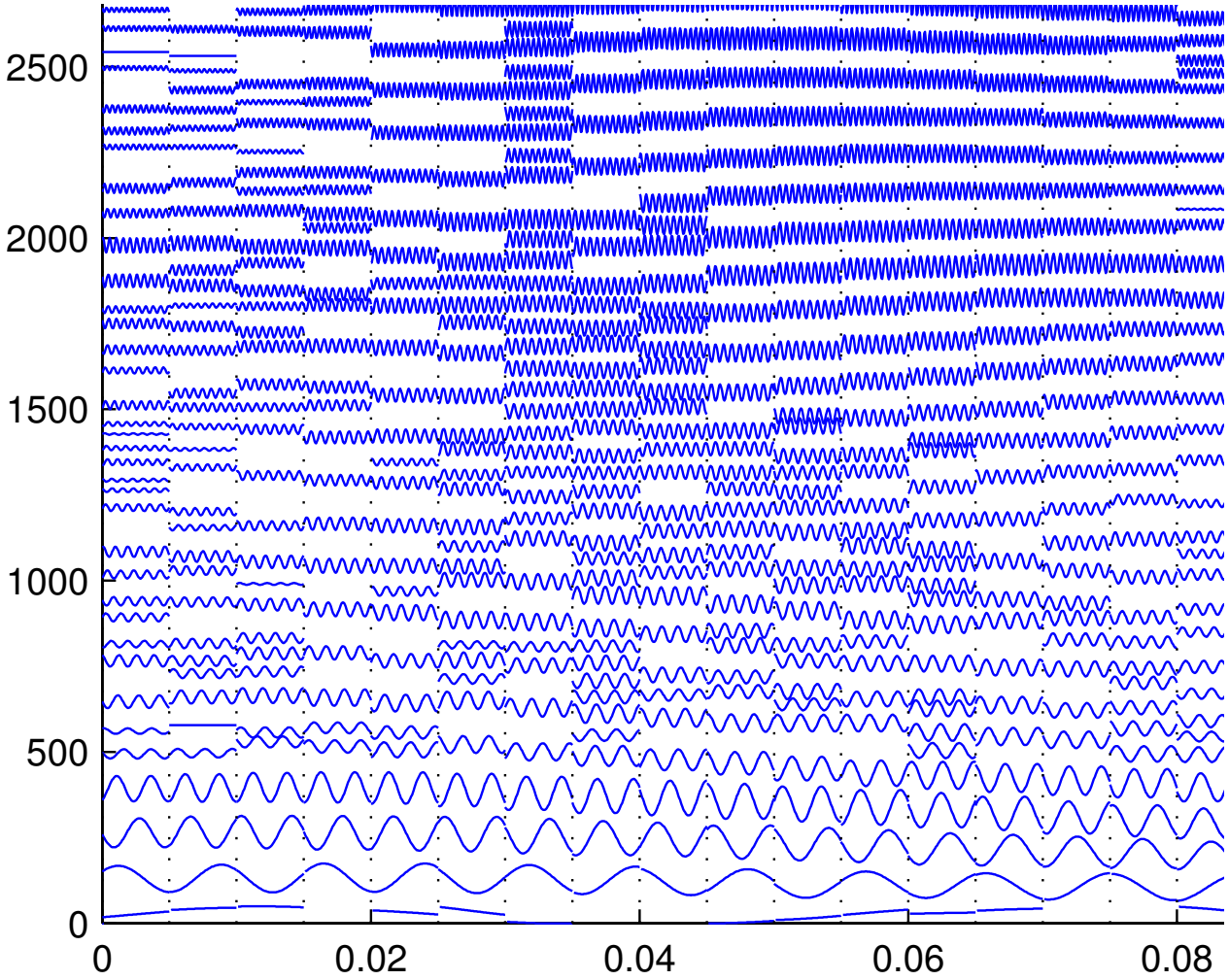
F_0 fällt schnell ($\approx$ 7 Oktaven/s)

Beobachtung:

Spektrum des stimmhaften Signals ist nicht harmonisch!

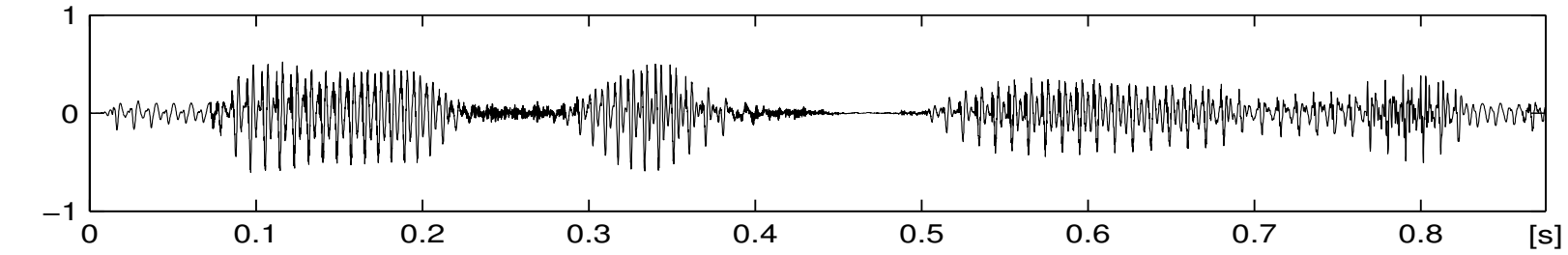
Grund:

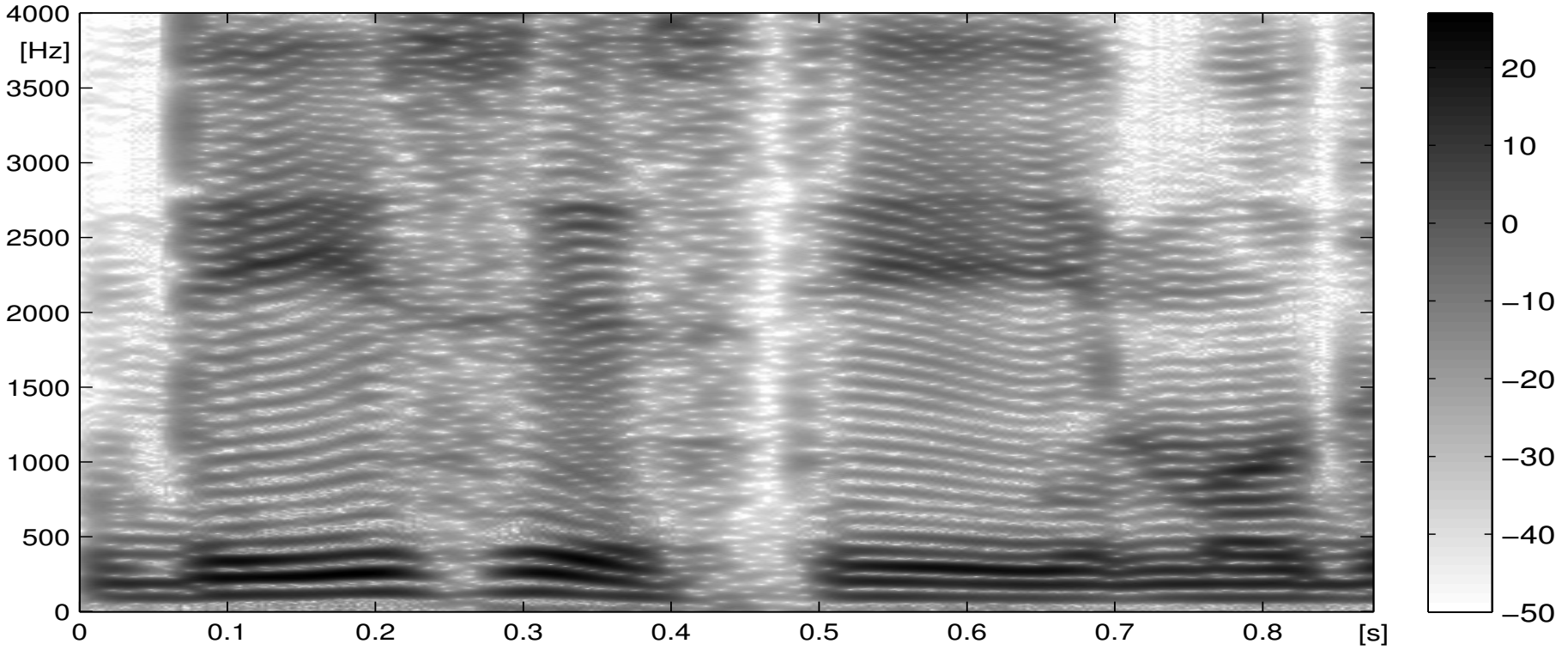
Stationaritätsbedingung der FT ist nicht erfüllt !



Spektrogramm eines Sprachsignals

$F_0 \approx 100\,\text{Hz}$





Probleme bei der Spektralanalyse von Sprachsignalen

- Signal stimmhaft: Spektrum aber nicht harmonisch! >>>

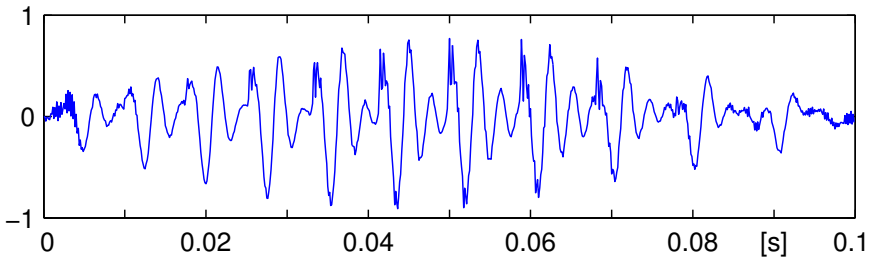
  a) Bereiche vermeintlich harmonisch (systematischer Fehler)
  Grund: Starke F_0-Änderung innerhalb Analyseabschnitt
  (Stationaritätsbedingung der FT ist verletzt) Lösung?

  b) Bereiche ohne Struktur (kein systematischer Fehler)
  Grund: schwache Frequenzkomponenten ungenau Lösung?

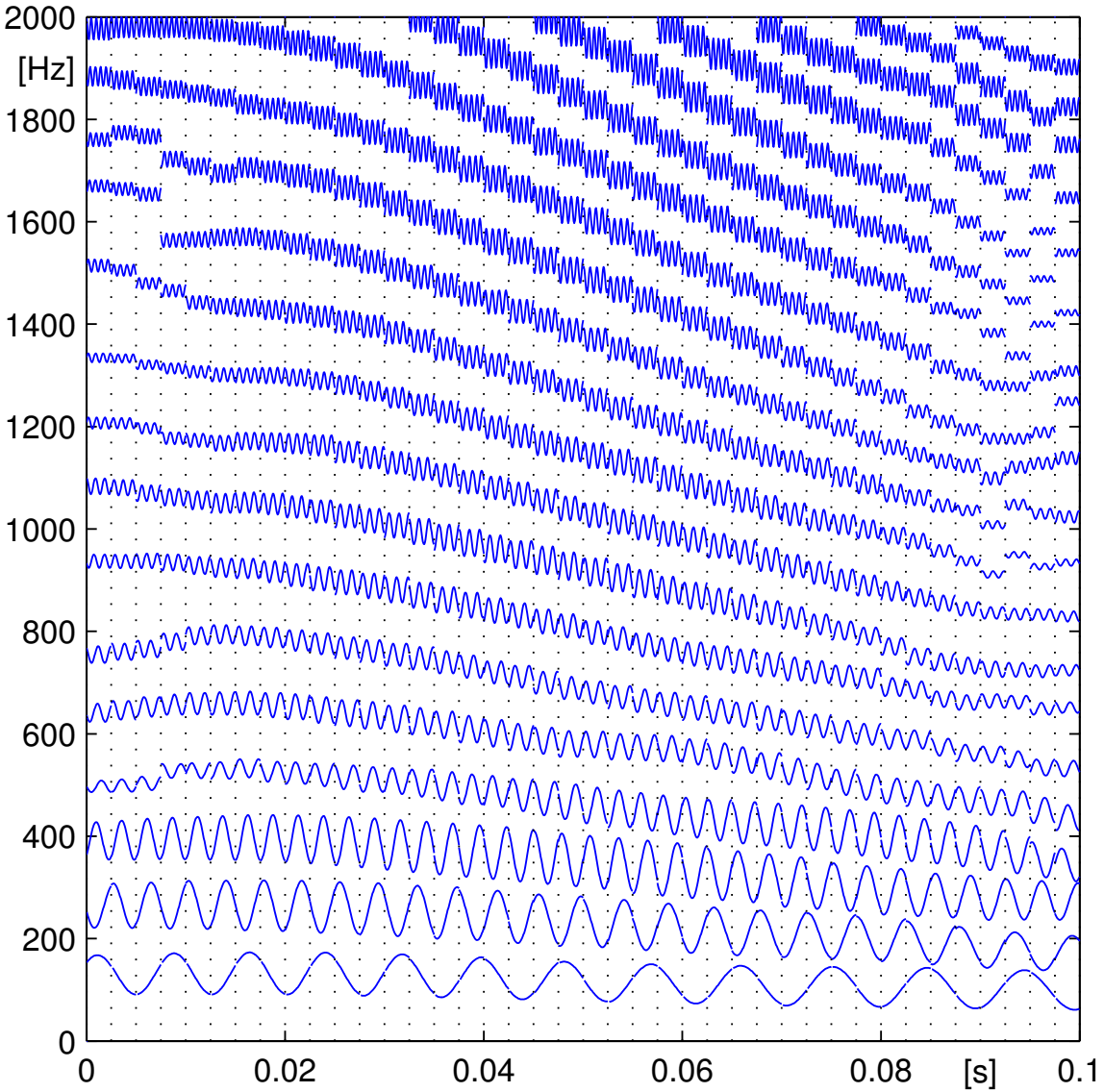
- Signal stimmlos: spektrale Zusammensetzung? >>>

Lösung dieser Probleme $\longrightarrow$ gute Schätzung des Spektrum

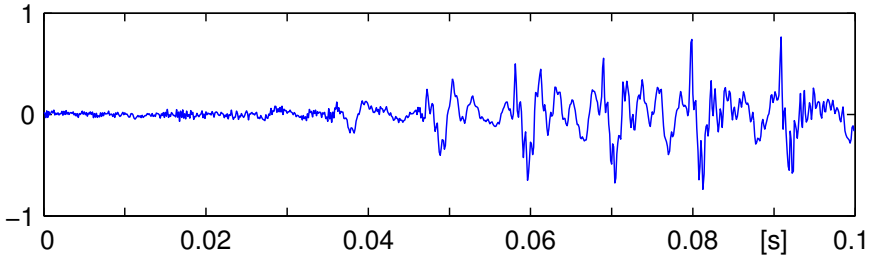
Gute Schätzung der Frequenzkomponenten

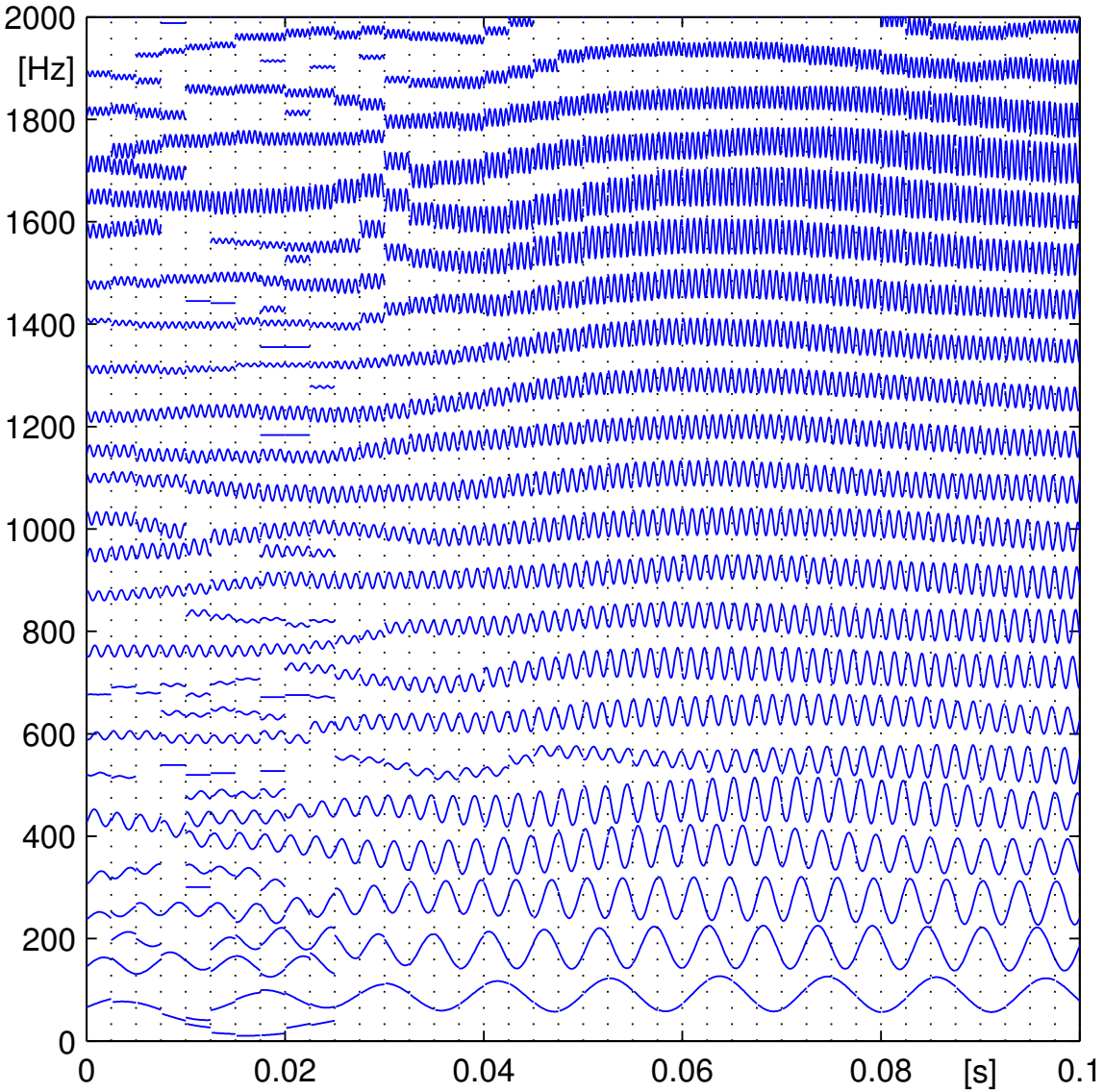


Nach abschnittweiser Stationarisierung liefert die Mehr-Fenster-Fourier-transformation für obiges Sprachsignals ein zuverlässiges Resultat



Gute Schätzung der Frequenzkomponenten





Rekonstruktion des Sprachsignal

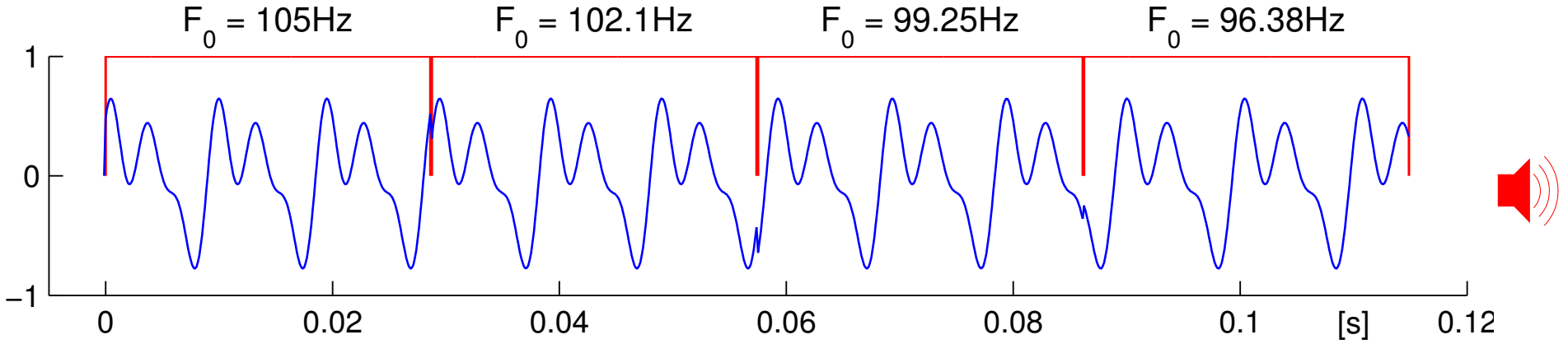
Gegeben: Schätzung der zeitabhängigen Signalkomponenten
(das “wirkliche” Spektrum)

Frage: Wie wird daraus das Signal rekonstruiert?
(ev. mit Dauer- und / oder F_0-Veränderung)

Signalrekonstruktion aus dem “wirklichen” Spektrum

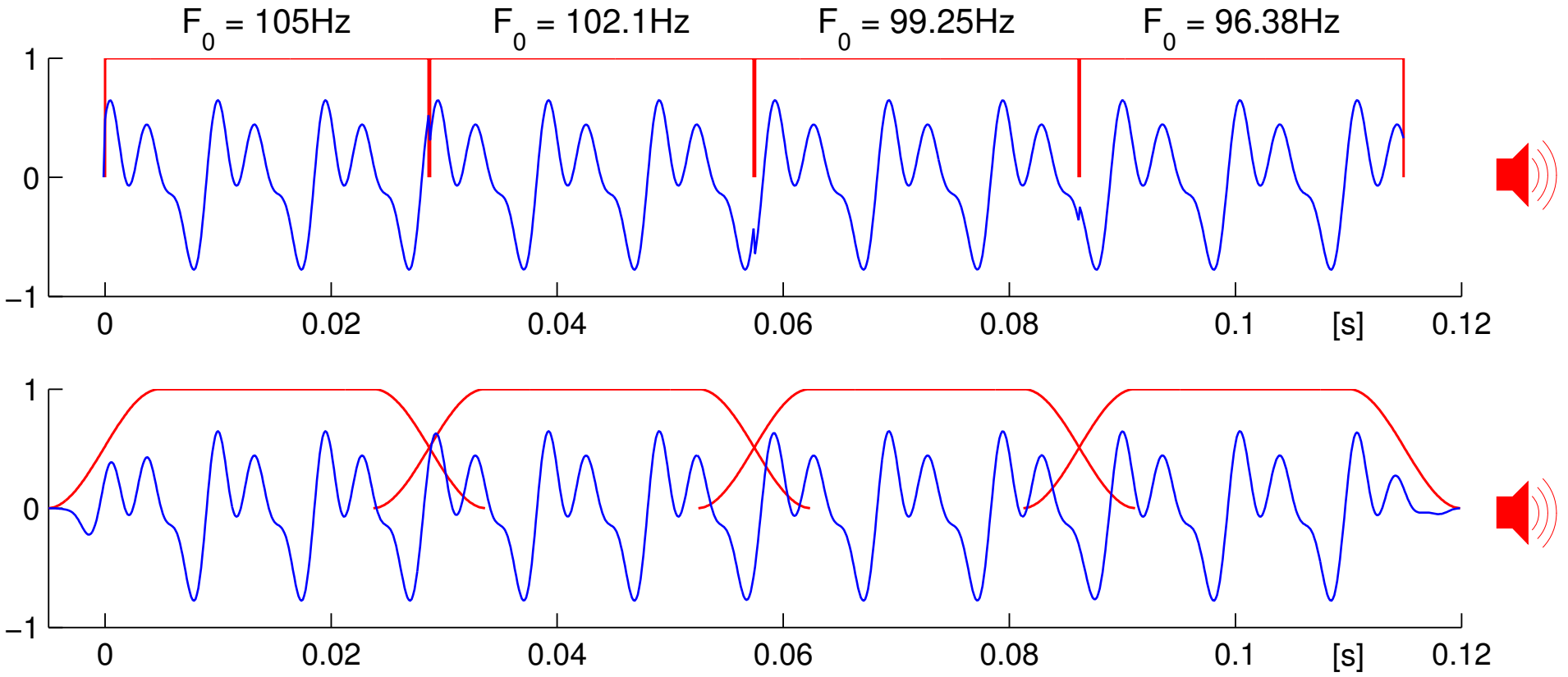
Prinzip: Rekonstruktion des Signals innerhalb eines Abschnittes durch Superposition der Sinussignale, die den Frequenzkomponenten entsprechen

Problem: An den Abschnittsgrenzen ergeben sich i.a. Unstetigkeitsstellen bzw. Diskontinuitäten

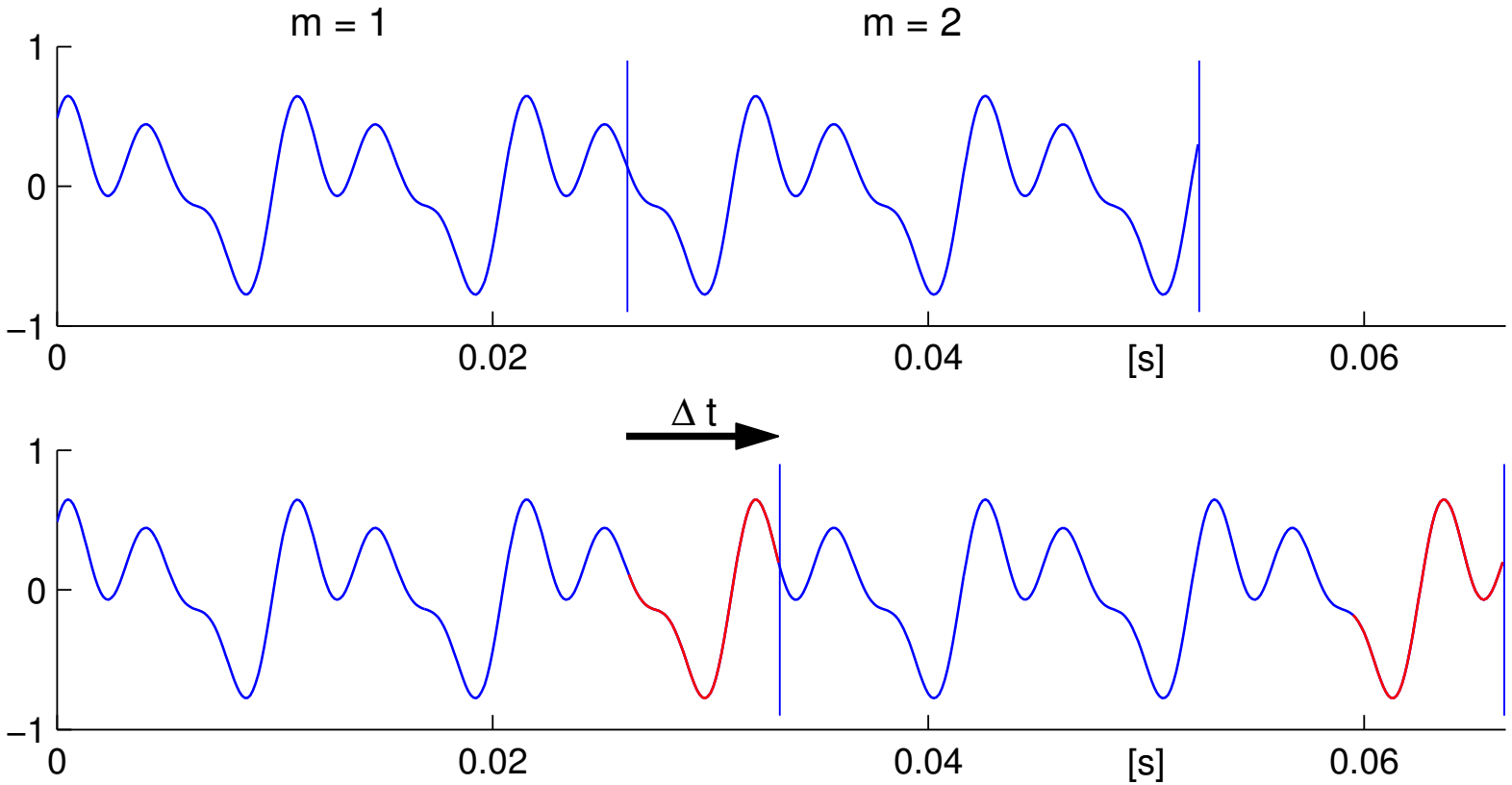


Rekonstruktion des Signals mittels OLA-Technik

Die abrupten Übergänge zwischen den Abschnitten werden durch “gleitende” ersetzt $\longrightarrow$ Überblenden (*windowed overlap-add*)



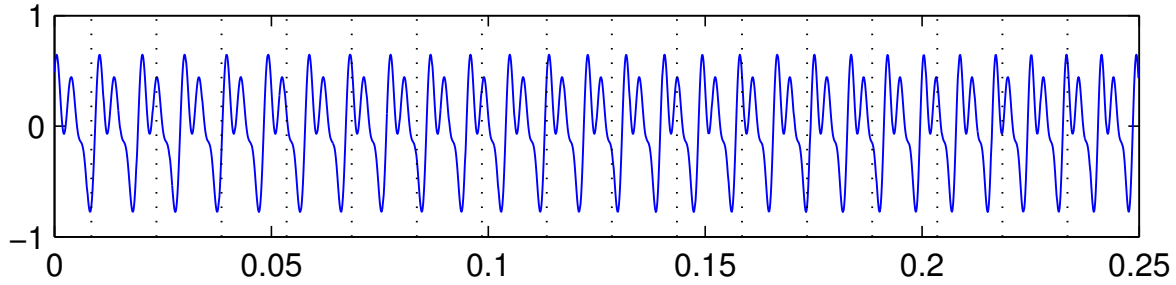
Dauerveränderung bei der Rekonstruktion des Signals



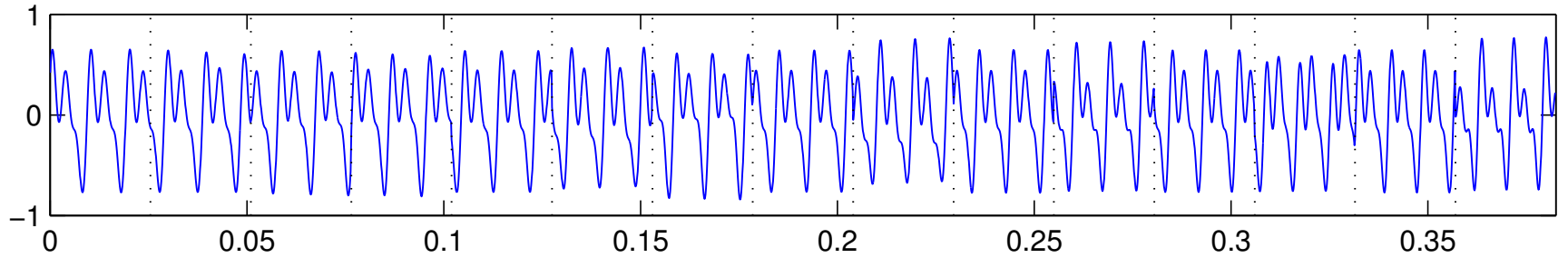
Kompensation der Phase der i-ten Komponente: $\Delta p_m(i) = 2\pi f_m(i)\,(m-1)\Delta t$

Dauerveränderung eines instationären Signals

Originalsignal: Grundfrequenzverlauf $100\,(1+t)$ Hz ($\approx$ 2 Okt/sec)



Verändertes Signal: Streckungsfaktor 1.7



>>>

Vermeiden grosser Zeitverschiebungen

Vorgehen bei der Dauer- und / oder Frequenzmodifikation:

Zeitliche Verschiebung beschränken! (auf maximal eine halbe Signalperiode)

Achtung: Implementation so konzipieren, dass sich Ungenauigkeiten der F_0-Bestimmung nur lokal auswirken!

Beispiel: Iterative F_0-Phasennachführung bei Dauermodifikation: $D \cdot Z_T$

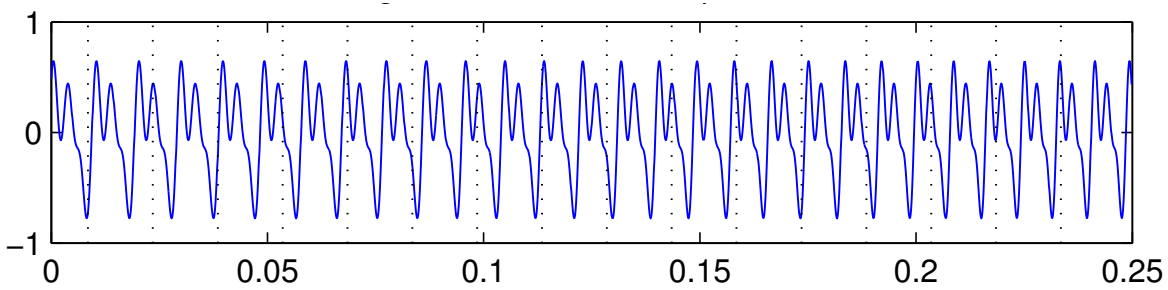
$$\Delta_{\varphi_o}(m+1) = \{\Delta_{\varphi_o}(m) + 2\pi\, F_0(m)\, D \cdot (Z_T - 1)\} \,\widetilde{\text{mod}}\, 2\pi$$

wobei $$a \,\widetilde{\text{mod}}\, b = \begin{cases} a \bmod b & \text{für} \quad a \bmod b < b/2 \\ a \bmod b - b & \text{sonst} \end{cases}$$

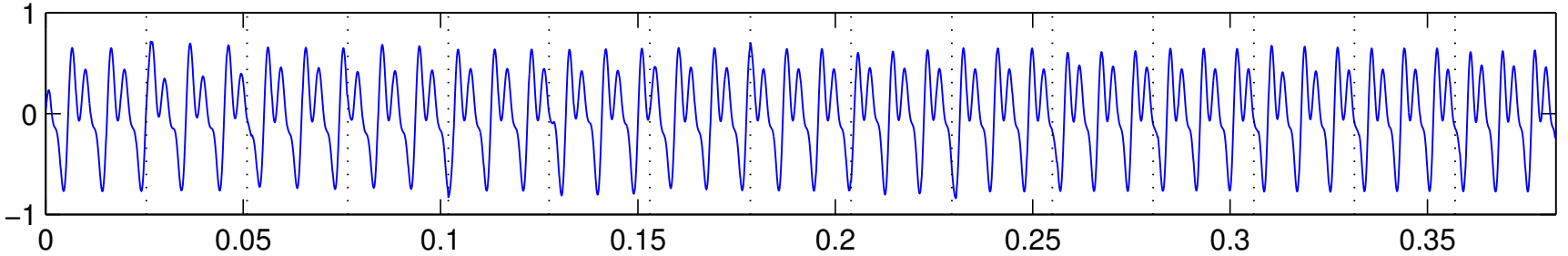
Dauerveränderung

absolute zeitliche Verschiebung < $T_0/2$

Originalsignal: Grundfrequenzverlauf $100\,(1+t)$ Hz ($\approx$ 1 Okt/sec)



Verändertes Signal: Streckungsfaktor 1.7



Frequenzveränderung

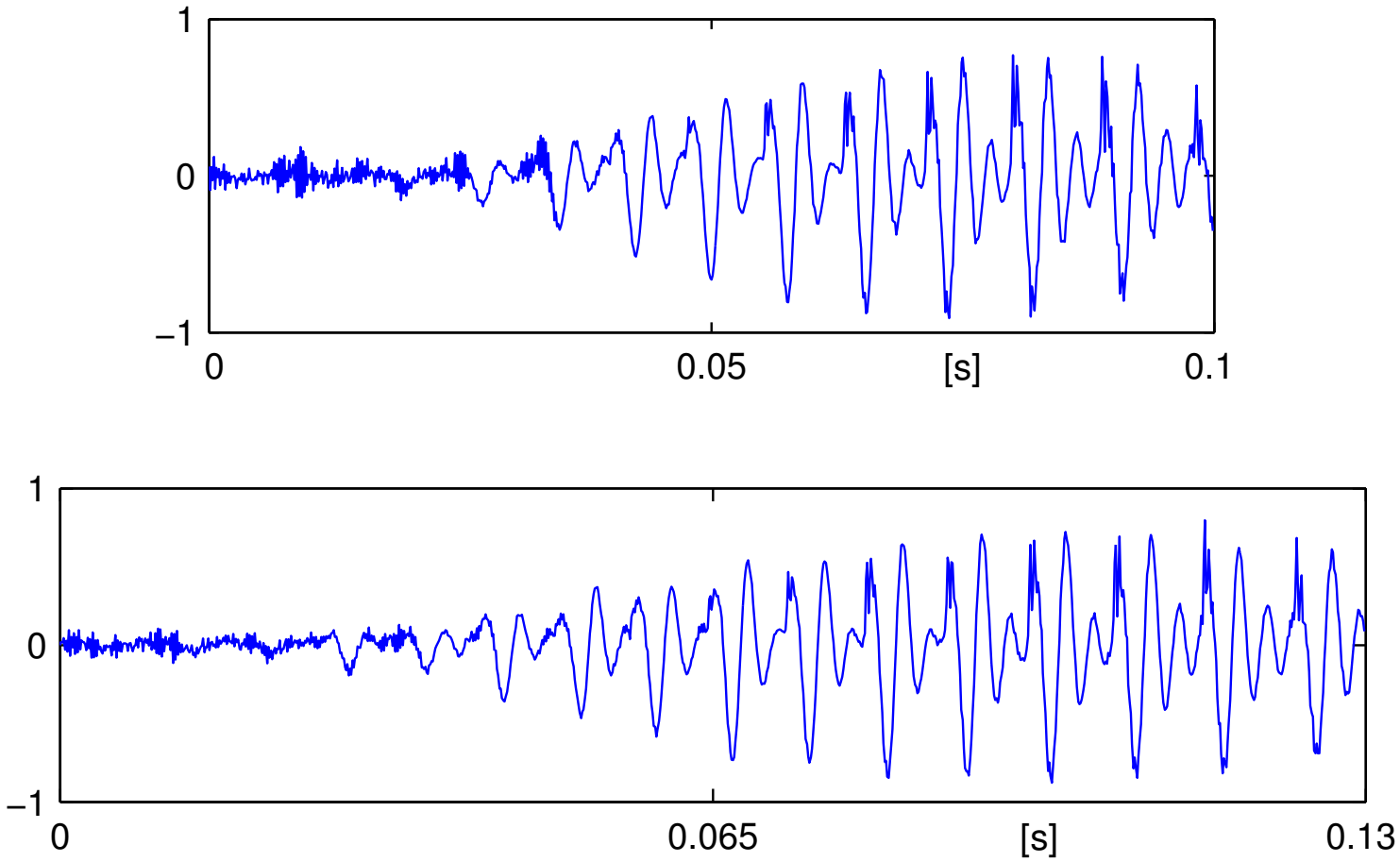
Prinzip: Skalieren der Frequenzkomponenten: $f_i' = Z_F \cdot f_i,$ für alle i

Achtung: Auch hier Kompensation der Phase nötig!
(analog zur Dauerveränderung)

Hörbeispiel: Originalsignal (wie vorher)

Frequenzen skaliert mit Faktor 1.6

Dauerveränderung mittels Fourier-Analyse-Synthese: 130 %

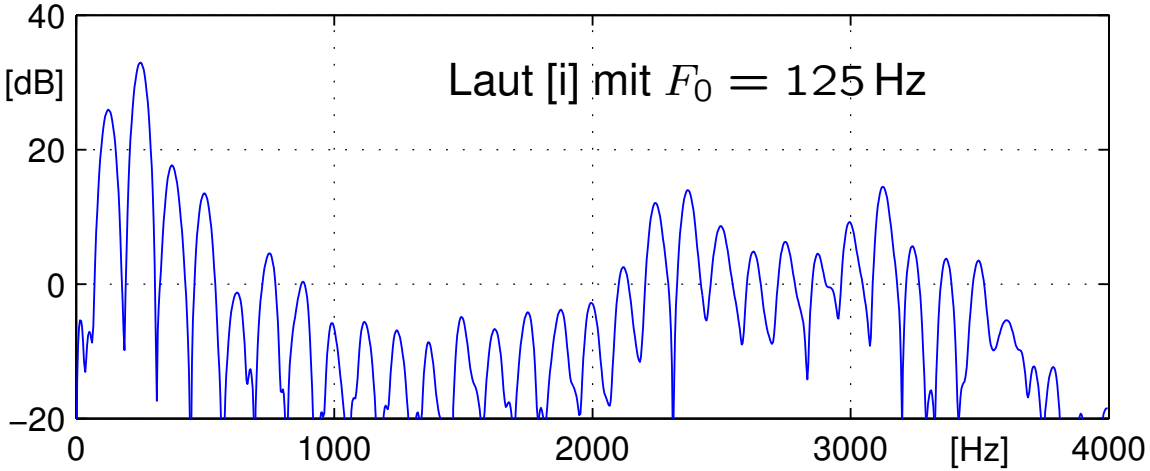


Dauer- und F_0-Veränderung bei Sprachsignalen

Hörbeispiele: Sprache (4 kHz Bandbreite)

- Originalsignal: 
- Dauerfaktor: 1.3 
- Frequenzfaktor: 0.7 
- 1.2 

Verändern der Tonhöhe in Sprachsignalen

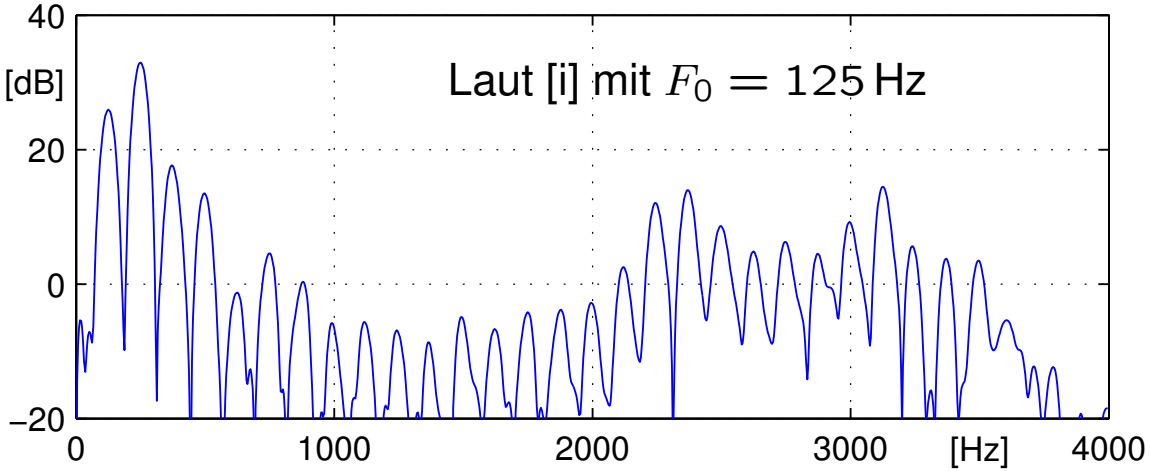


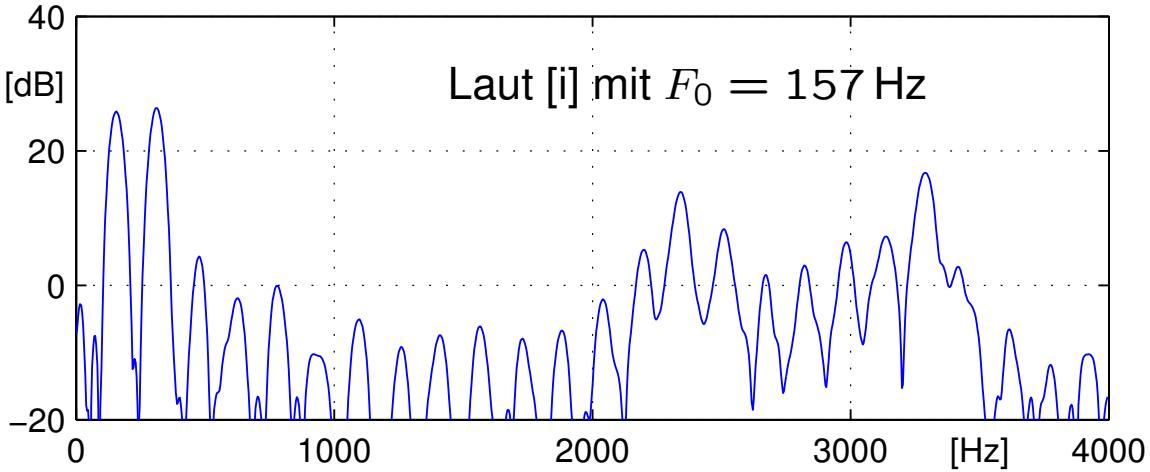
Tonhöhe nimmt zu

$\longrightarrow$ Abstand der Frequenzlinien nimmt zu

Wie wirkt sich das Verändern der Tonhöhe aus?

Verändern der Tonhöhe in Sprachsignalen



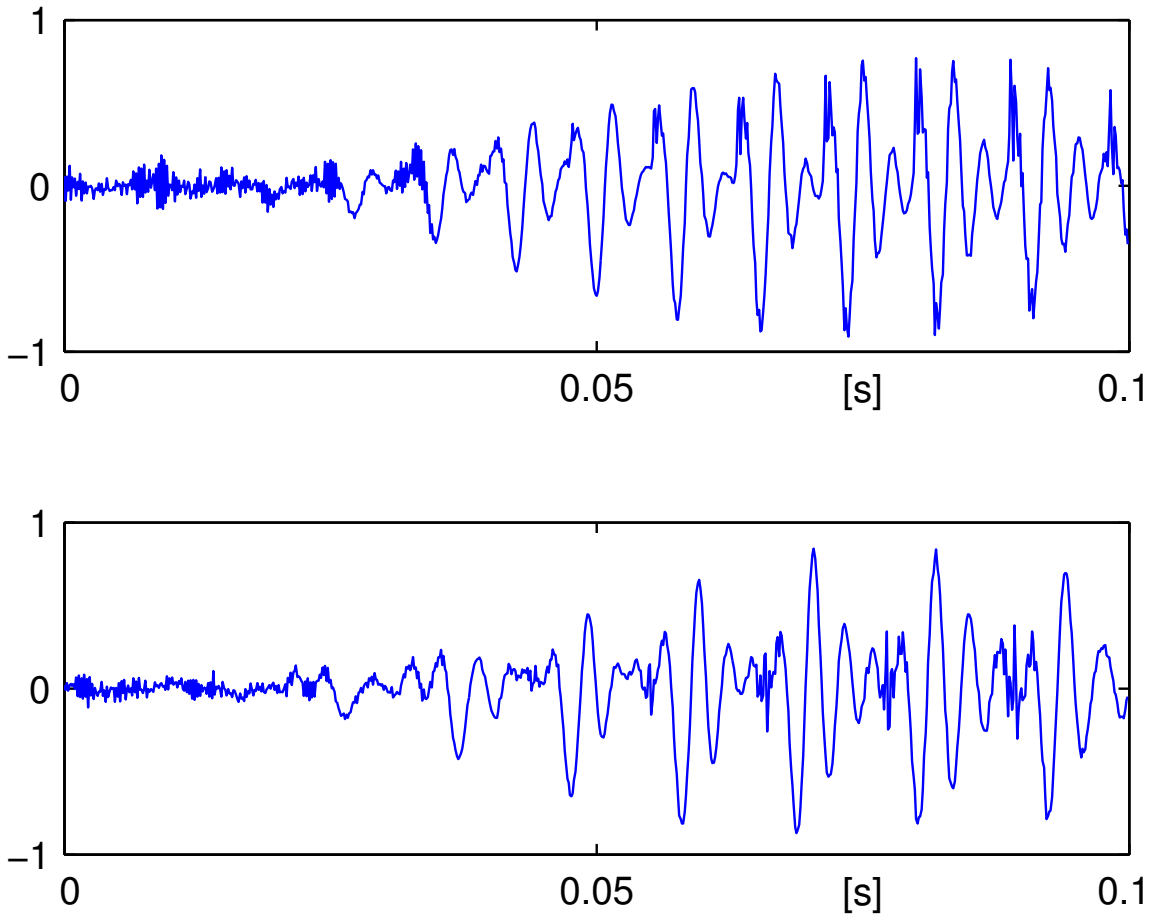


Tonhöhe nimmt zu

$\longrightarrow$ Abstand der Frequenzlinien nimmt zu

Spektrale Enveloppe und Formanten müssen erhalten bleiben ! >>>

Mittels Fourier-Analyse-Synthese F_0 auf 73 % reduziert



Veränderung von Dauer und F_0 mittels Fourier-Analyse-Synthese

Originalsignal (8 kHz Abtastrate)

		Grundfreqenz		
		80%	100%	125%
Dauer	80%			
	100%			
	125%			

Zusammenfassung

Dauer- und Grunfrequenzänderung mittels Fourier-Analyse-Synthese:

- Genaue Schätzung der spektralen Zusammensetzung
  - Abschnitte mit grosser F_0-Änderung (Stationarisierung)
  - Signalkomponenten mit rel. geringer Leistung (mehrfach Analyse)
- Frequenzveränderung:
  - Skalieren der Frequenz der Komponenten
  - Beibehalten der spektralen Enveloppe
- Dauerveränderung und Rekonstruktion:
  - Zeitverschiebung auf $T_0/2$ beschränken
  - Überblenden der Synthese-Abschnitte (windowed overlap-add)

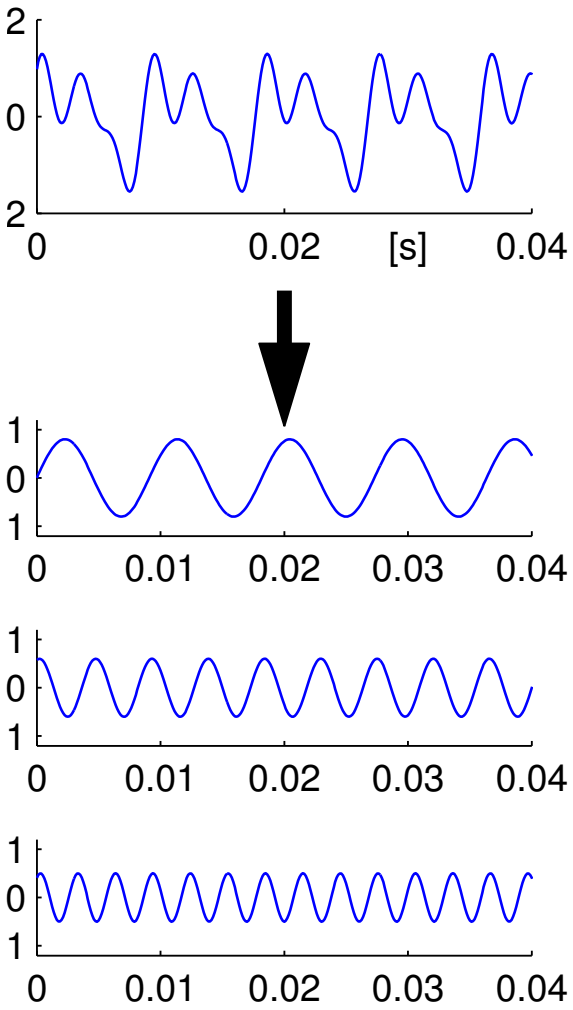
Thema der nächsten Lektion

Spracherkennung

Zur Übersicht der Vorlesung *Sprachverarbeitung I* >>>

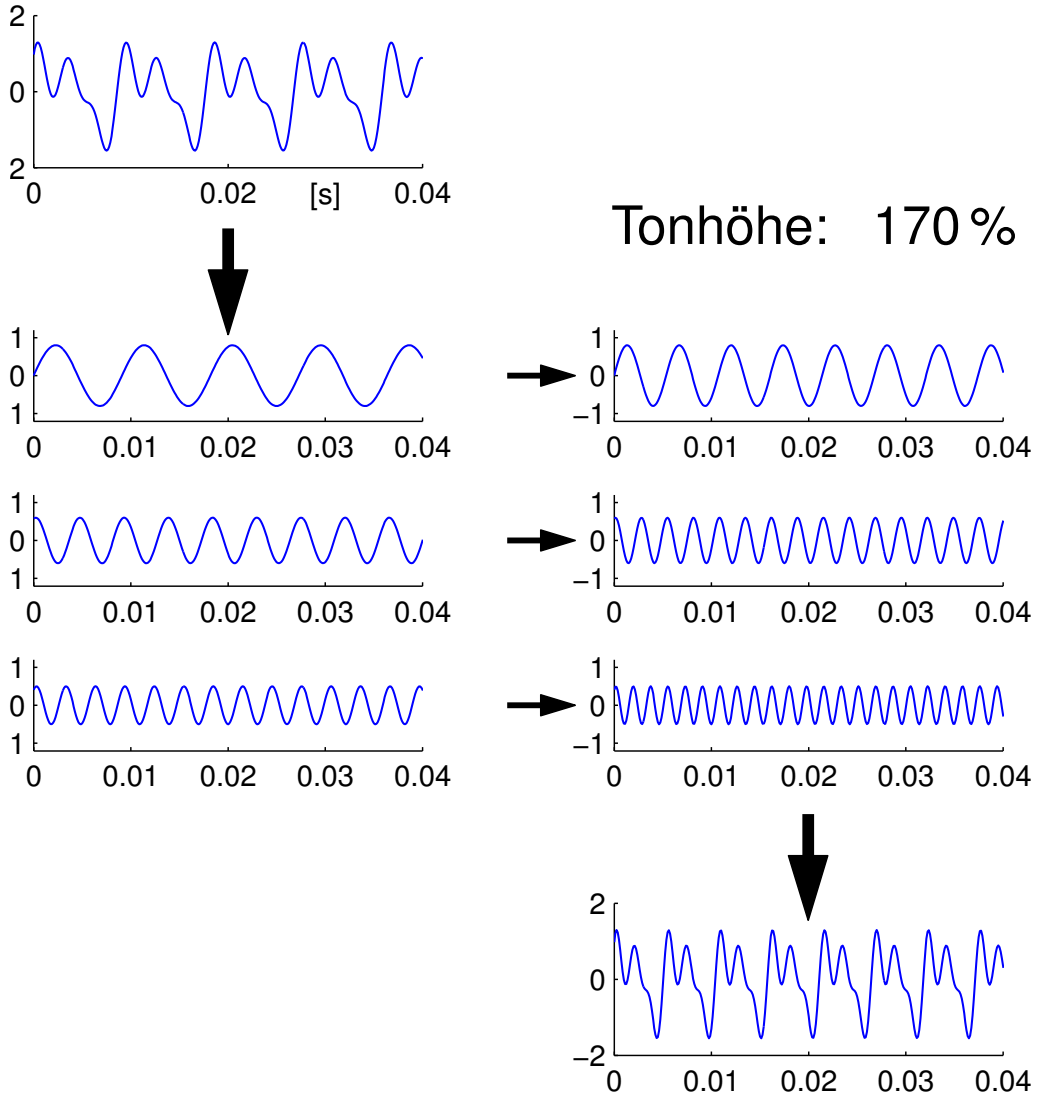
Fourier-Analyse

Zerlegung in Sinuskomponenten



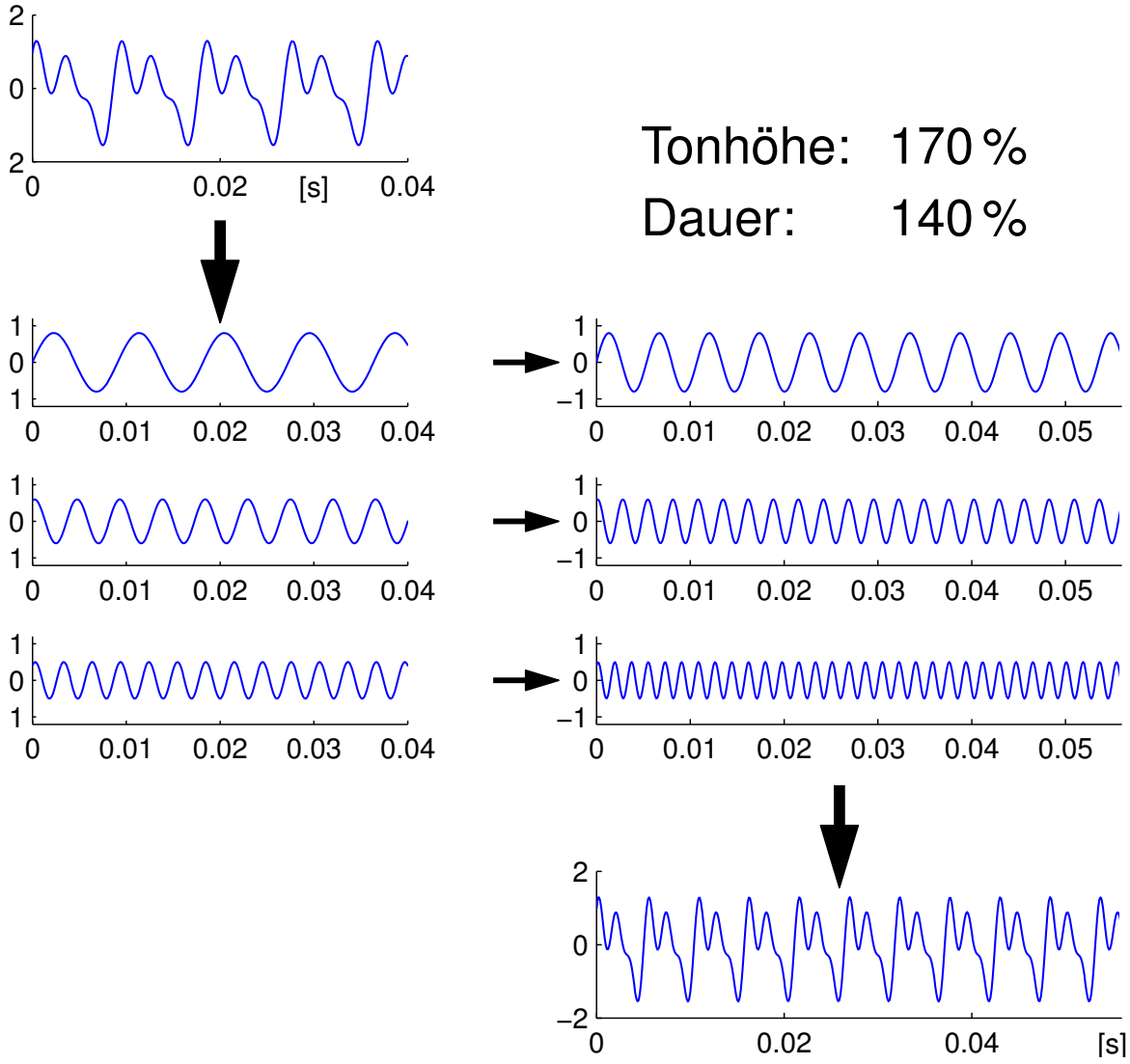
Fourier-Synthese

Superposition von
frequenzveränderten
Sinuskomponenten



Fourier-Synthese

Superposition von
frequenzveränderten und
verlängerten / verkürzten
Sinuskomponenten



Fourier-Analyse-Synthese

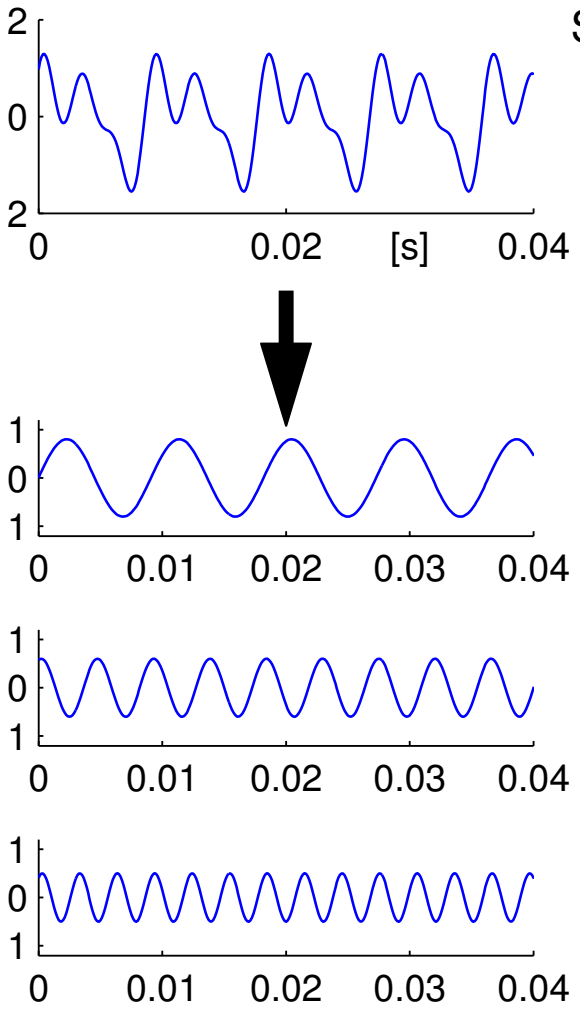
Tonhöhen- und Daueränderung eines Signals $x(n)$:

1. Zerlegen in Sinuskomponenten $\longrightarrow Y_i = \{f_i, a_i, p_i\}$, mit $i = 1, \ldots, I$
   (Ermitteln der spektralen Zusammensetzung)
2. Frequenzänderung $\longrightarrow f_i' = z_f \cdot f_i$, für alle i
3. Daueränderung $\longrightarrow d' = z_d \cdot d$
   (Generieren von Sinuskomponenten der Dauer d')
4. Superposition der generierten Sinuskomponenten
   $\longrightarrow x'(n)$

<<<

Fourier-Analyse

Zerlegung in Sinuskomponenten

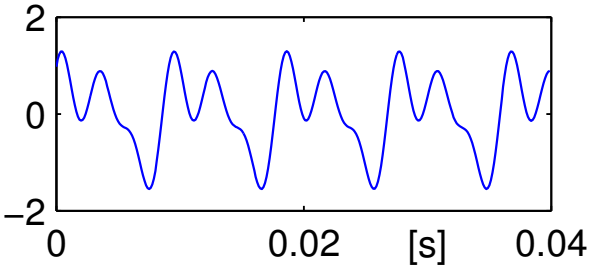


Signal mit 3 Komponenten
($F_0 = 110\,\text{Hz}$)

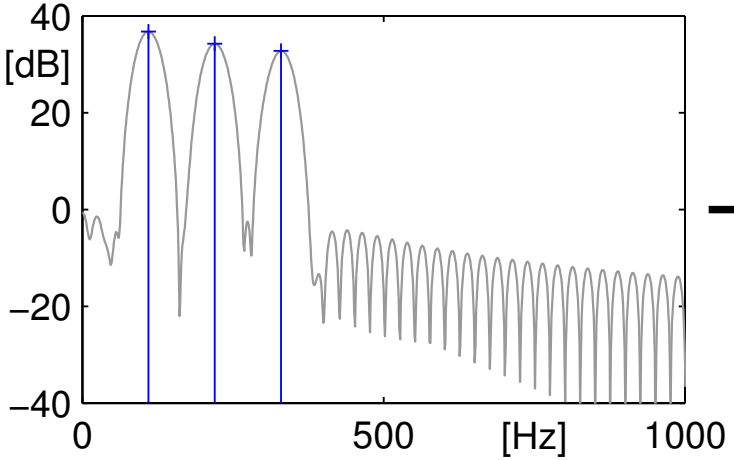
Resultat: 3 Komponenten
(nicht N aus DFT)

<<<

Schätzen der spektralen Zusammensetzung



Signal mit 3 Komponenten
($f_s = 8000\,\text{Hz}$, $F_0 = 110\,\text{Hz}$)

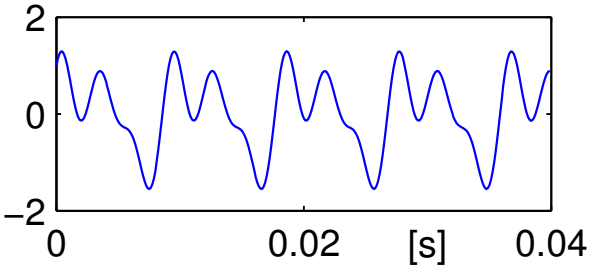


1. Komponente: $f_1 = 110\,\text{Hz}$ / $a_1 = \ldots$ / $p_1 = \ldots$
2. Komponente: $f_2 = 220\,\text{Hz}$ / $a_2 = \ldots$ / $p_2 = \ldots$
3. Komponente: $f_3 = 330\,\text{Hz}$ / $a_3 = \ldots$ / $p_3 = \ldots$

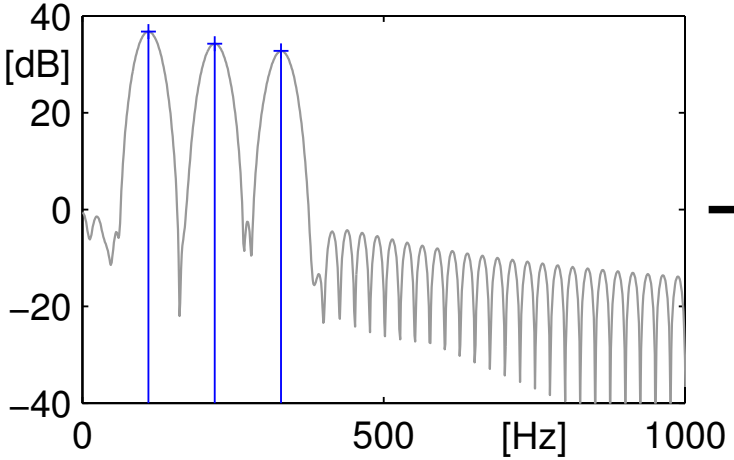
Fensterlänge ? >>>

Auflösung der DFT ? >>>

Schätzen der spektralen Zusammensetzung



Signal mit 3 Komponenten
($f_s = 8000\,\text{Hz}$, $F_0 = 110\,\text{Hz}$)

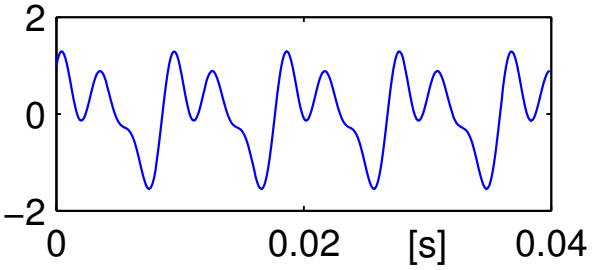


1. Komponente: $f_1 = 110\,\text{Hz}$ / $a_1 = \ldots$ / $p_1 = \ldots$
2. Komponente: $f_2 = 220\,\text{Hz}$ / $a_2 = \ldots$ / $p_2 = \ldots$
3. Komponente: $f_3 = 330\,\text{Hz}$ / $a_3 = \ldots$ / $p_3 = \ldots$

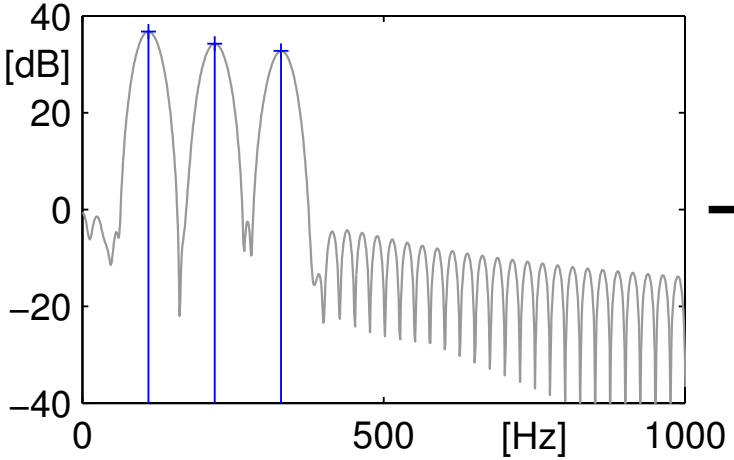
Fensterlänge $\geq 2.5 \cdot T_0$

Auflösung der DFT ? >>>

Schätzen der spektralen Zusammensetzung



Signal mit 3 Komponenten
($f_s = 8000\,\text{Hz}$, $F_0 = 110\,\text{Hz}$)

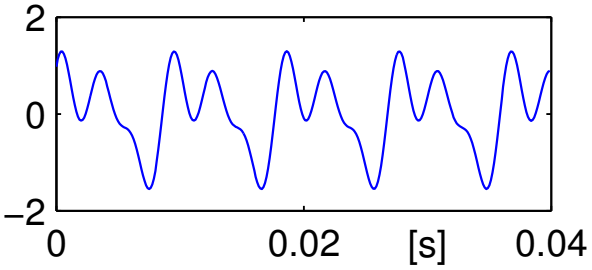


1. Komponente: $f_1 = 110\,\text{Hz}$ / $a_1 = \ldots$ / $p_1 = \ldots$
2. Komponente: $f_2 = 220\,\text{Hz}$ / $a_2 = \ldots$ / $p_2 = \ldots$
3. Komponente: $f_3 = 330\,\text{Hz}$ / $a_3 = \ldots$ / $p_3 = \ldots$

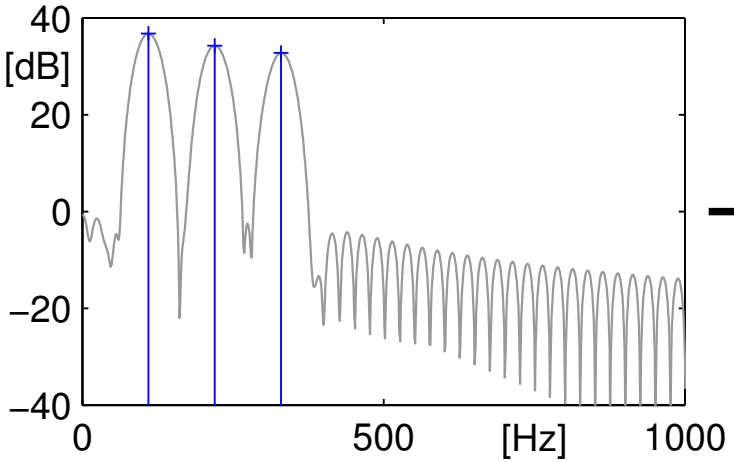
Fensterlänge $\geq 2.5 \cdot T_0$
hochauflösende DFT

<<<

Vergrösserung der Dauer

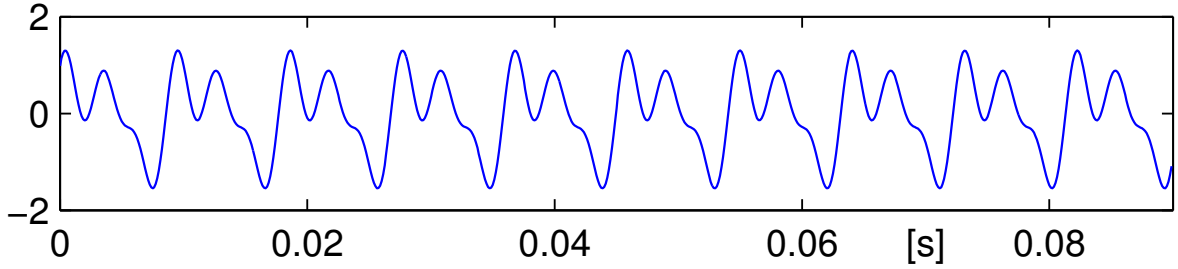


Signal mit 3 Komponenten
($f_s = 8000\,\text{Hz}$, $F_0 = 110\,\text{Hz}$)



hochauflösende DFT !

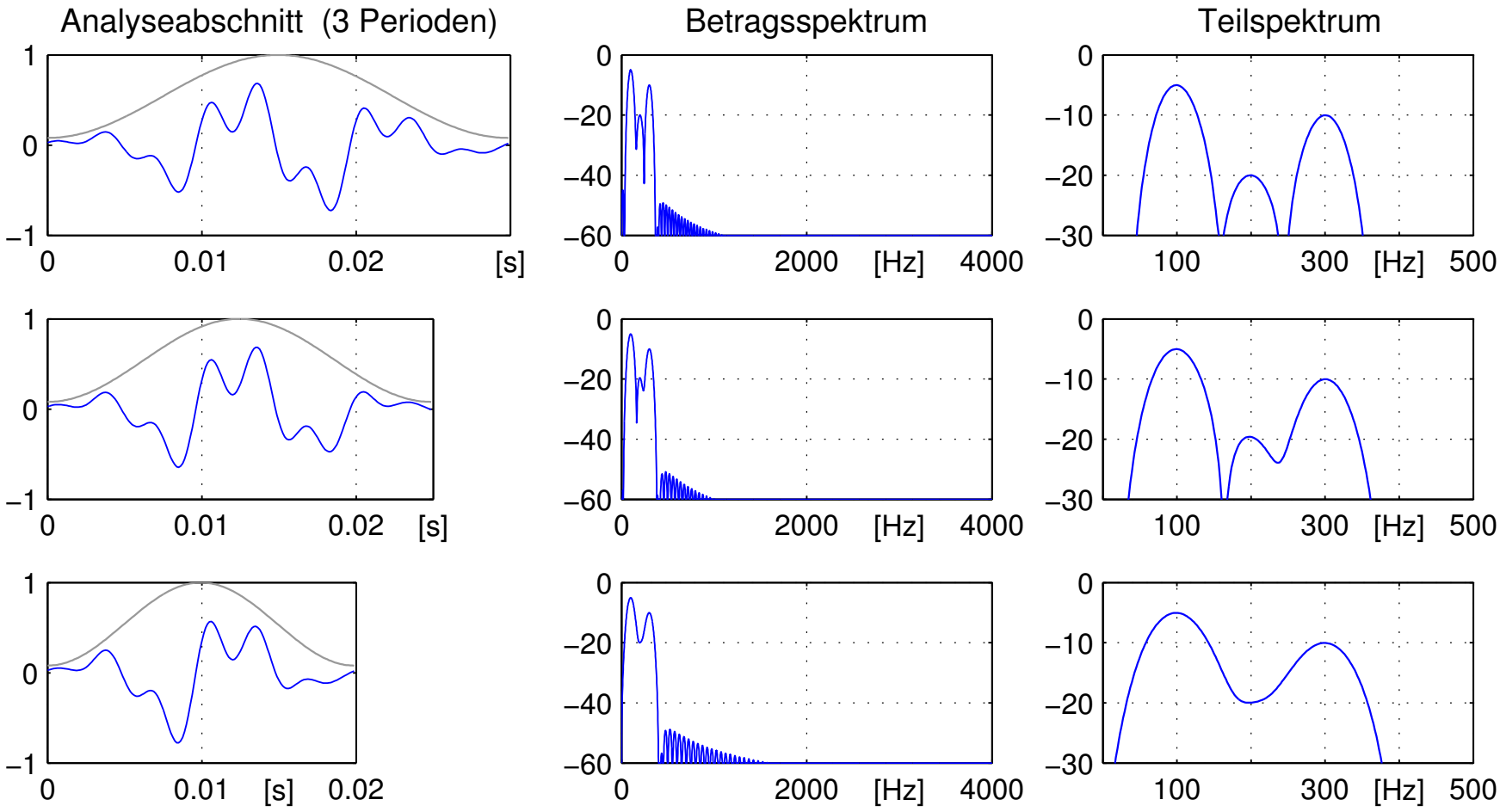
Signal verlängert auf 225 % Dauer



<<<

Einfluss der Länge des Hamming-Fensters

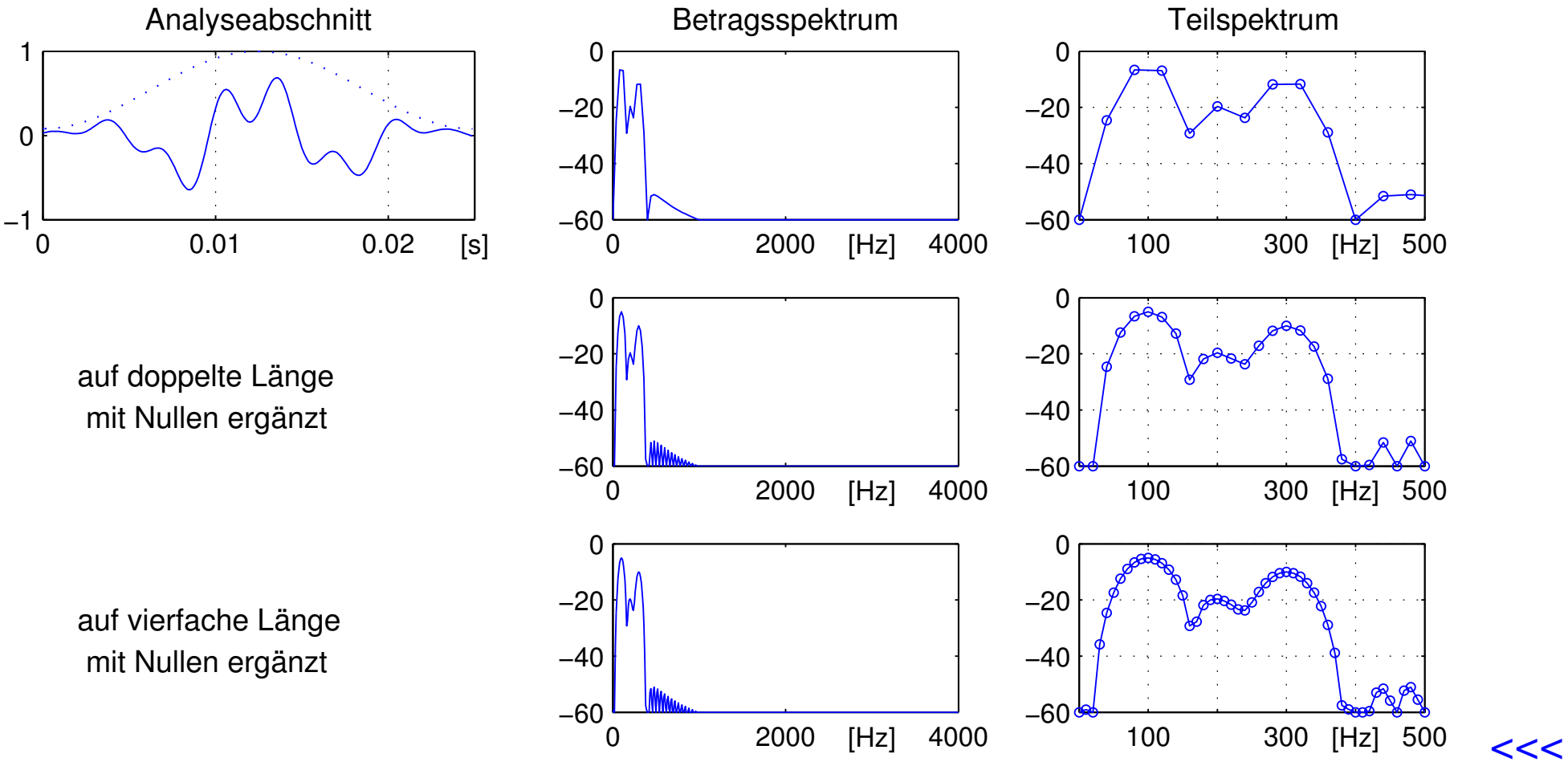
Fourieranalyse eines periodischen Signals ($F_0 = 100$ Hz) mit versch. Fensterlängen



<<<

Frequenz-Auflösung der diskreten Fourier-Transformation

Erhöhung der Genauigkeit durch Ergänzung mit Nullen (*zero padding*)

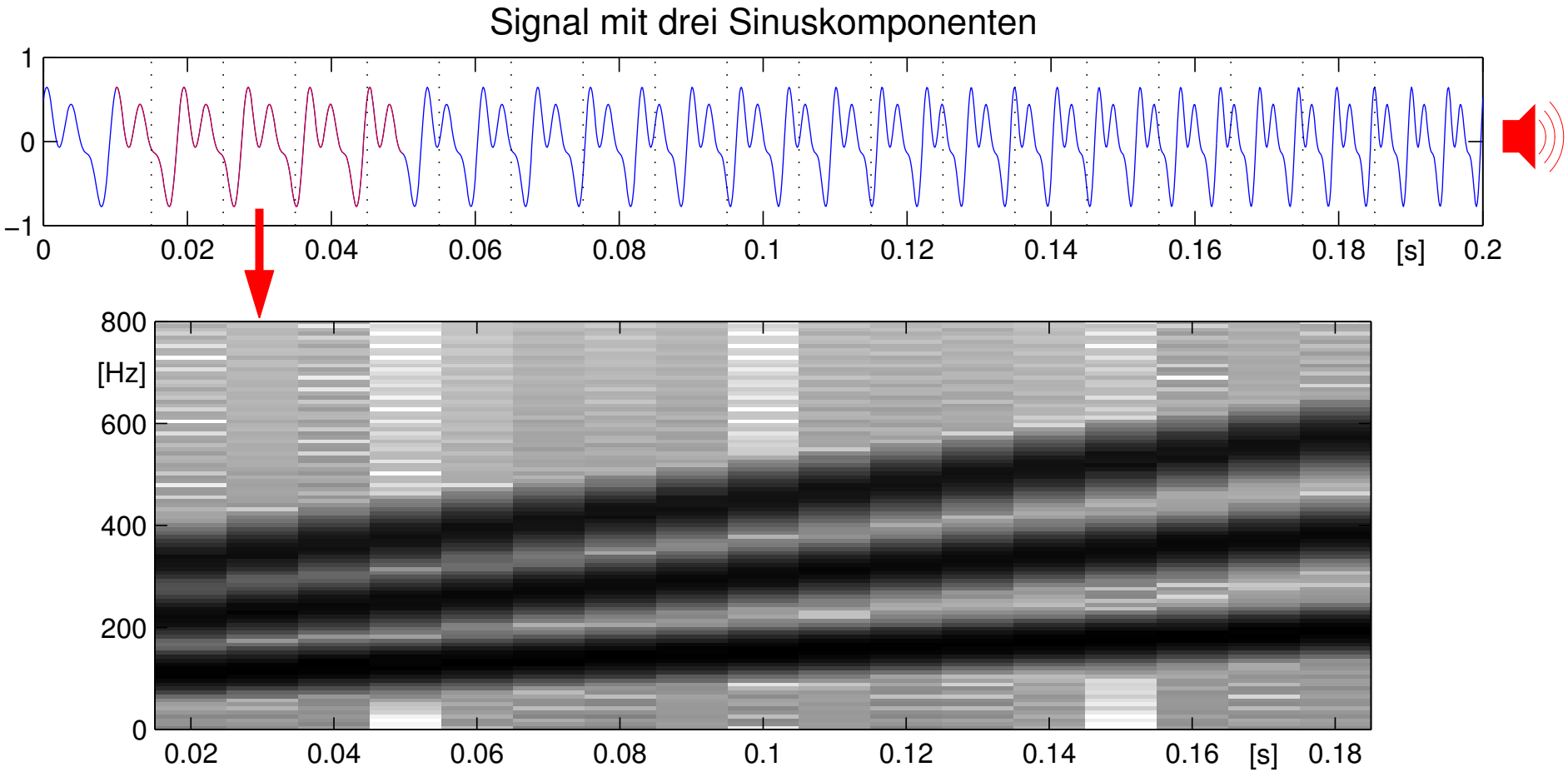


auf doppelte Länge mit Nullen ergänzt

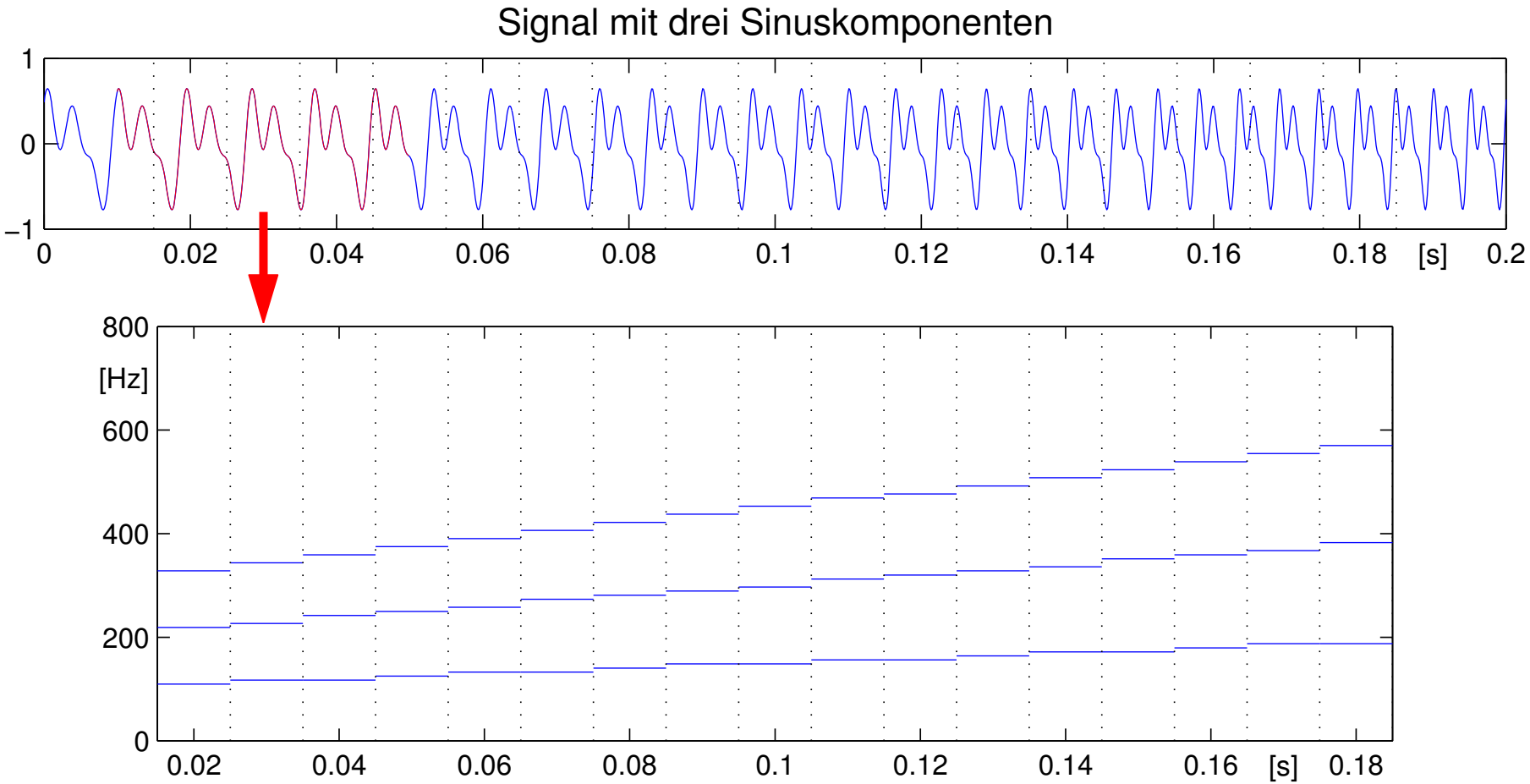
auf vierfache Länge mit Nullen ergänzt

<<<

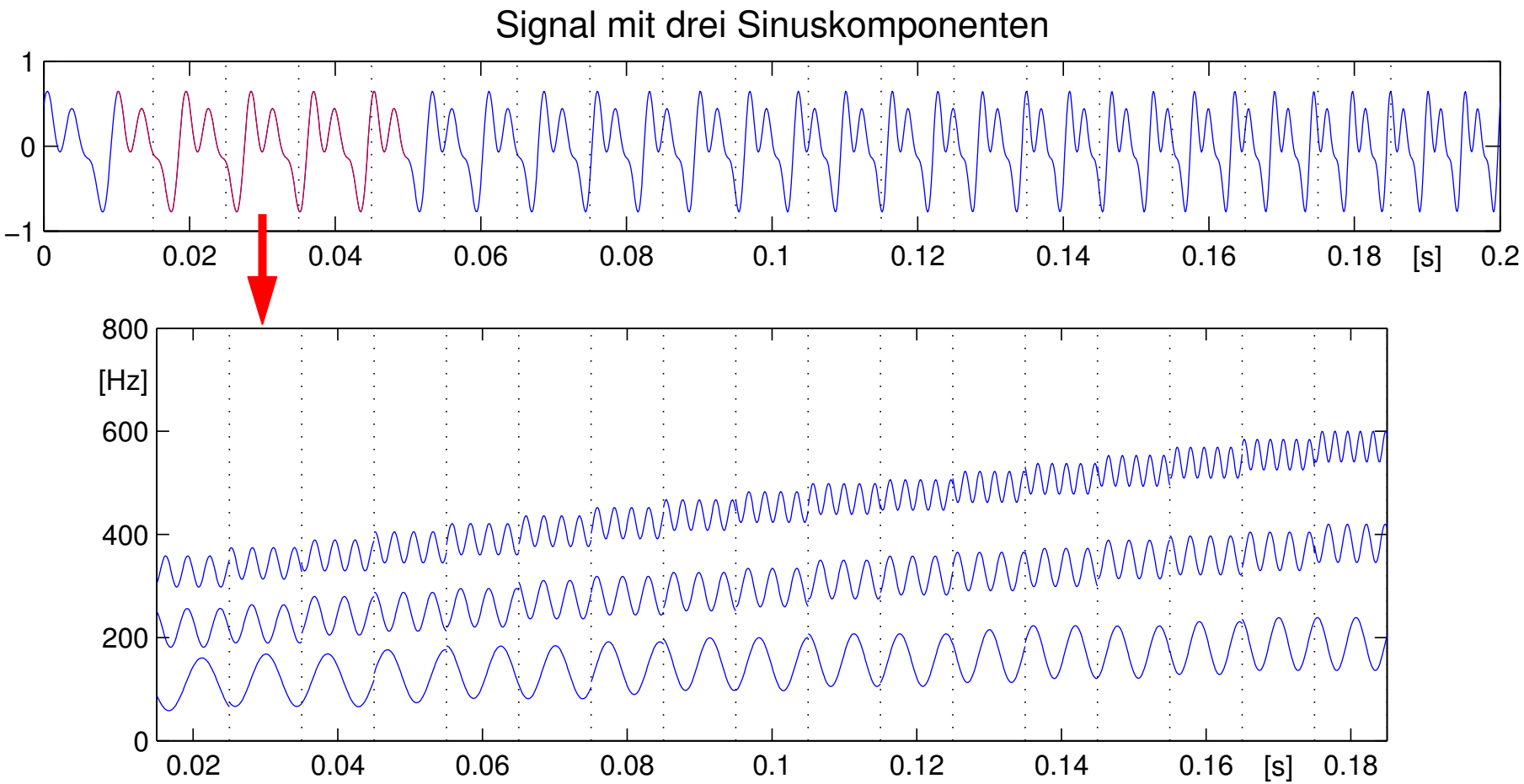
Darstellung der Kurzzeit-Fourier-Transform. instationärer Signale



Darstellung des “wirklichen Spektrums” harmonischer Signale

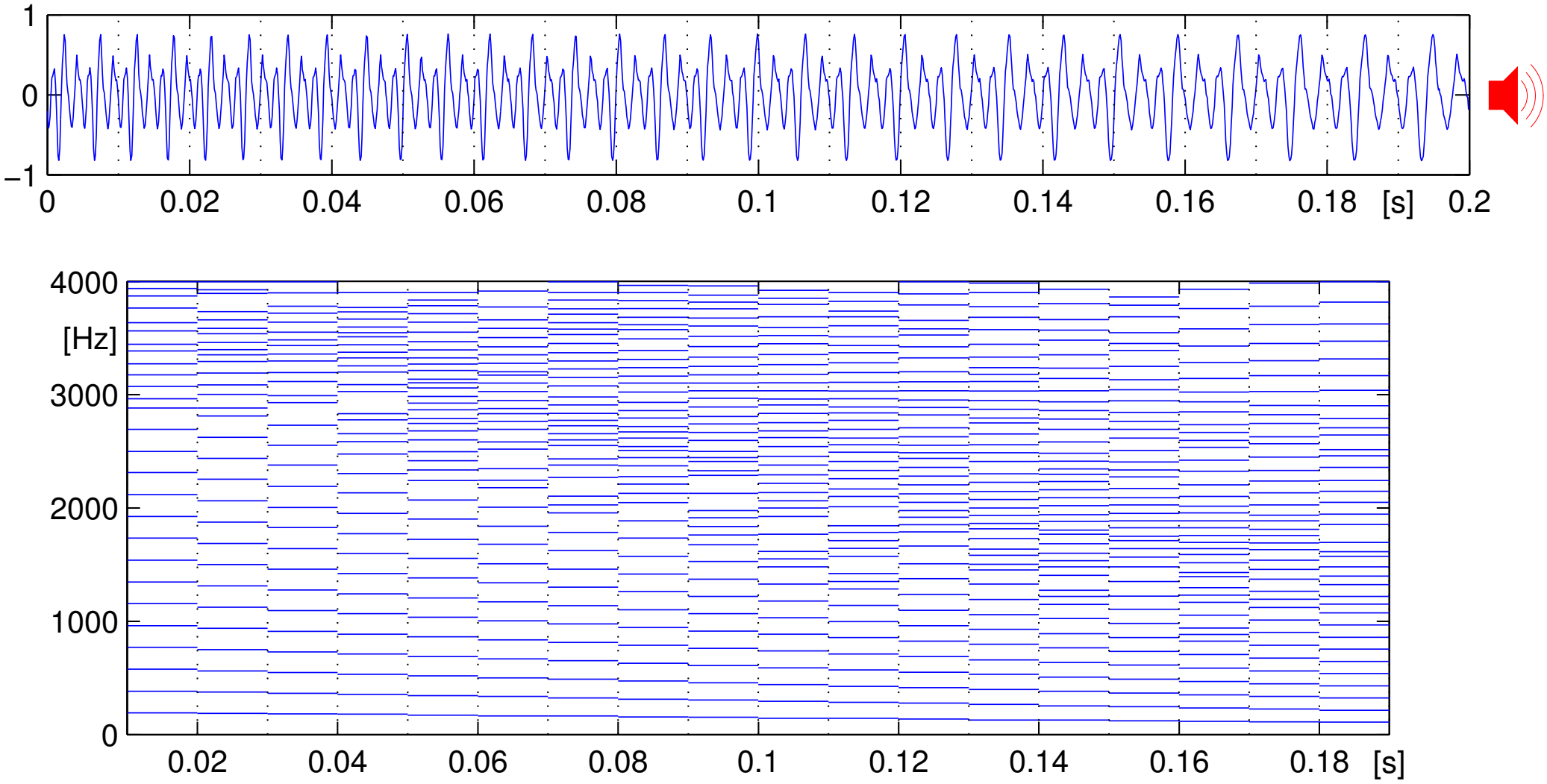


“Wirkliches Spektrum” mit Amplituden und Phaseninformation

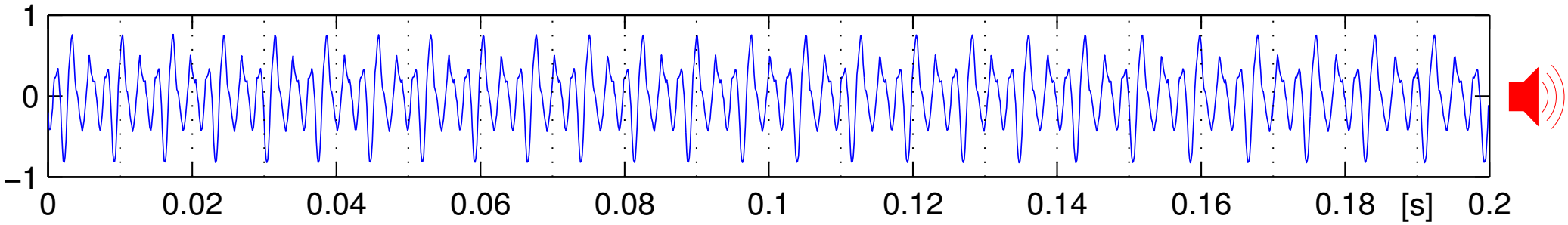


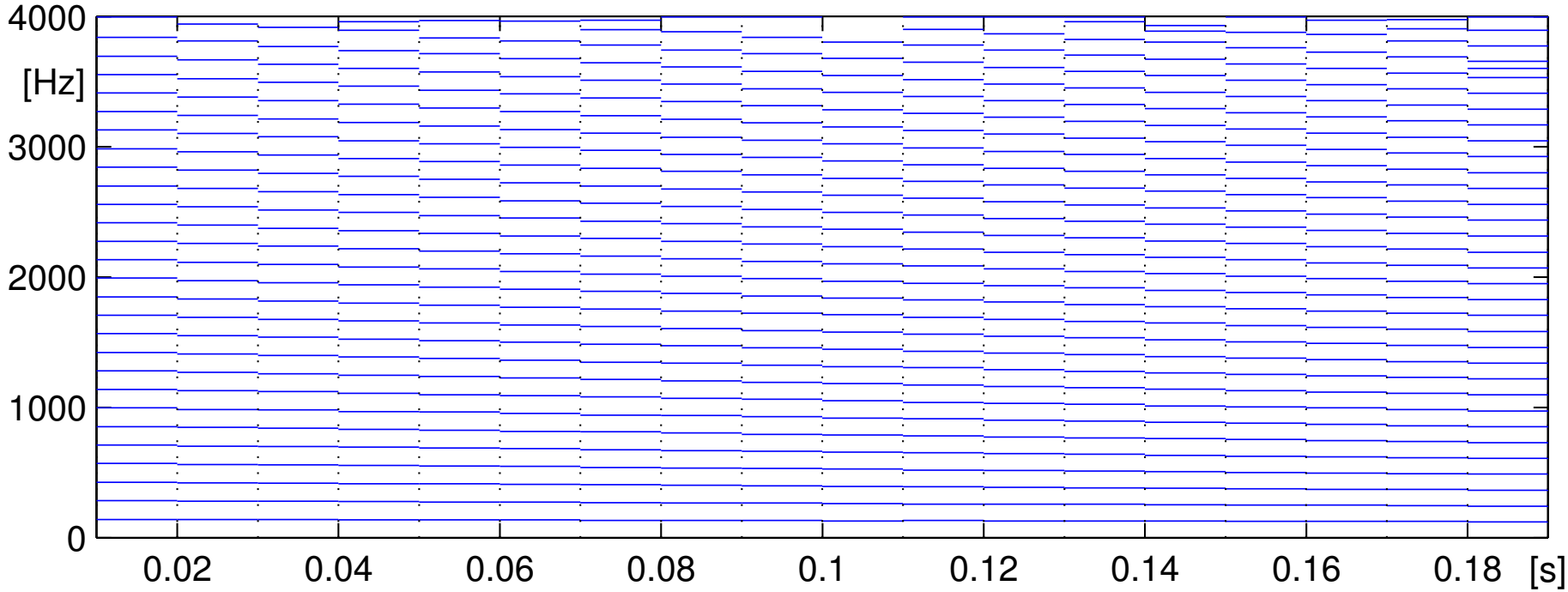
<<<

Veränderung der Grundfrequenz: –5 Okt/s

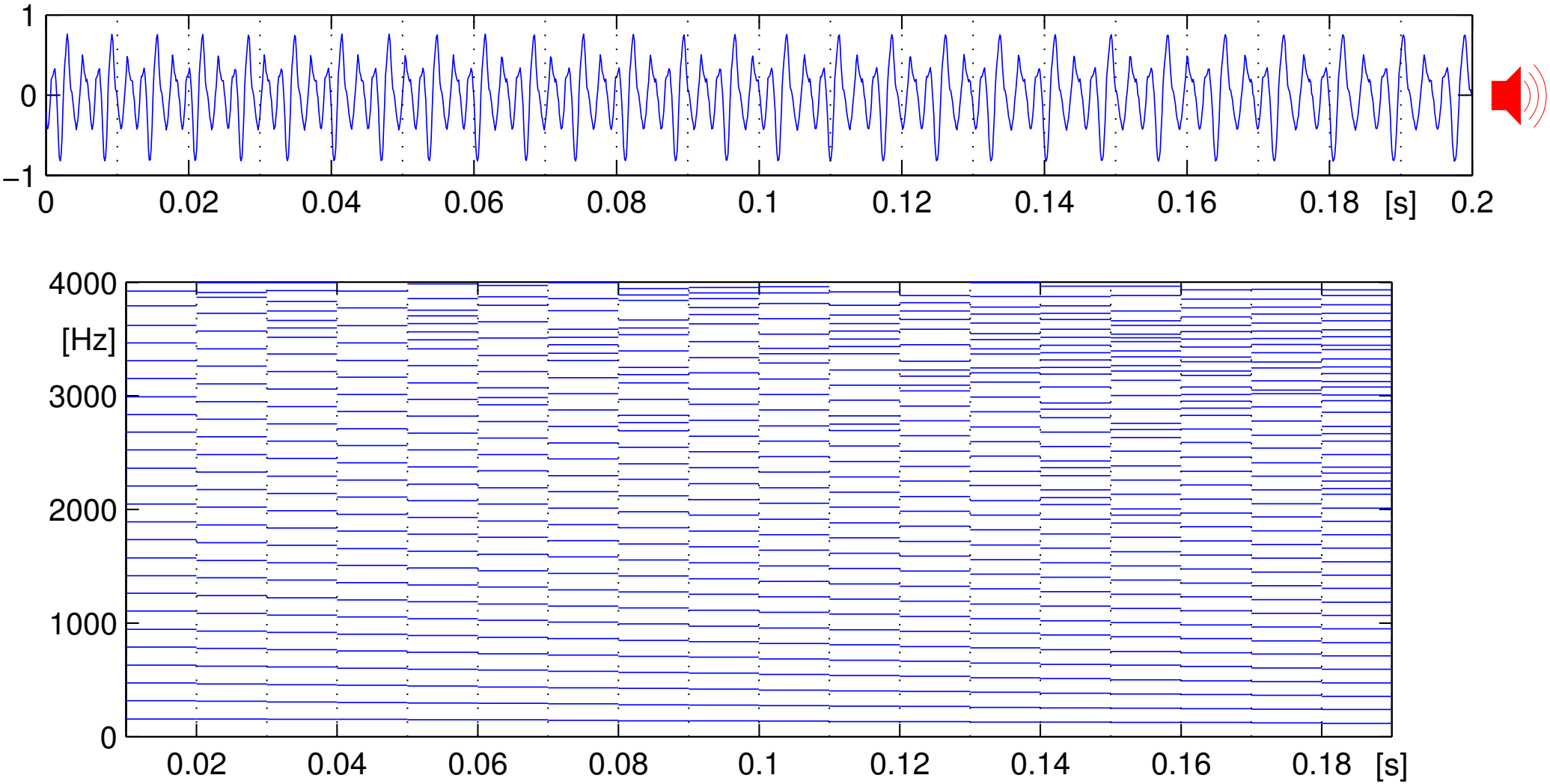


Veränderung der Grundfrequenz: –1 Okt/s

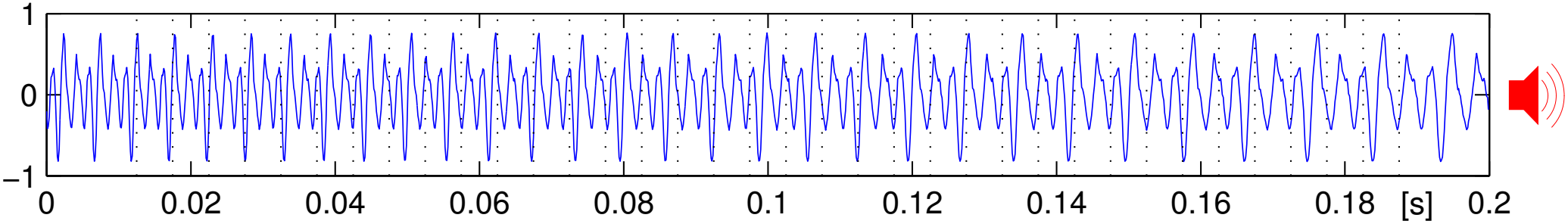


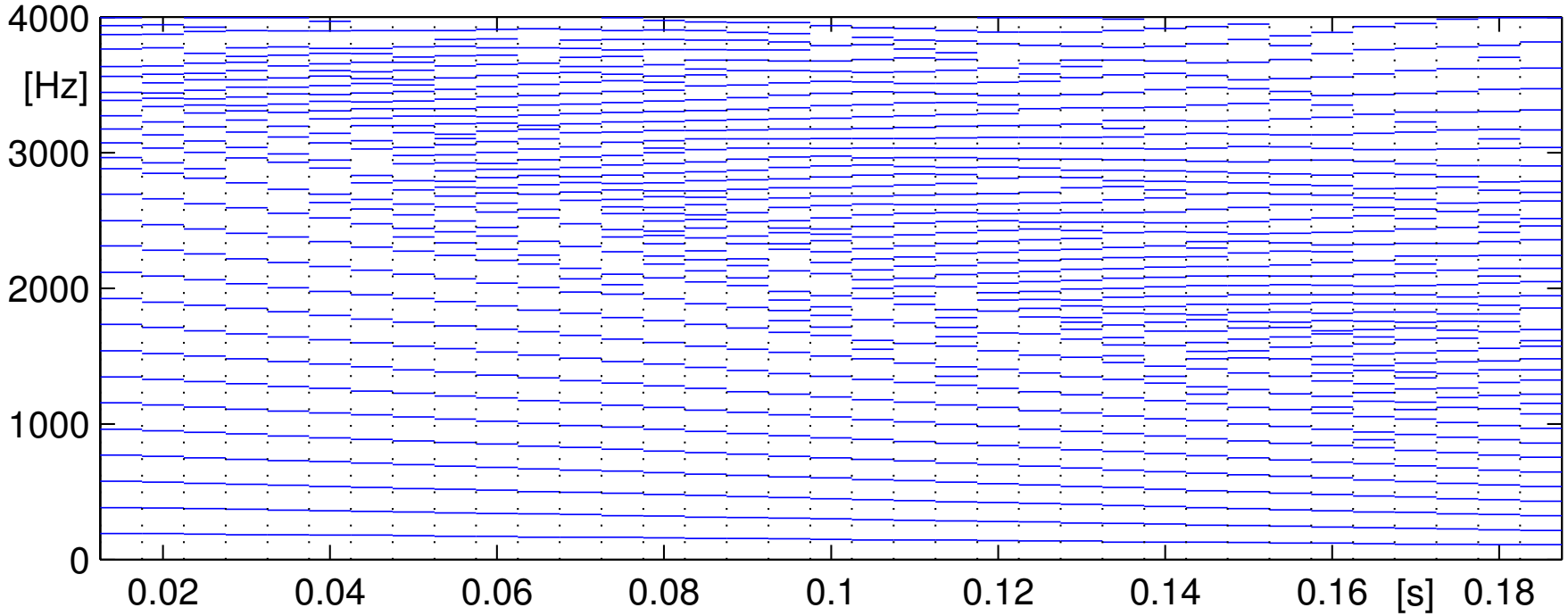


Veränderung der Grundfrequenz: –2 Okt/s

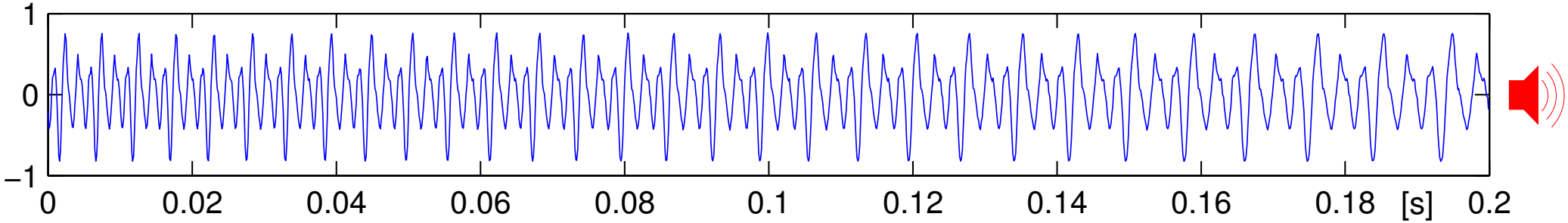


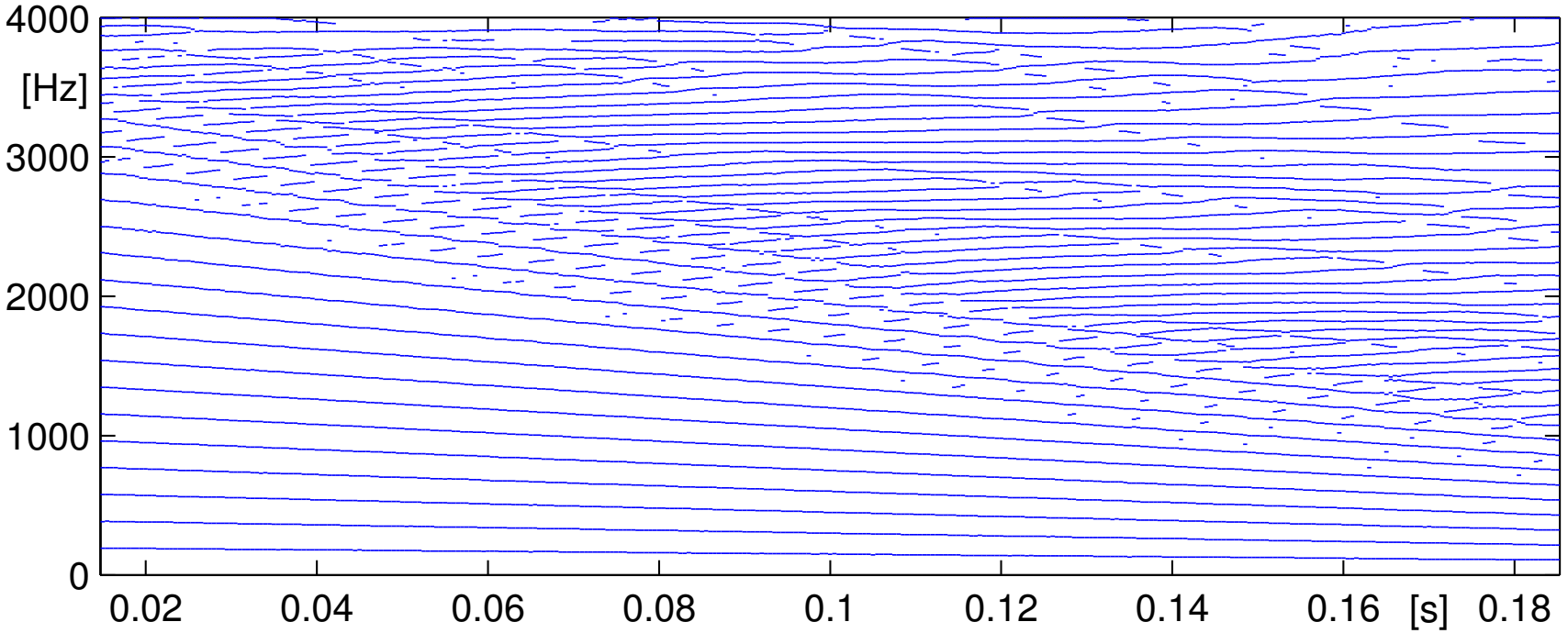
Veränderung der Grundfrequenz: –5 Okt/s





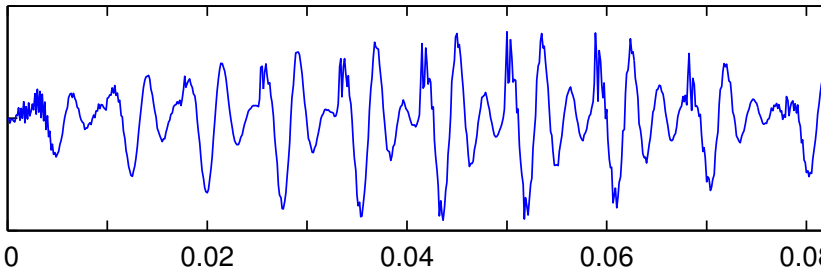
Veränderung der Grundfrequenz: –5 Okt/s





<<<

Spektrum eines Sprachsignals mit fallendem F_0



Sprachsignal:

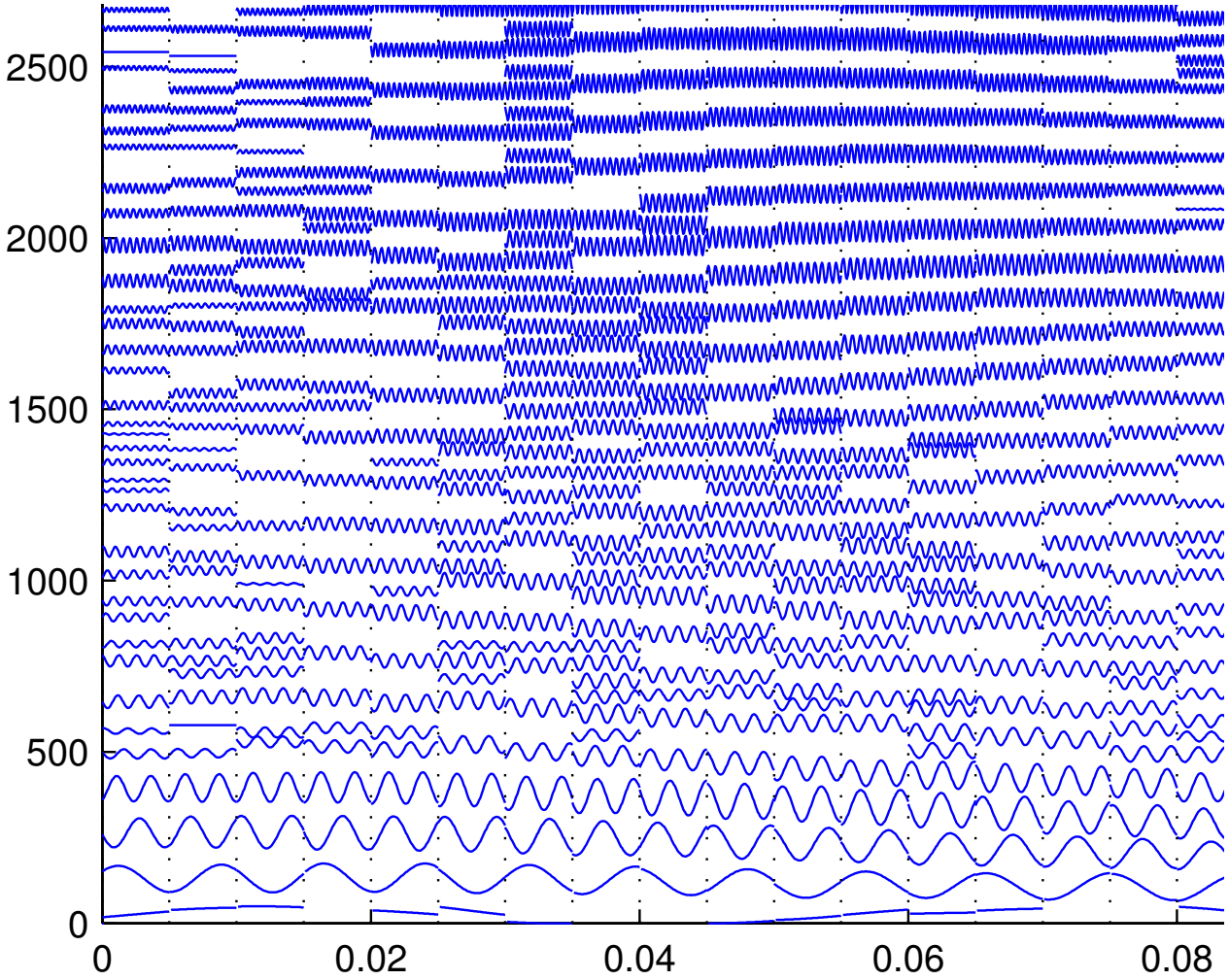
F_0 fällt schnell ($\approx$ 7 Oktaven/s)

Beobachtung:

Spektrum des stimmhaften Signals ist nicht harmonisch!

Grund:

Stationaritätsbedingung der FT ist nicht erfüllt !



<<<

Spektralanalyse für Signale mit schnellen Grundfrequenzveränderungen

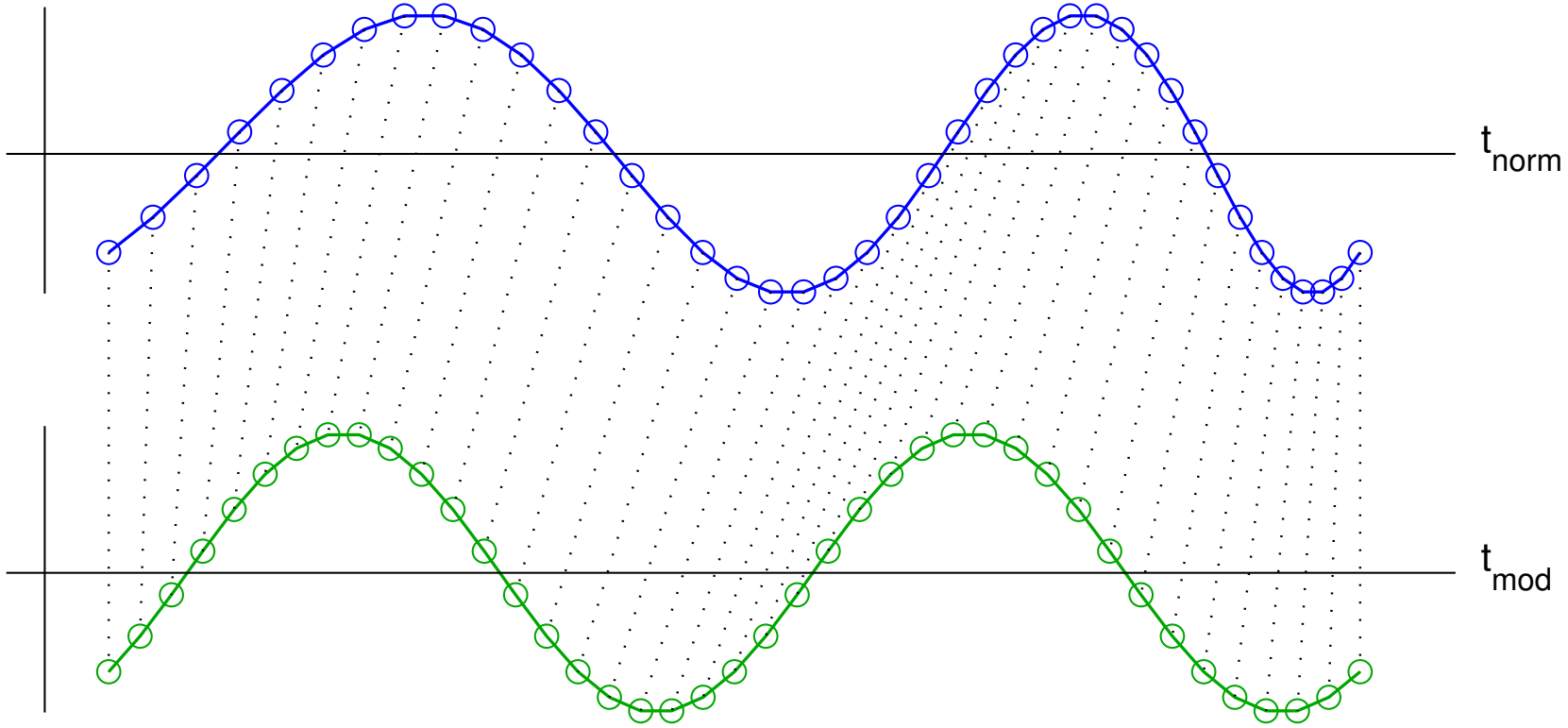
Unter der Annahme, dass für einen kurzen Signalabschnitt (Analysefenster) der Verlauf der Grundfrequenz durch eine Gerade approximierbar ist, also $F_0(t) \approx F_0(0) + \dot{F}_0\, t$ ist, lässt sich das Spektrum des Signals auf zwei Arten bestimmen:

a) Abtastratenkonversion vor der Fouriertransformation mit der variablen Abtastfrequenz: $\dot{f}_s \approx \dot{F}_0$ >>>

b) Fouriertransformation mit einem Satz orthogonaler Funktionen, deren Frequenz gilt: $\dot{f}_k \approx \dot{F}_0$. >>>

<<<

Reduktion der Instationarität durch variables Abtasten

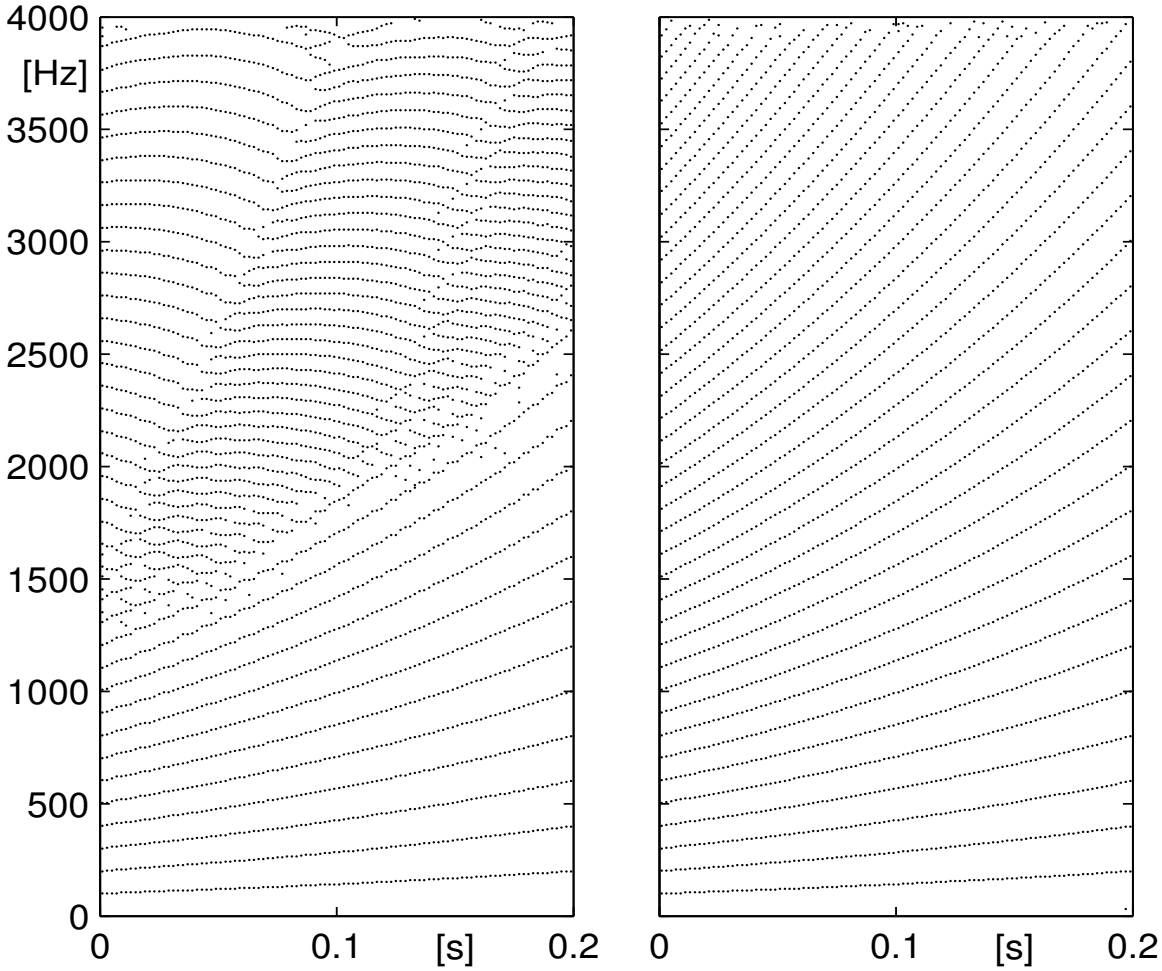


Nötig: Abtastratenwandler für variable Abtastraten

Spektrum eines Signals mit starkem F_0-Anstieg

Analyse eines Signals
mit F_0-Anstieg
von 1 Okt. pro 0.2 s:

- ohne Stationarisierung (links)
- mit Stationarisierung (rechts)



<<<

Fouriertransformation für Signale mit grossem $\dot{F}_0$

Implizite Transformation der Zeitachse in der DFT, indem die Punkte

$$n = \{0, 1, \ldots, N-1\}$$

der diskreten Zeitachse durch die Punkte

$$\tilde{n} = n + \varphi(n)$$

der modifizierten Zeitachse ersetzt werden. Die DFT ist dann zu schreiben als:

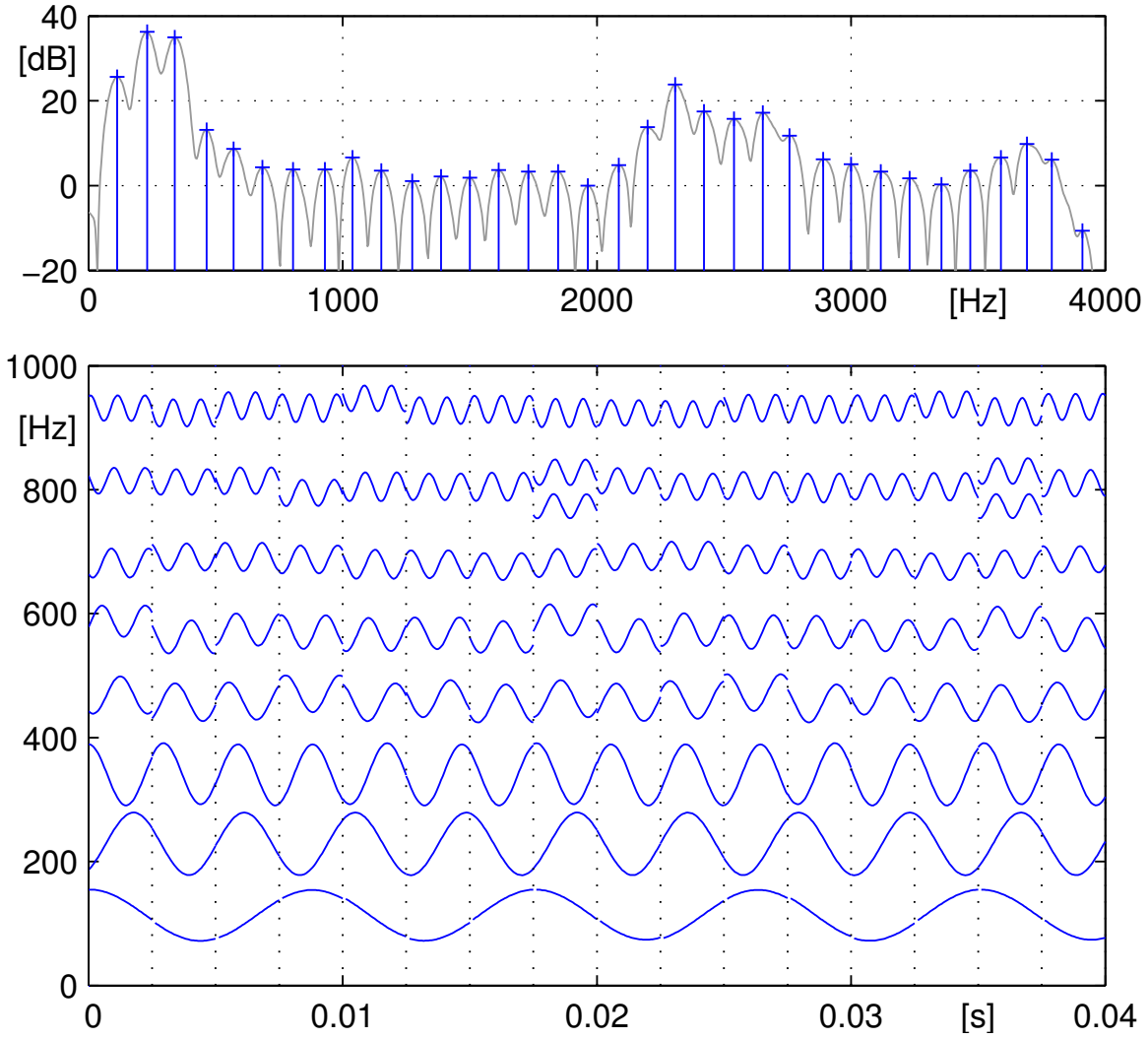
$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi k(n+\varphi(n))/N}$$

Achtung:
- Aliasing-Effekte nicht vermeidbar
- Term $\varphi(n)$ verhindert Anwendung der FFT

<<<

Analyse schwacher spektraler Komponenten

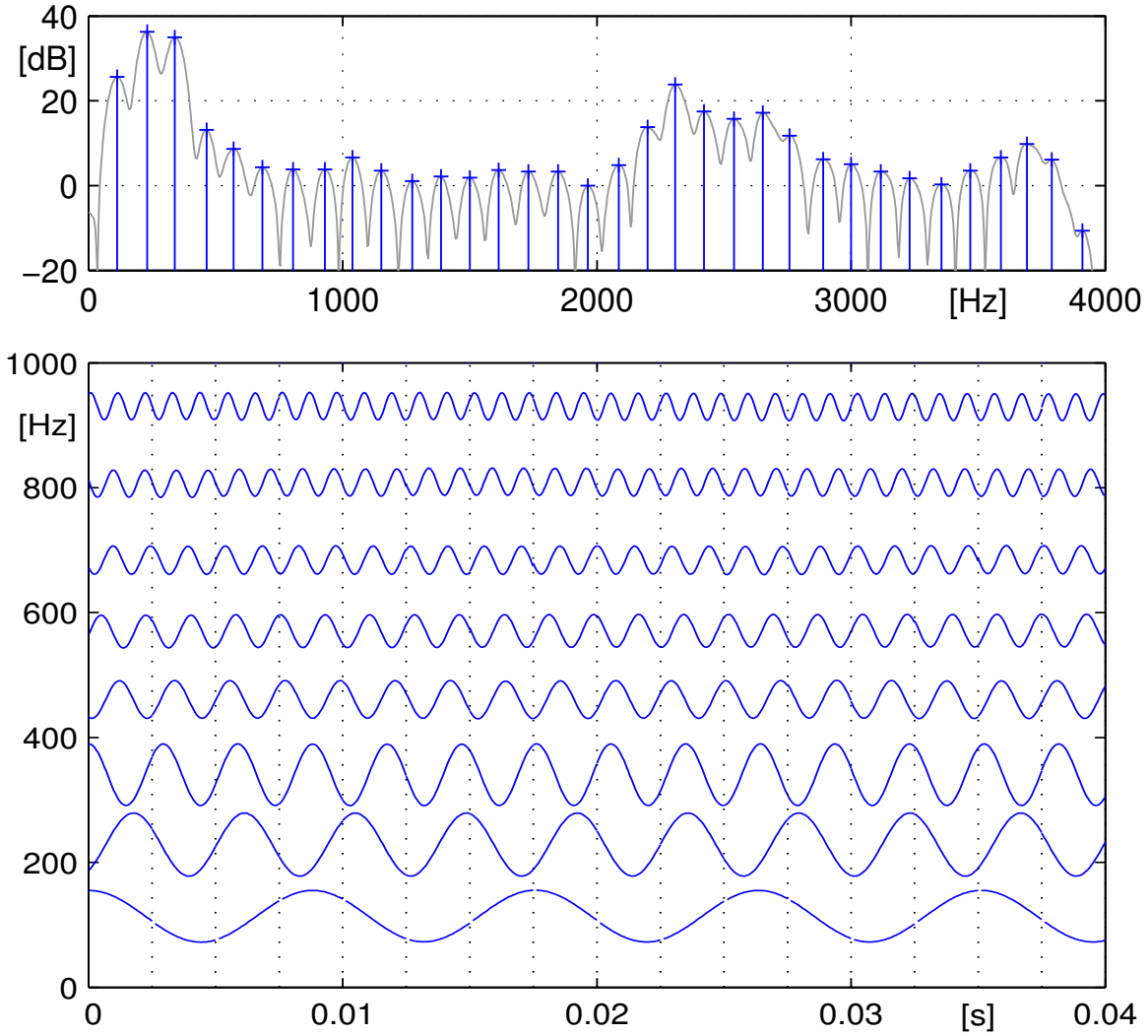
Fouriertransformation liefert für schwache Frequenzkomponenten eines Sprachsignals keine zuverlässigen Werte



Analyse schwacher spektraler Komponenten

Fouriertransformation liefert für schwache Frequenzkomponenten eines Sprachsignals keine zuverlässigen Werte

→ Analyse mehrerer, zeitlich leicht verschobener Fenster und Nachverarbeitung



<<<

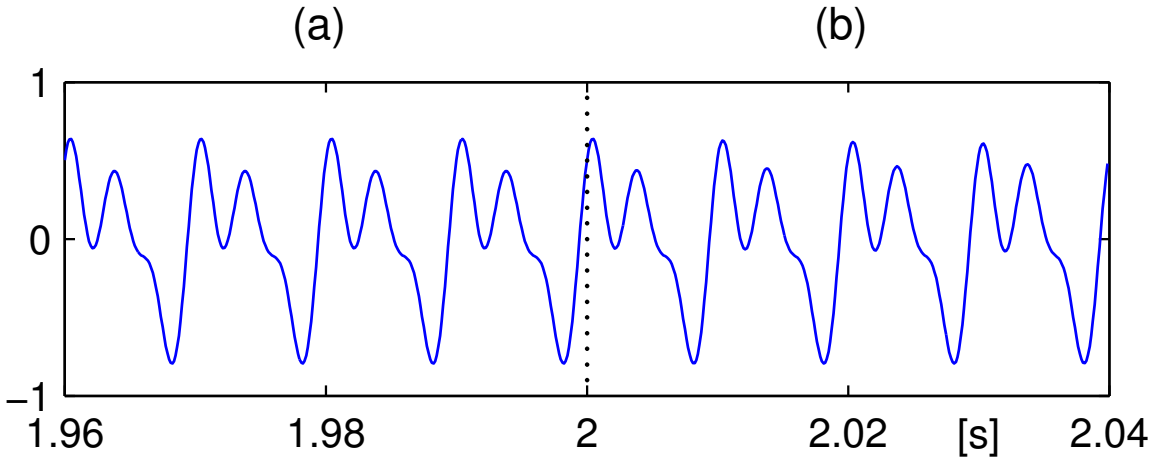
Auswirkung der akkumulierten Zeitverschiebung

Frequenzkomp. in Abschnitt (a):

100, 200, 300 Hz

Frequenzkomp. in Abschnitt (b):

100, 201, 301 Hz



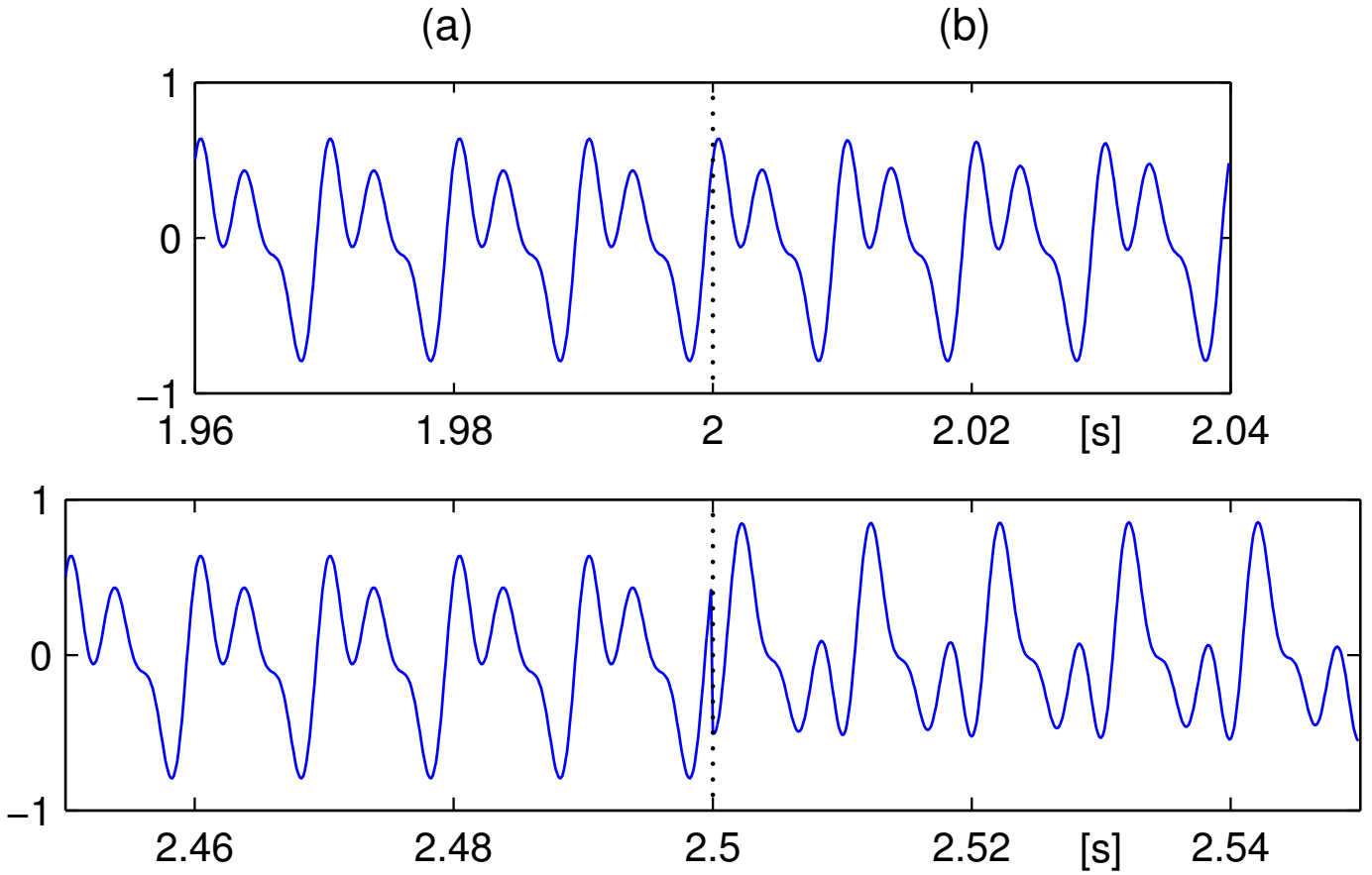
Auswirkung der akkumulierten Zeitverschiebung

Frequenzkomp. in Abschnitt (a):

100, 200, 300 Hz

Frequenzkomp. in Abschnitt (b):

100, 201, 301 Hz



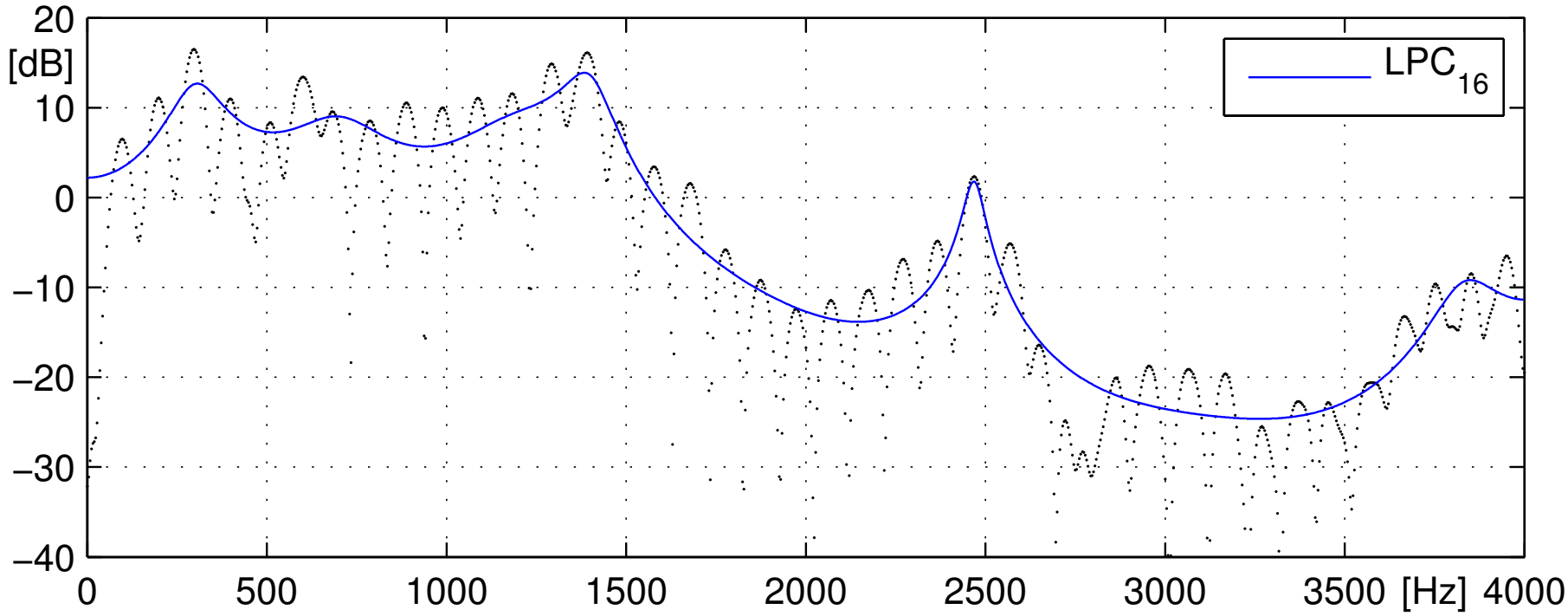
zeitliche Streckung um 25 %

$\longrightarrow$ akkum. Zeitverschieb.: 0.5 s

Anmerkung: Ist nur bei periodischen Signalabschnitten ein Problem !

<<<

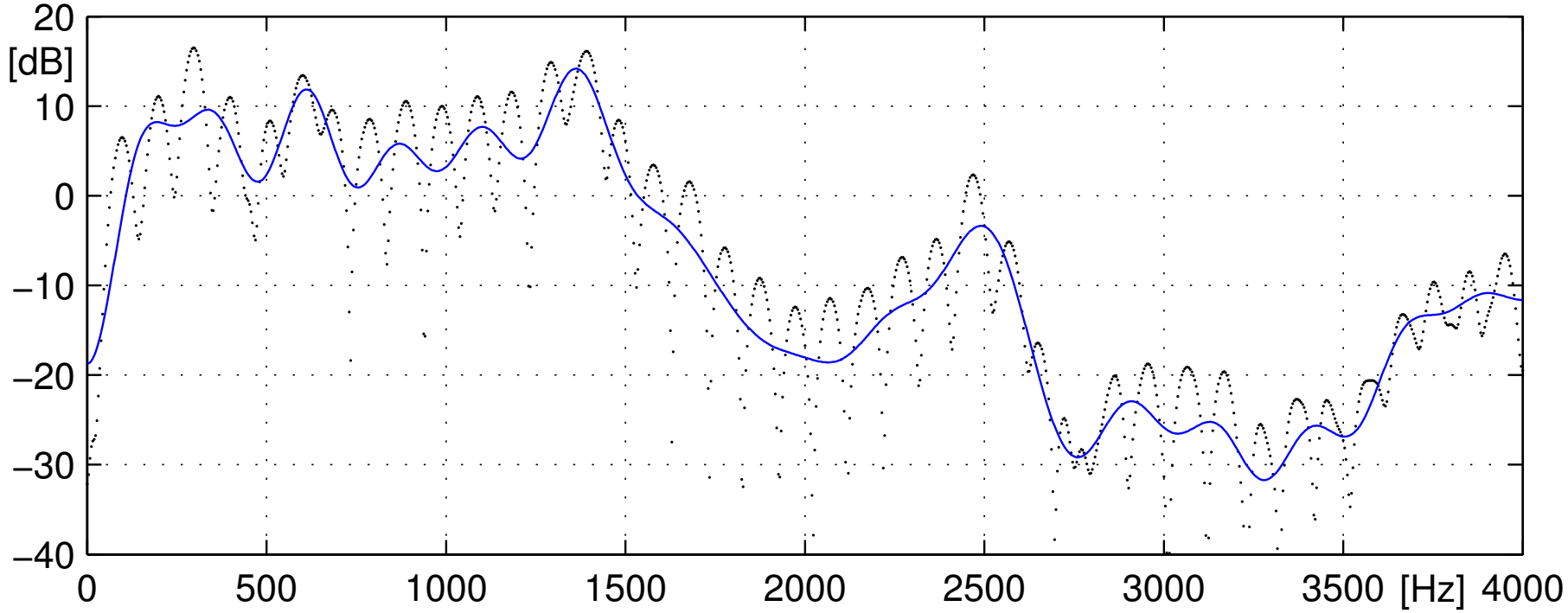
Schätzen der spektralen Enveloppe mittels linearer Prädiktion



Problem: Spektrale Maxima sind oft sehr schmal und hoch (Allpol-Filter), insbesondere bei hohen Stimmen

$\longrightarrow$ verursacht Töne oder Geräusche!

Schätzen der spektralen Enveloppe mittels cepstraler Glättung



Problem: Schmale spektrale Maxima oft zu tief geschätzt !

$\longrightarrow$ Laute werden undeutlich ! Lösung:

Schätzen der spektralen Enveloppe

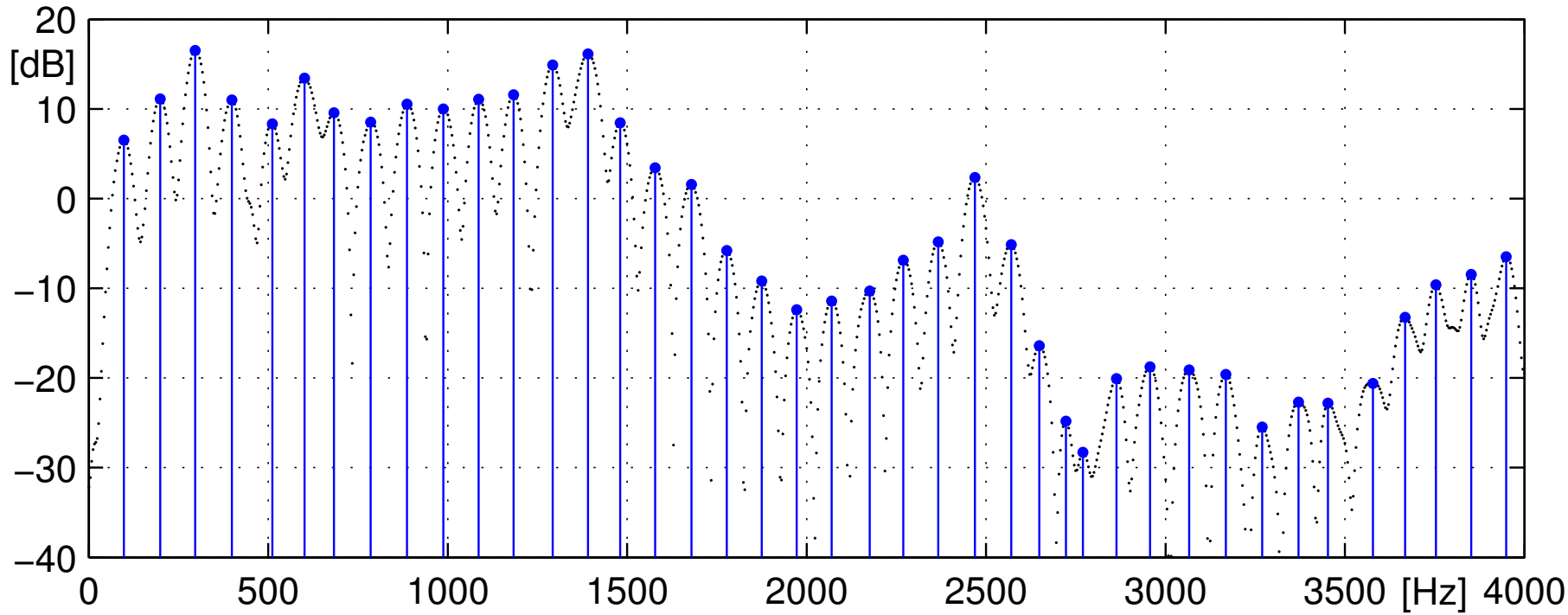
Gute Schätzung:

- Cepstrale Glättung des LPC-Spektrums
- Schätzen der Enveloppe aus dem "wirklichen" Spektrum (durch geeignete Interpolation)

>>>

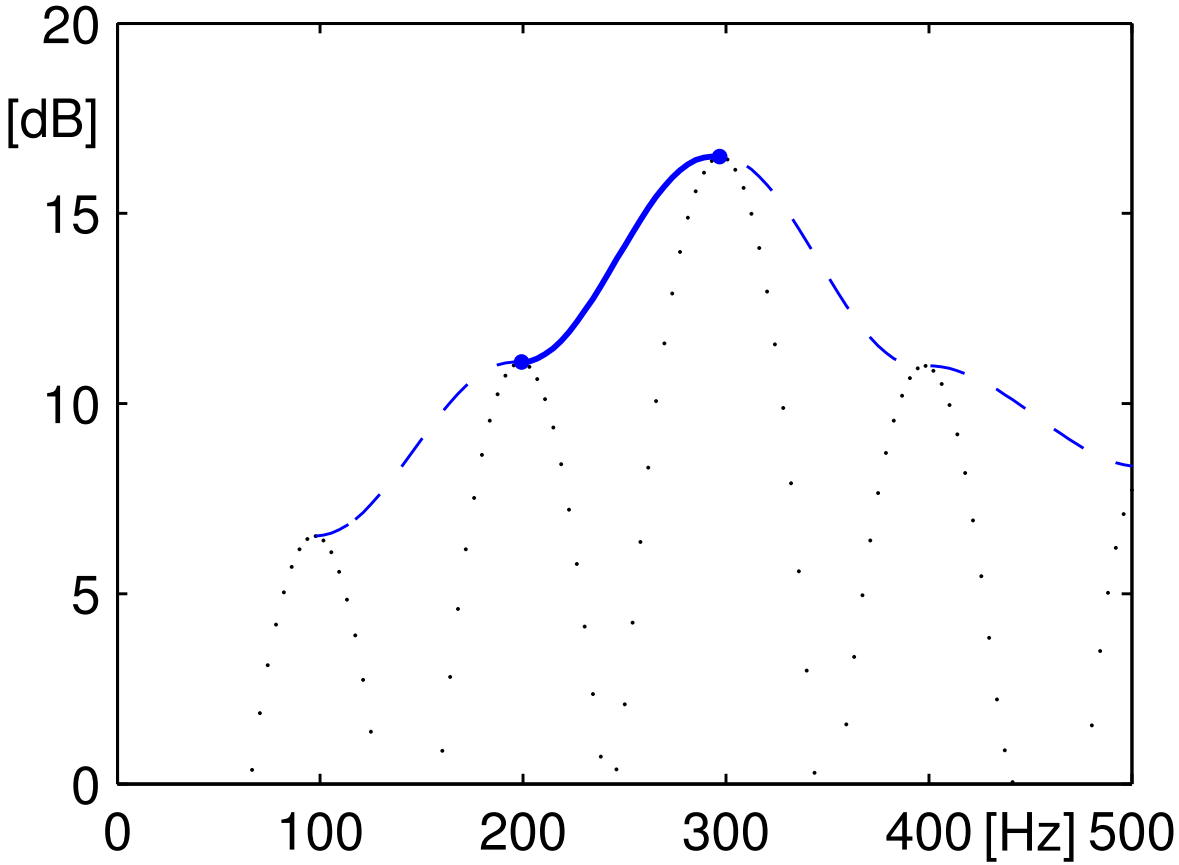
<<<

Schätzen der Enveloppe aus dem “wirklichen” Spektrum

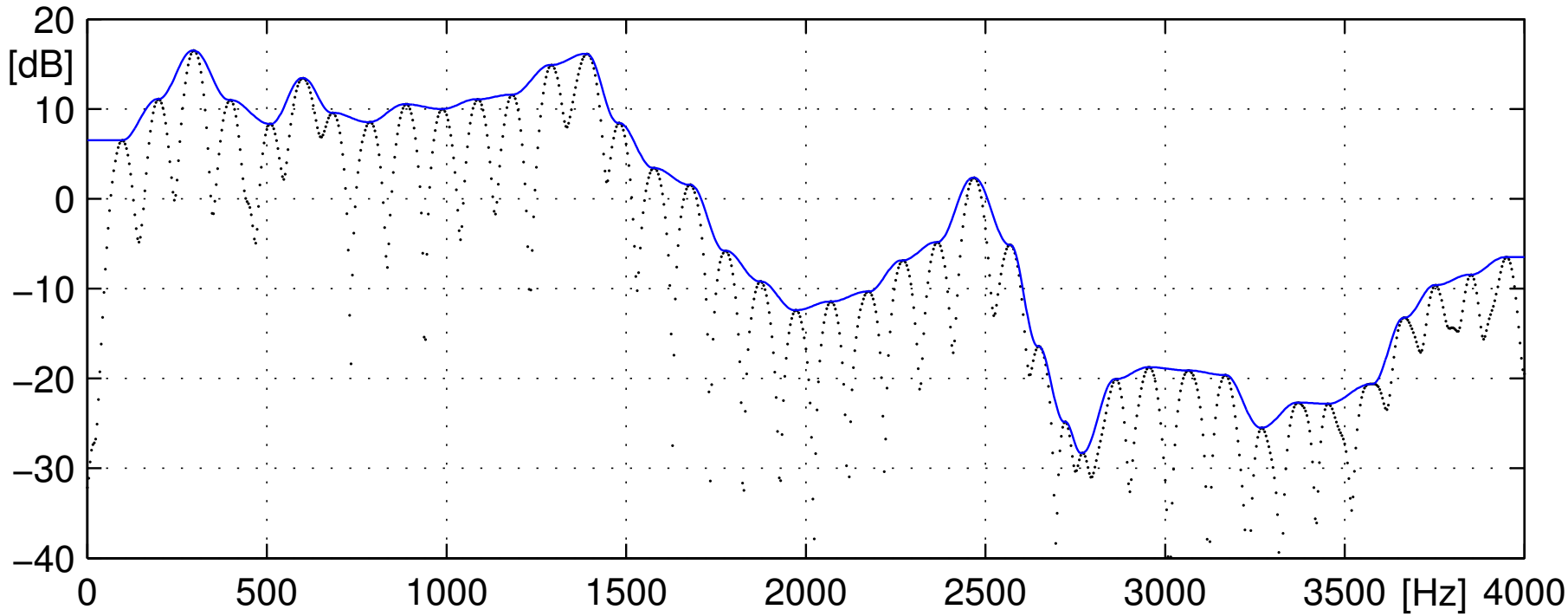


Frage: Wie soll interpoliert werden ?

Cos-Interpolation des “wirklichen” Spektrums



Enveloppe aus Cos-Interpolation des “wirklichen” Spektrums



<<<