Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Sprachverarbeitung I / 7 HS 2016

Sprachsynthese: Signalproduktion

Buch: Kapitel 9.1 und 9.2

Beat Pfister

Sprachverarbeitung I / 7

Vorlesung: **Sprachsynthese** (Teil I.2)

Phono-akustische Stufe

Sprachsignalproduktion

Übung: Sprachsignalproduktion mittels LPC-Analyse-Synthese

Produktion des Sprachsignals

Es gibt unzählige Verfahren zur Sprachsignalproduktion.
Sie lassen sich in drei Gruppen einteilen:

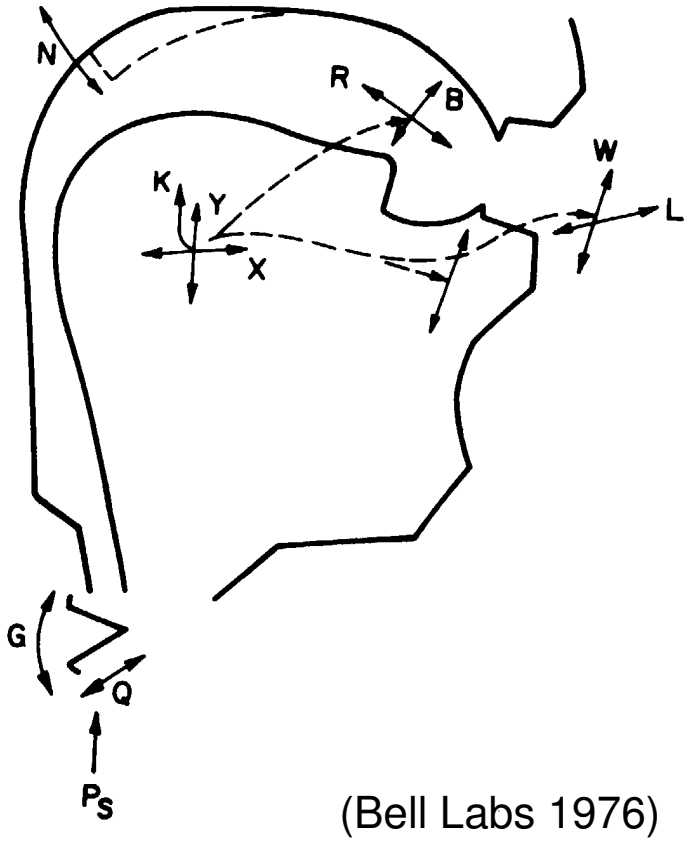
- Artikulatorischer Ansatz
- Signalmodellierung
- Verkettungsansatz

Produktion des Sprachsignals

Es gibt unzählige Verfahren zur Sprachsignalproduktion.
Sie lassen sich in drei Gruppen einteilen:

- Artikulatorischer Ansatz
- Signalmodellierung
- Verkettungsansatz

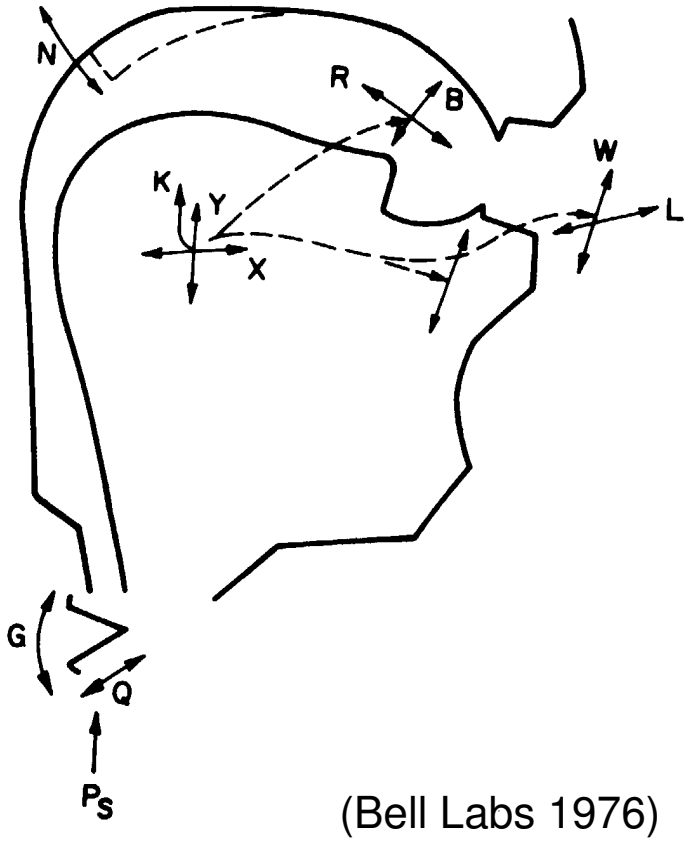
Artikulatorischer Ansatz der Sprachproduktion



(Bell Labs 1976)

Hörprobe in englischer Sprache

Artikulatorischer Ansatz der Sprachproduktion



(Bell Labs 1976)

I can read stories and speak them aloud. I do not understand what the words mean when I read them, but I can guess which words are important and which words are not, by rules I have been given.

Artikulatorisch motivierte Sprachproduktion

Ermittlung der Modellparameter nicht aus dem bewegten Vokaltrakt (d.h. aus Röntgenfilmen), sondern aus dem Sprachsignal

Uni Göttingen (2002): Hörprobe in deutscher Sprache

Artikulatorisch motivierte Sprachproduktion

Ermittlung der Modellparameter nicht aus dem bewegten Vokaltrakt (d.h. aus Röntgenfilmen), sondern aus dem Sprachsignal

Uni Göttingen (2002):

Peter macht neue Töne.
Rainer malt bunte Bilder.
Werner deutet die neuen Werte.

Produktion des Sprachsignals

Es gibt unzählige Verfahren zur Sprachsignalproduktion.
Sie lassen sich in drei Gruppen einteilen:

- Artikulatorischer Ansatz
- Signalmodellierung
- Verkettungsansatz

Signalmodellierung

Ansatz: Erzeugung eines Signals, das wie ein Sprachsignal tönt
und deshalb u.a. die folgenden Eigenschaften haben muss:

- Breitbandigkeit
- Resonanzfrequenzen (Formanten)
- Periodizität (stimmhaft / stimmlos)
- Quasi-Stationarität >>>
- Kontinuität

Realisationsmöglichkeiten:

a) LPC-Sprachproduktionsmodell

b) Formantmodell

c) HMM-Sprachsynthese

Signalmodellierung

Ansatz: Erzeugung eines Signals, das wie ein Sprachsignal tönt
und deshalb u.a. die folgenden Eigenschaften haben muss:

- Breitbandigkeit
- Resonanzfrequenzen (Formanten)
- Periodizität (stimmhaft / stimmlos)
- Quasi-Stationarität >>>
- Kontinuität

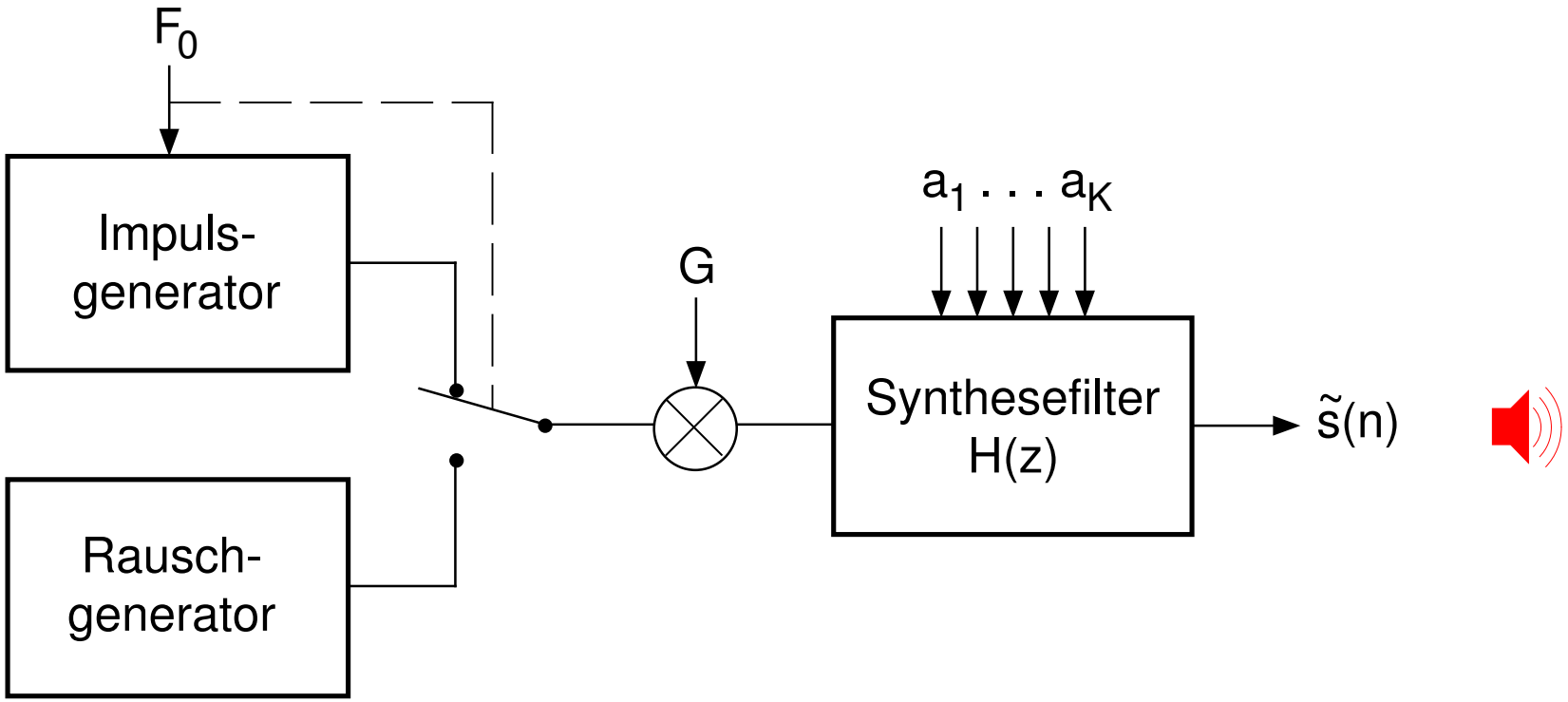
Realisationsmöglichkeiten:

a) LPC-Sprachproduktionsmodell

b) Formantmodell

c) HMM-Sprachsynthese

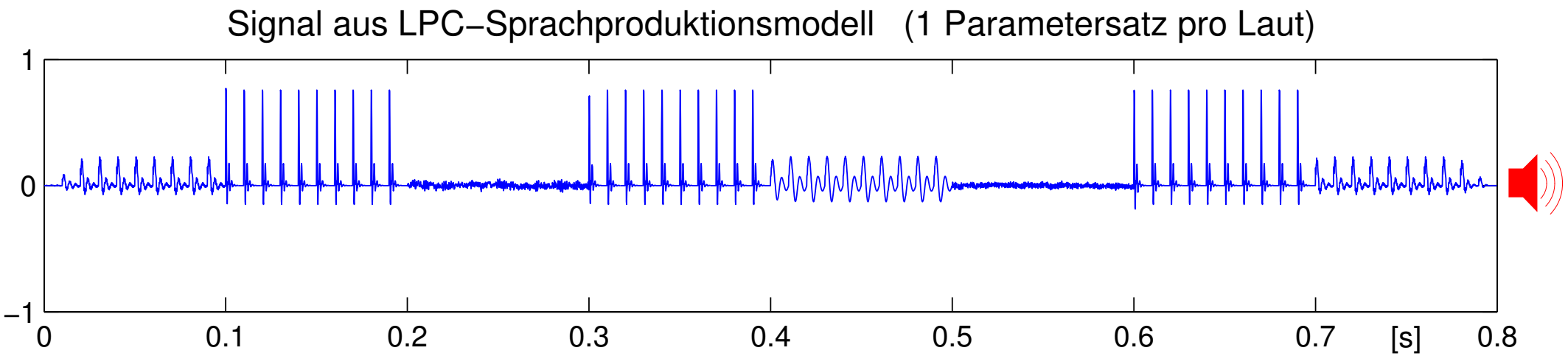
Sprachsignalproduktion mit dem LPC-Modell



Sprachsignalproduktion mit dem LPC-Modell

Synthetische Laute: [a] [l] [n] [f] [x]

Synthetisiertes Wort:



Signalmodellierung

Ansatz: Erzeugung eines Signals, das wie ein Sprachsignal tönt
und deshalb u.a. die folgenden Eigenschaften haben muss:

- Breitbandigkeit
- Resonanzfrequenzen (Formanten)
- Periodizität (stimmhaft / stimmlos)
- Quasi-Stationarität
- Kontinuität

>>>

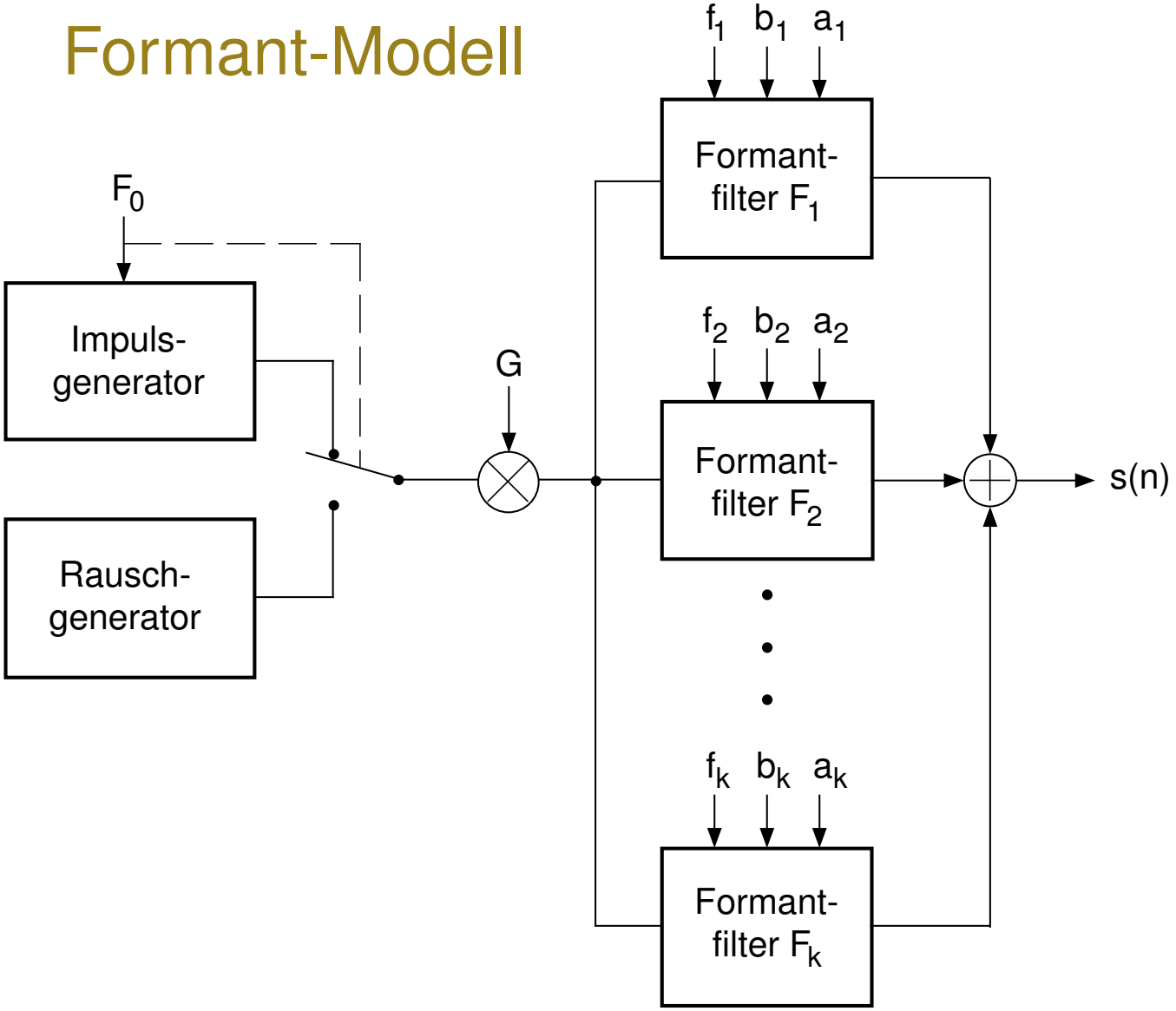
Realisationsmöglichkeiten:

a) LPC-Sprachproduktionsmodell

b) Formantmodell

c) HMM-Sprachsynthese

Formant-Modell



Englisch: (MIT 1983)

“Klattalk”

Deutsch: (Infovox 1995)

Männerstimme

Frauenstimme

Signalmodellierung

Ansatz: Erzeugung eines Signals, das wie ein Sprachsignal tönt
und deshalb u.a. die folgenden Eigenschaften haben muss:

- Breitbandigkeit
- Resonanzfrequenzen (Formanten)
- Periodizität (stimmhaft / stimmlos)
- Quasi-Stationarität
- Kontinuität

>>>

Realisationsmöglichkeiten:

a) LPC-Sprachproduktionsmodell

b) Formantmodell

c) HMM-Sprachsynthese

HMM-basierte Sprachsignalproduktion

Ansatz: Modellierung der Laute mittels kontextabhängiger HMM:

Hidden-Markov-Modelle zur Beschreibung des zeitlichen Verlaufes des Spektrums der Laute

Hörprobe von University of Miami, 2006 (ohne Grundfrequenz)

+ auch nicht-stationäre Laute beschreibbar
+ Stimmtransformation einfach machbar (reduziert Qualität!)
– Signalqualität durch Signalmodellierung limitiert

Produktion des Sprachsignals

Es gibt unzählige Verfahren zur Sprachsignalproduktion.
Sie lassen sich in drei Gruppen einteilen:

- Artikulatorischer Ansatz
- Signalmodellierung
- Verkettungsansatz

Verkettungsansatz

Prinzip: Erzeugung von Sprachsignalen durch Aneinanderfügen von Segmenten aus natürlichen Sprachsignalen

Vorteil: innerhalb eines Segmentes ist das Sprachsignal perfekt

Forderung: keine hörbaren Diskontinuitäten an Segmentstossstellen >>>
(bzgl. Spektrum und Grundfrequenz)

$\longrightarrow$ mehrere Laute pro Segment (Polyphone)

z.B. Diphonsegmente >>>

Sprachsignalproduktion nach Verkettungsansatz

Nötige Eigenschaften des produzierten Sprachsignals:

- Kontinuität: Stossstellen der verketteten Segmente sollen nicht hörbar sein
- Prosodie: Laute sind im Sprachsignal nicht überall gleich!

  $\longrightarrow$ Segmente müssen nicht nur die richtigen Laute enthalten, sondern auch hinsichtlich Dauer, Tonhöhe und Intensität passen!

Verkettungsansatz

Anforderungen an die Sprachsegmente:

- mehrere Laute pro Segment (Polyphone)
- minimale Diskontinuitäten an Segmentstossstellen
- prosodische Veränderung (Dauer und F_0)

Fragen:

1. Welche Grundelemente sind nötig? (Lautinventar)
2. Aus was für Sprachsignalen sind sie zu entnehmen?
3. Wie sind die Schnittpunkte festzulegen?
4. Wie lassen sie sich prosodisch verändern?

Verkettungsansatz

Anforderungen an die Sprachsegmente:

- mehrere Laute pro Segment (Polyphone)
- minimale Diskontinuitäten an Segmentstossstellen
- prosodische Veränderung (Dauer und F_0)

Fragen:

1. Welche Grundelemente sind nötig? (Lautinventar)
2. Aus was für Sprachsignalen sind sie zu entnehmen?
3. Wie sind die Schnittpunkte festzulegen?
4. Wie lassen sie sich prosodisch verändern?

Festlegen des Lautinventars für die Signalproduktion

Grundsätze:

1. grosses Lautinventar (feine Differenzierung)
   $\longrightarrow$ potentiell bessere Sprachqualität
2. Lautinventar jedoch so, dass Folge von Grundelementen aus phonologischer Darstellung ableitbar
   (phonologische Darstellung basiert auf Aussprache-Wörterbuch!)

Anforderung:

- Lautinventar muss mindestens alle Phoneme und die stellungsbedingten Allophone umfassen: z.B. [ç] und [x]
- Auswahl freier Allophone: z.B. [r] oder [ʀ]

Option: feinere Differenzierung der Laute mittels Kontext >>>

Wahl der Art der Grundelemente

Anforderung: jede mögliche Lautfolge muss durch eine passende Folge von Grundelementen darstellbar sein
(es kann fast jeder Laut nach jedem folgen, insb. an Wortübergängen)

Realisation: a) kompakter Satz von Grundelementen (je ein Exemplar)

$\longrightarrow$
- kurze Grundelemente (Diphone)
- prosodische Veränderung nötig

b) viele Sprachsignale $\Rightarrow$ "Korpus-Synthese"

SVOX AG (2005)

$\longrightarrow$
- variantenreicher Satz von Grundelementen
- optimale Auswahl der Segmente
  (beste Qualität, da nur wenig prosodische Veränderung nötig)

Verkettungsansatz

Anforderungen an die Sprachsegmente:

- mehrere Laute pro Segment (Polyphone)
- minimale Diskontinuitäten an Segmentstossstellen
- prosodische Veränderung (Dauer und F_0)

Fragen:

1. Welche Grundelemente sind nötig? (Lautinventar)
2. Aus was für Sprachsignalen sind sie zu entnehmen?
3. Wie sind die Schnittpunkte festzulegen?
4. Wie lassen sie sich prosodisch verändern?

Sprachsignale zur Gewinnung von Grundelementen

Anforderungen an Sprachsignale ergeben sich hinsichtlich:

- Sprache und dialektale Färbung (Deutsch ist in A, CH und D unterschiedlich)
- Sprechart und Sprechstil (präzise, gleichbleibende Artikulation)
- Stimmliche Eigenschaften des Sprechers / der Sprecherin
- natürliche Wörter / Ausdrücke oder Logatome (Kunstwörter)
- Signalqualität

Verkettungsansatz

Anforderungen an die Sprachsegmente:

- mehrere Laute pro Segment (Polyphone)
- minimale Diskontinuitäten an Segmentstossstellen
- prosodische Veränderung (Dauer und F_0)

Fragen:

1. Welche Grundelemente sind nötig? (Lautinventar)
2. Aus was für Sprachsignalen sind sie zu entnehmen?
3. Wie sind die Schnittpunkte festzulegen?
4. Wie lassen sie sich prosodisch verändern?

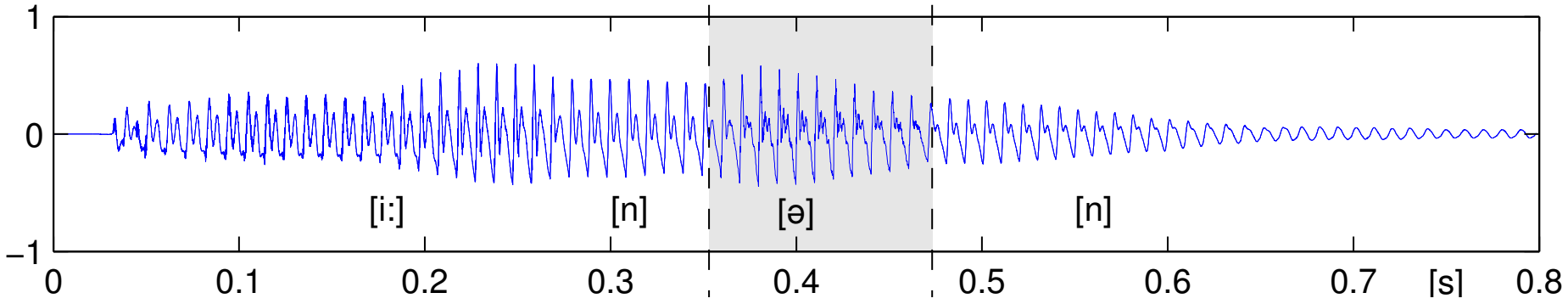
Extraktion eines Diphons

Gegeben: Sprachsignal des Wortes “ihnen” (in Lautschrift: [iːnən])

Absicht: Extraktion des Diphons [ən]

Frage: Wo ist zu schneiden ? Wo ist die “Mitte” von [ə] ?
Hilft das Spektrogramm ?

>>>



Extraktion eines Diphons

Problem: Die Lautmitte (Anfangs- bzw. Endpunkt eines Diphons) ist weder im Frequenz- noch im Zeitbereich zu sehen !

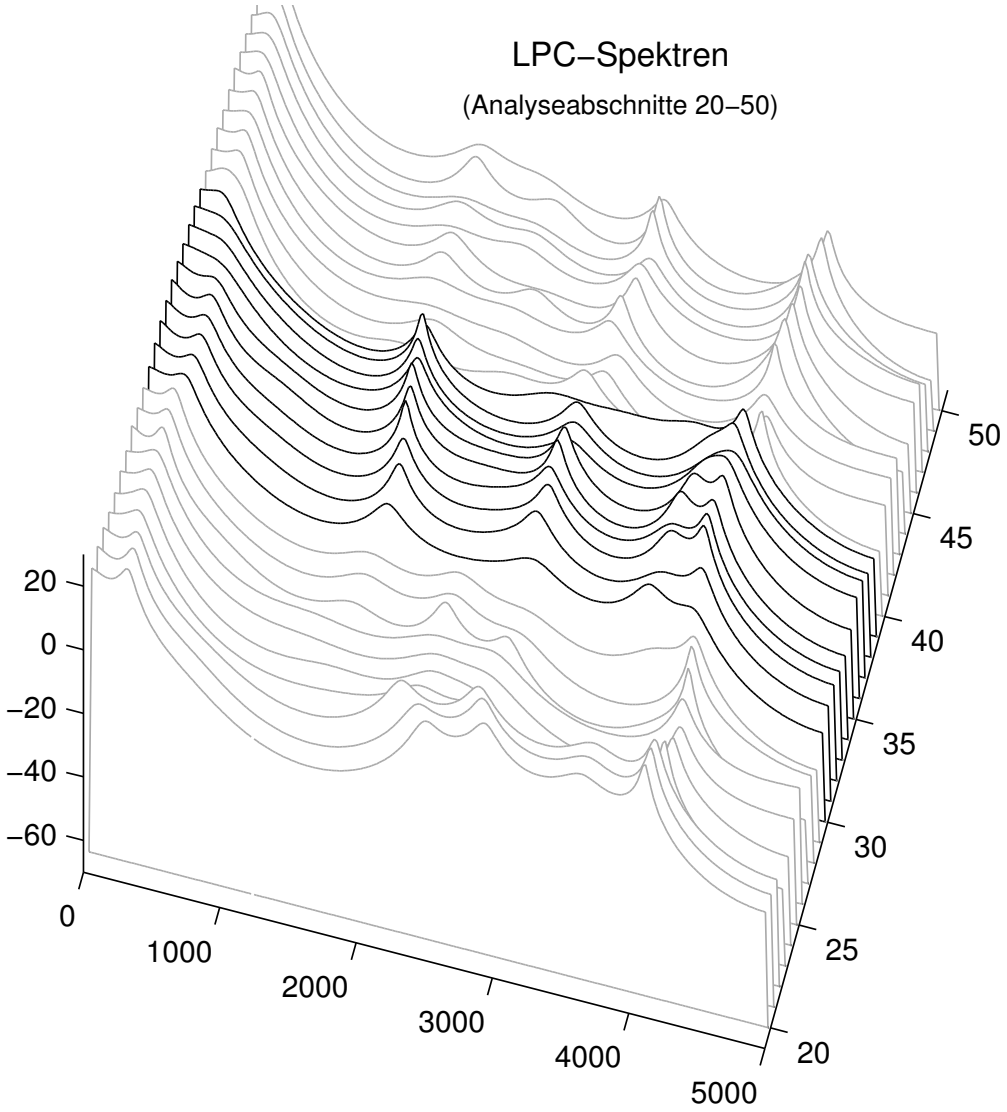
Lösung: a) Globale Optimierung der Schnittpunkte, damit die Diskontinuitäten an den Diphonstossstellen möglichst klein sind.

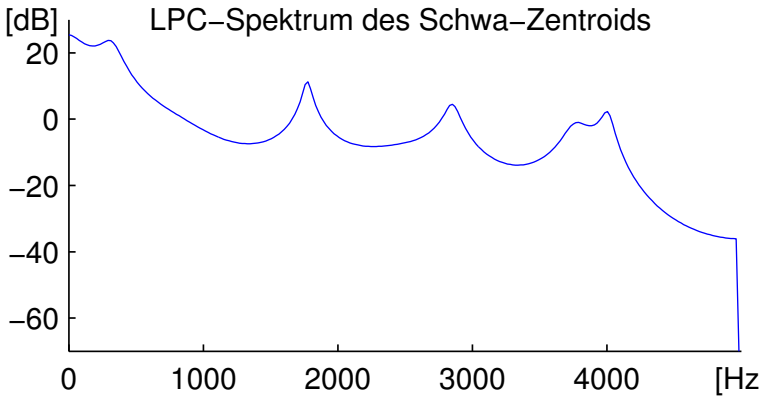
$\longrightarrow$ machbar, aber nicht ausreichend!

(Qualität der Laute ist auch zu berücksichtigen)

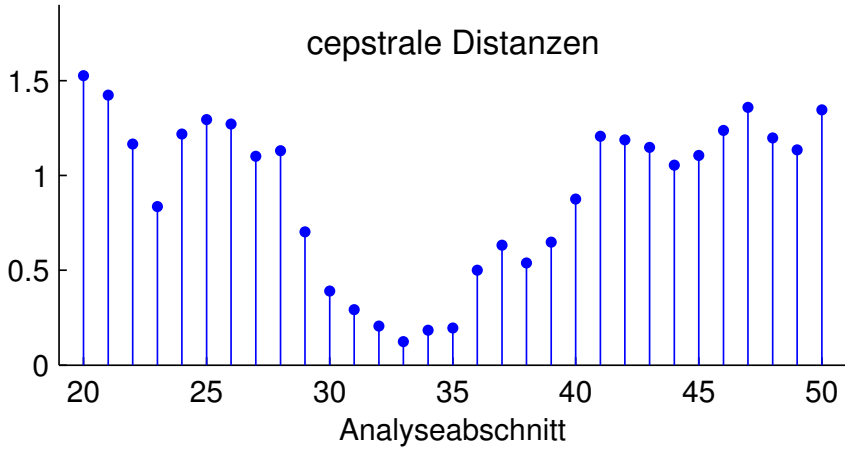
b) Bestimmung der Lautmitte mittels Laut-Zentroid

Bestimmung der Schnittstelle in Laut [ə] mittels Zentroid von [ə]





>>>



Verkettungsansatz

Anforderungen an die Sprachsegmente:

- mehrere Laute pro Segment (Polyphone)
- minimale Diskontinuitäten an Segmentstossstellen
- prosodische Veränderung (Dauer und F_0)

Fragen:

1. Welche Grundelemente sind nötig? (Lautinventar)
2. Aus was für Sprachsignalen sind sie zu entnehmen?
3. Wie sind die Schnittpunkte festzulegen?
4. Wie lassen sie sich prosodisch verändern?

Prosodische Veränderung der Grundelemente

Aufgabe: Einstellung von Dauer, Grundfrequenz und Intensität der Sprachsignalsegmente (Diphone) vor der Verkettung

Methoden:

a) via LPC-Analyse-Synthese (siehe Übungen 5, 6 und 7)

b) Fourier-Analyse-Synthese

c) PSOLA-Technik im Zeitbereich

Prosodische Veränderung der Grundelemente

Aufgabe: Einstellung von Dauer, Grundfrequenz und Intensität der Sprachsignalsegmente (Diphone) vor der Verkettung

Methoden:

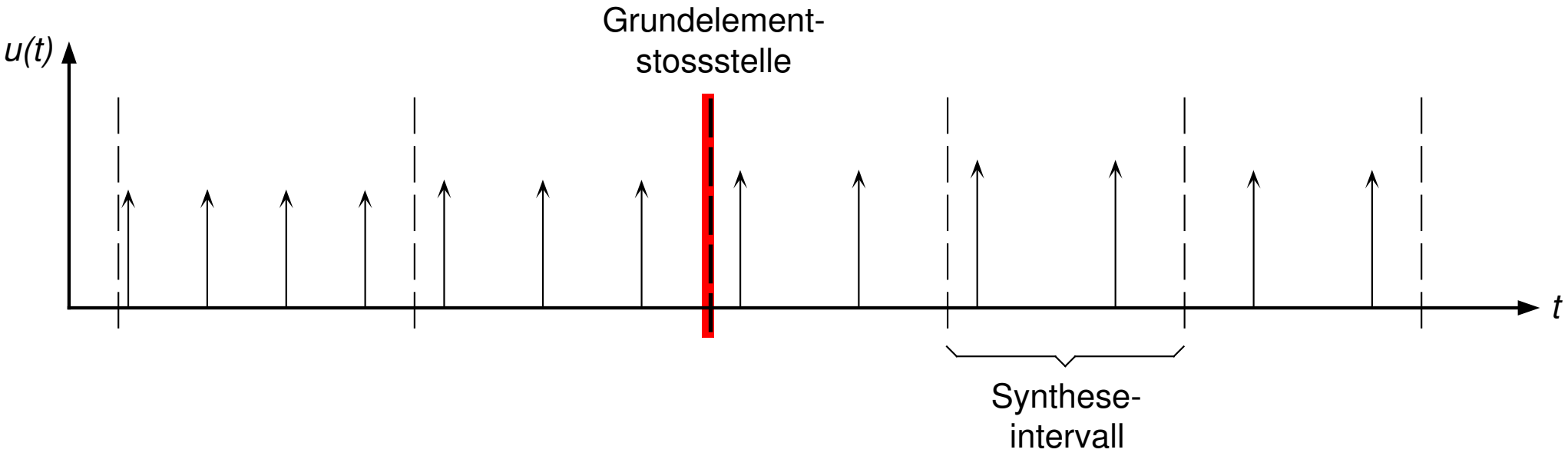
a) via LPC-Analyse-Synthese (siehe Übungen 5, 6 und 7)

b) Fourier-Analyse-Synthese

c) PSOLA-Technik im Zeitbereich

LPC-Analyse-Synthese

Veränderung der prosodischen Grössen von Grundelementen (z.B. Diphonen)



Achtung: Impulsabstände bei Syntheseintervall- und Grundelementgrenzen richtig!

Prosodische Veränderung der Grundelemente

Aufgabe: Einstellung von Dauer, Grundfrequenz und Intensität der Sprachsignalsegmente (Diphone) vor der Verkettung

Methoden:

a) via LPC-Analyse-Synthese (siehe Übungen 5, 6 und 7)

b) Fourier-Analyse-Synthese

c) PSOLA-Technik im Zeitbereich

Fourier-Analyse-Synthese

Prinzip der Frequenz- und Daueränderung via Fourier-Analyse-Synthese >>>

Dauer und Grundfrequenz eines Sprachsignals verändern:

1. Abschnittweise Zerlegung des Signals in Sinuskomponenten
2. Veränderung der Frequenz der Sinuskomponenten
3. Rekonstruktion der Signalabschnitte mit gewünschter Dauer

Fourier-Analyse-Synthese von Sprachsignalen:

- ist grundsätzlich einfach, aber . . . (Lektion 9; Anhang B)
- ergibt gute Signalqualität
- ist aber sehr aufwändig

Prosodische Veränderung der Grundelemente

Aufgabe: Einstellung von Dauer, Grundfrequenz und Intensität der Sprachsignalsegmente (Diphone) vor der Verkettung

Methoden:

a) via LPC-Analyse-Synthese (siehe Übungen 5, 6 und 7)

b) Fourier-Analyse-Synthese

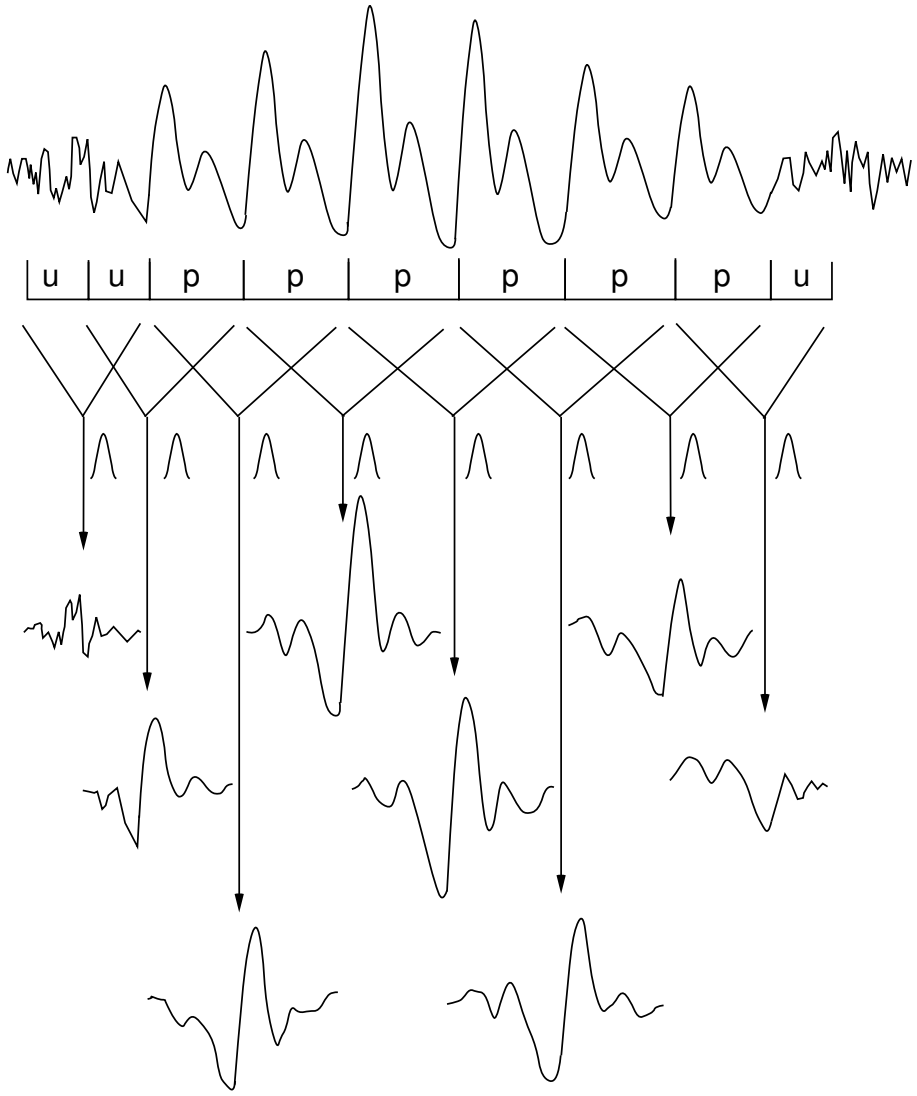
c) PSOLA-Technik im Zeitbereich

Zeitbereichs-PSOLA

(pitch-synchronous overlap-add)

Periodensynchrone
Segmentierung
des Sprachsignals

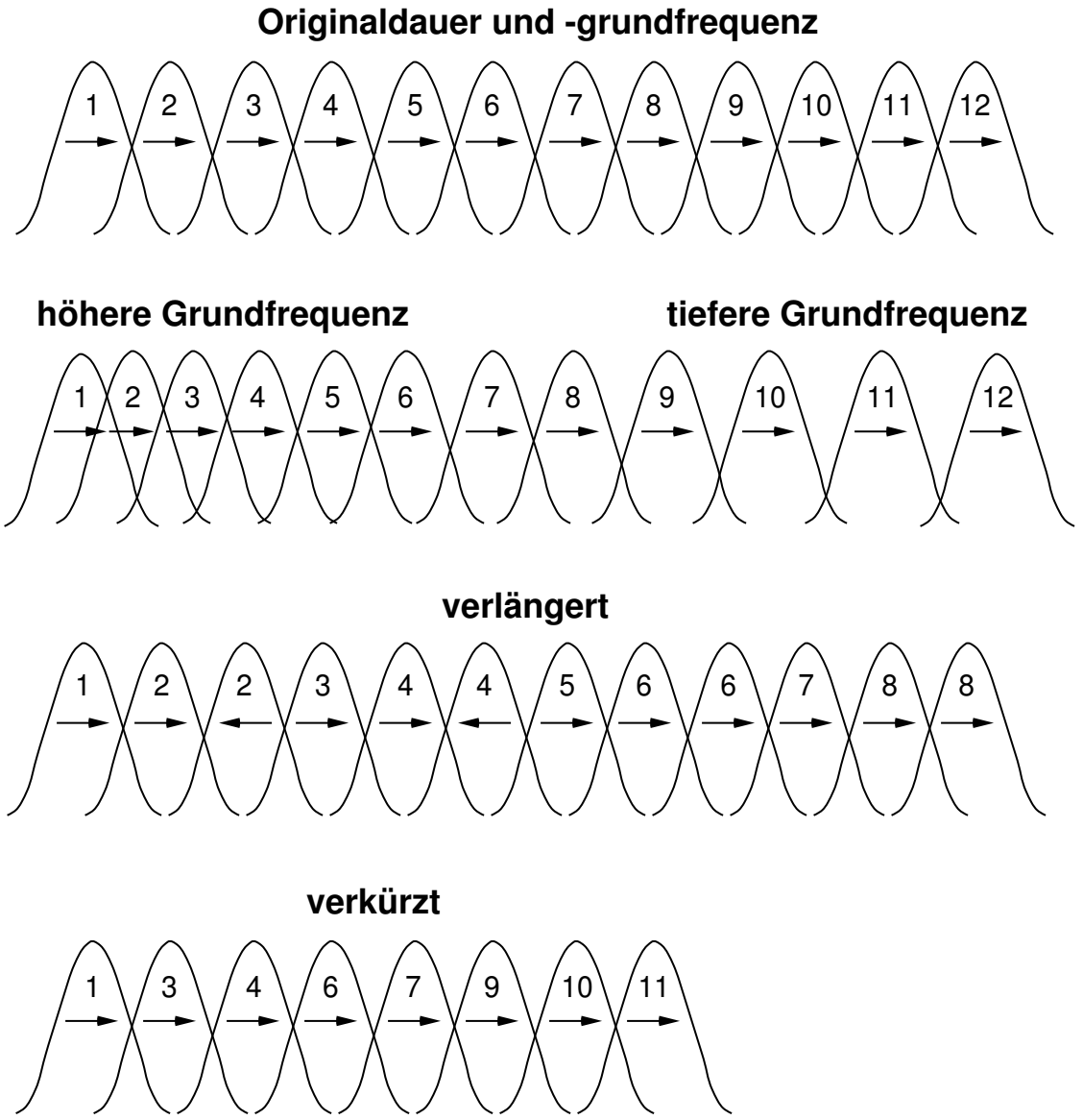
→ Doppelperioden-Segmente
(mit Hanning-Fenster multipliziert)



Zeitbereichs-PSOLA

(pitch-synchronous overlap-add)

Modifikation von
Dauer und Grundfrequenz
durch Verdoppeln,
Weglassen oder Verschieben
von Doppelperioden-Segmenten



Zusammenfassung

Verfahren zur Sprachsignalgenerierung in der Sprachsynthese

- Artikulatorischer Ansatz
- Signalmodellierung
- Verkettungsansatz

Methoden für die prosodische Veränderung von Sprachsignalen

(bei Verkettungsansatz nötig)

- LPC-Analyse-Synthese
- Fourier-Analyse-Synthese
- PSOLA-Technik im Zeitbereich

Thema der nächsten Lektion:

Steuerung der Prosodie

(Grundfrequenz und Dauer der Laute)

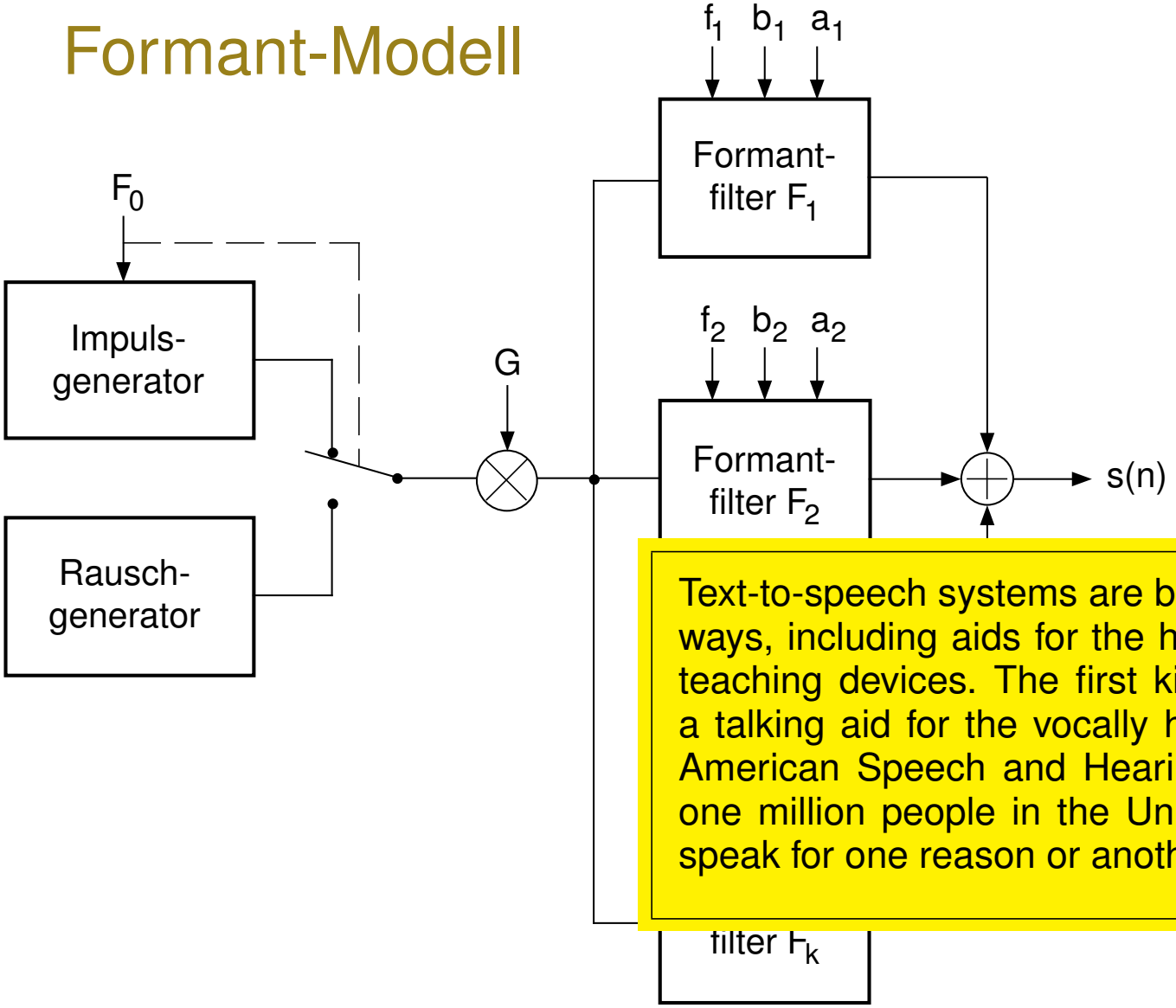
Zur Übersicht der Vorlesung *Sprachverarbeitung I* >>>

Quasistationarität der Laute

Laut	Original	LPC-Anal-Synth	LPC (stationär)
[a]			
[e]			
[i]			

<<<

Formant-Modell



Englisch: (MIT 1983)

“Klattalk”

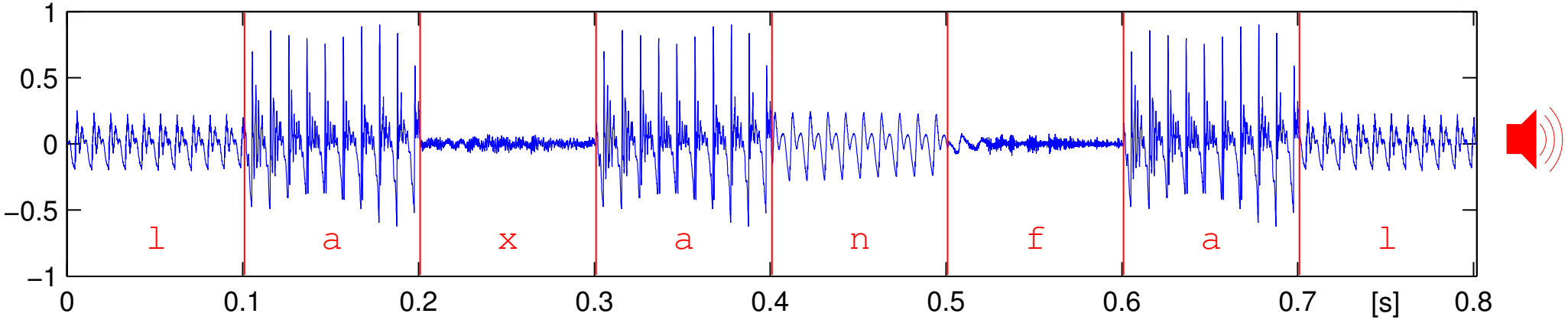
Text-to-speech systems are beginning to be applied in many ways, including aids for the handicapped, medical aids and teaching devices. The first kind of aid to be considered is a talking aid for the vocally handicapped. According to the American Speech and Hearing Association there are over one million people in the United States who are unable to speak for one reason or another.

<<<

Verkettung von Lautelementen

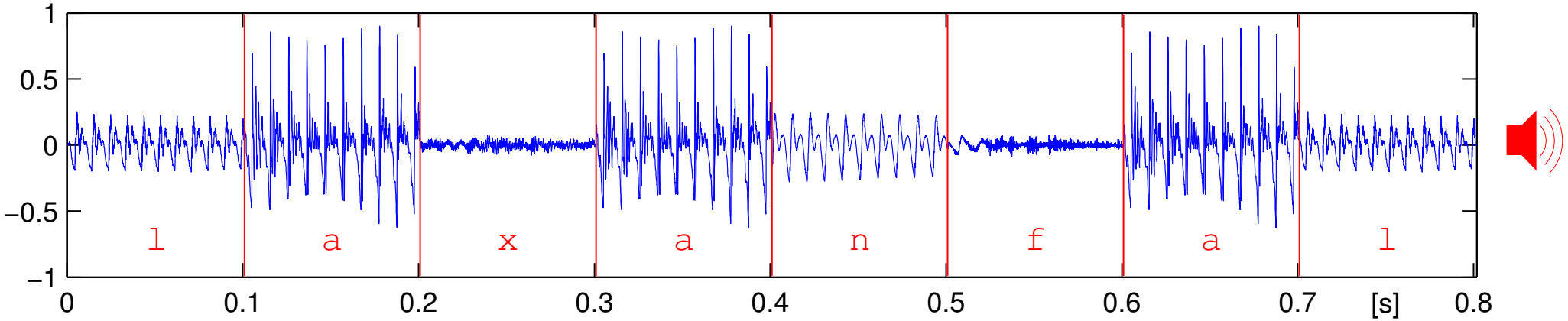
Natürliche Lautsegmente: [a] [l] [n] [f] [x]

Synthetisiertes Wort:

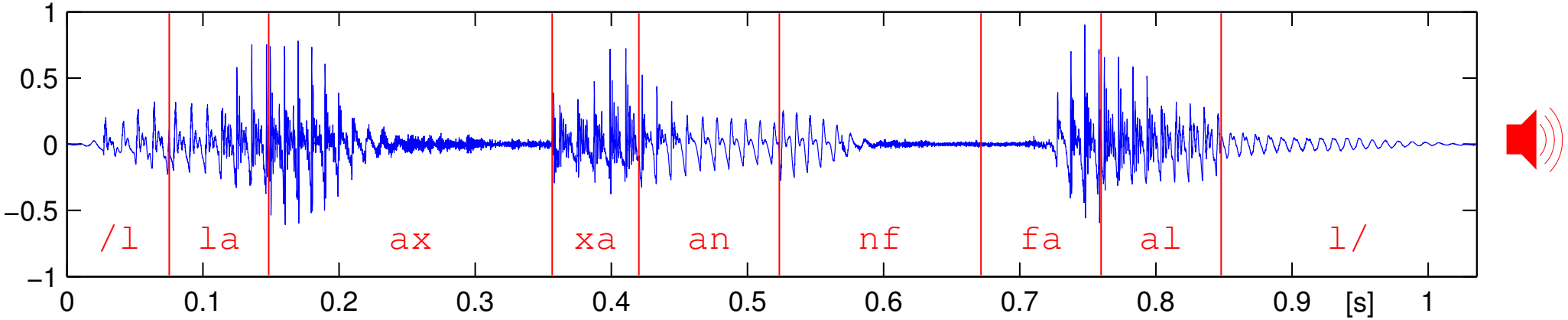


<<<

Verkettung von Lautelementen



Verkettung von Diphonelementen



<<<

Feinere Lautdifferenzierung mittels Kontext

Absicht: Berücksichtigen, dass Laute unterschiedlich gesprochen werden

z.B. [n] in “Bahn” [baːn] und “Bann” [ban]

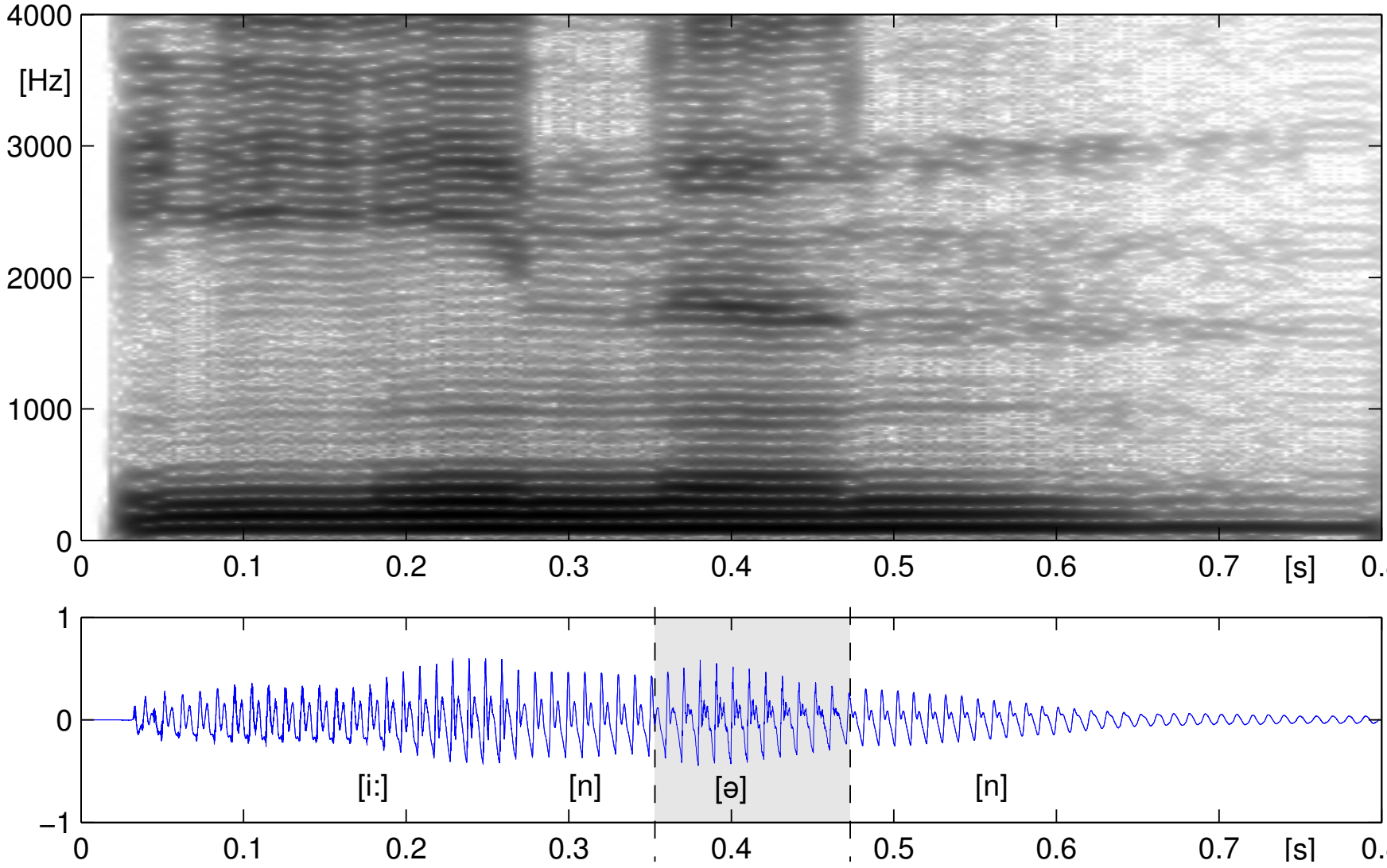
Lösung: [n] nach Langvokal $\longrightarrow$ [n_1]

[n] nach Kurzvokal $\longrightarrow$ [n_2]

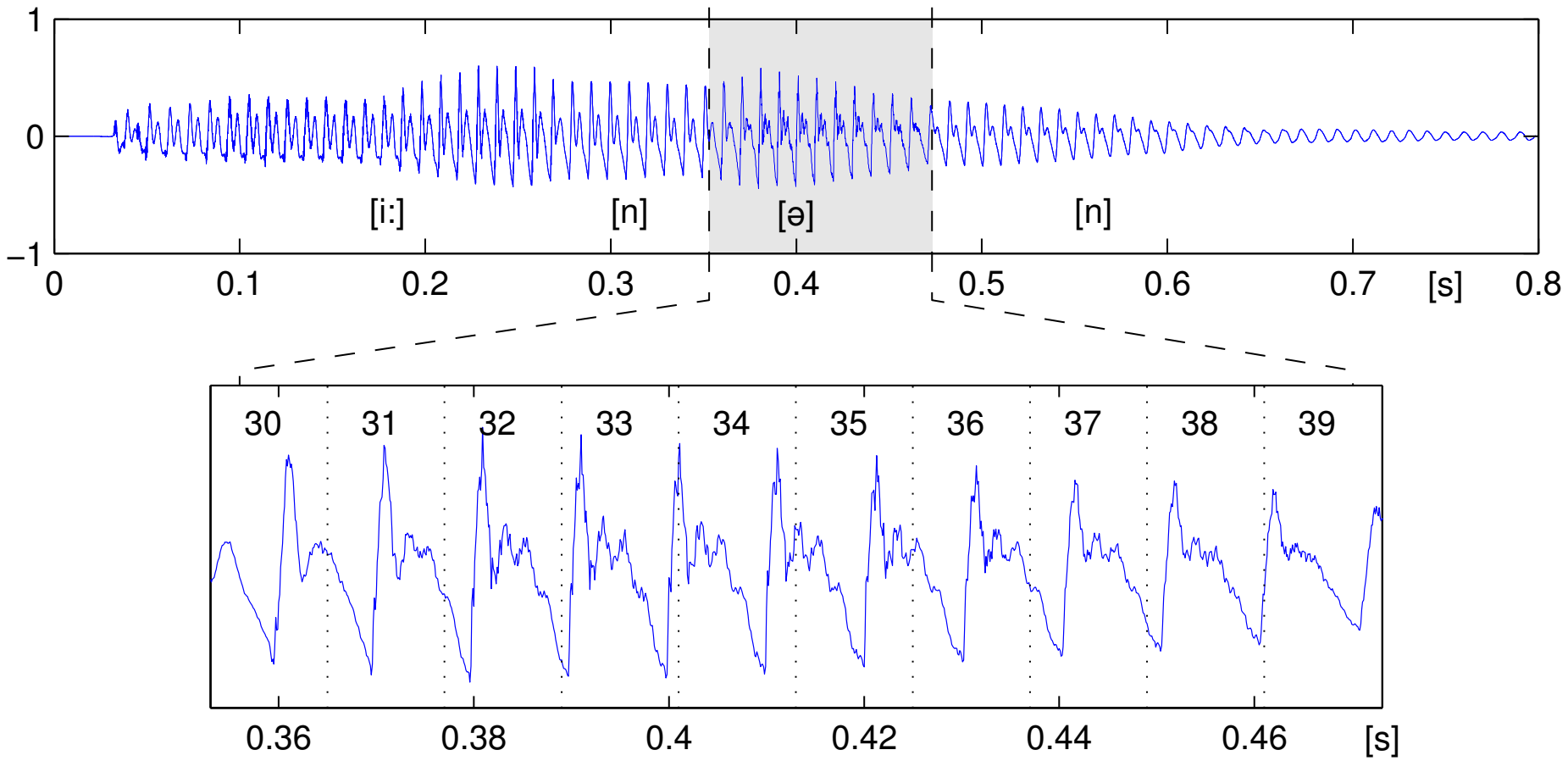
Verkettung: “Bahn” [baːn] $\longrightarrow$ [baːn_1] $\longrightarrow$ [/b] [baː] [aːn_1] [n_1/]

“Bann” [ban] $\longrightarrow$ [ban$_2$] $\longrightarrow$ [/b] [ba] [an_2] [n_2/]

<<<


4000
[Hz]
3000
2000
1000
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
[s]
0.8
1
0
−1
[i:]
[n]
[ə]
[n]
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
[s]
0.8

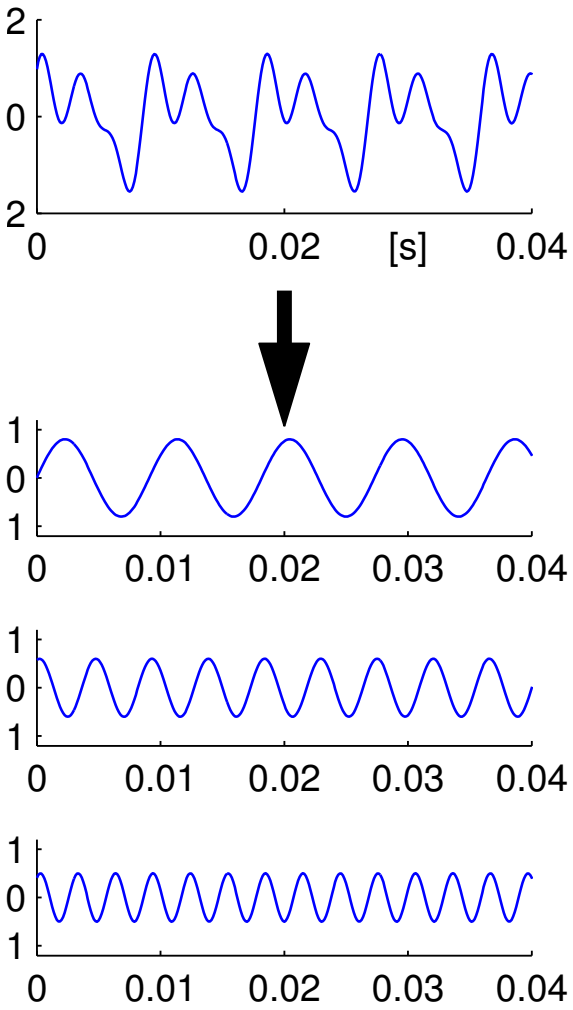
<<<


1
0
−1
[i:]
[n]
[ə]
[n]
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
[s]
0.8
30
31
32
33
34
35
36
37
38
39
0.36
0.38
0.4
0.42
0.44
0.46
[s]

<<<

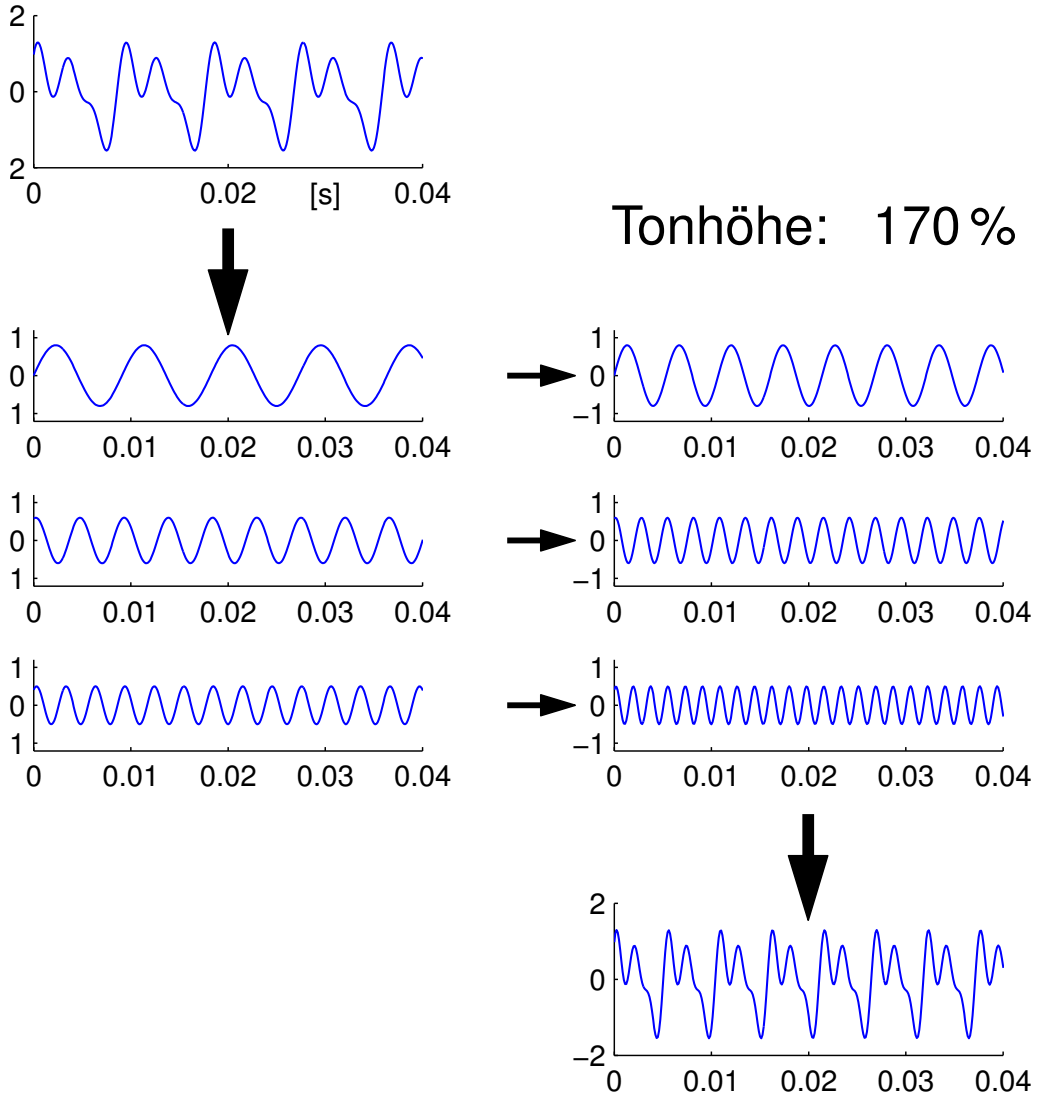
Fourier-Analyse

Zerlegung in Sinuskomponenten



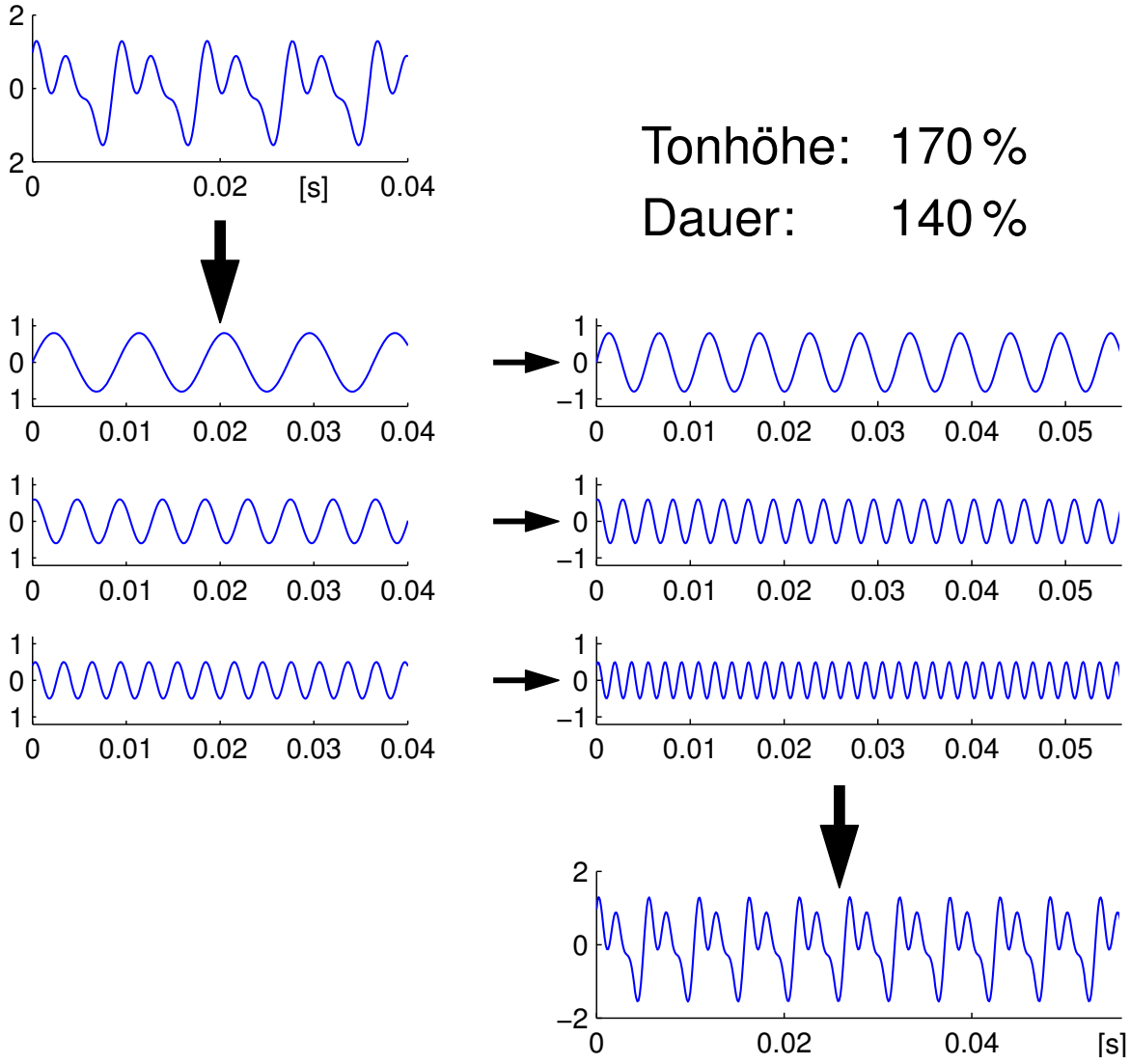
Fourier-Synthese

Superposition von
frequenzveränderten
Sinuskomponenten



Fourier-Synthese

Superposition von
frequenzveränderten und
verlängerten / verkürzten
Sinuskomponenten



<<<