

Sprachverarbeitung I / 8 HS 2016

Sprachsynthese: Prosodiesteuerung

Buch: Kapitel 9.3

Beat Pfister



Sprachverarbeitung I / 8

Vorlesung: **Sprachsynthese** (Teil I.3)

Prosodiesteuerung

- Prosodie in der Sprachsynthese
- Dauersteuerung
- Grundfrequenzsteuerung
- Intensitätssteuerung

Übung: Dauersteuerung

Prosodie der Sprache

Linguistische Funktion:

- Kennzeichnung —→ Unterscheidung von Aussage / Frage / Befehl
- Gewichtung —→ Hervorheben wichtiger Wörter und Silben
- Gruppierung —→ Gliederung des Sprachflusses in sinnvolle Einheiten
(Sprechgruppen / prosodische Phrasen)

Ausserlinguistische Funktion:

Prosodie bestimmt, ob Stimme traurig, wütend, fröhlich usw. wirkt

Steuerung der Prosodie

Gegeben: a) linguistische Information (phonologische Darstellung)
b) sonstige Information (z.B. Sprechgeschwindigkeit, Frauenstimme, ...)

Gesucht: Werte der prosodischen Grössen für jeden Laut

- Dauer
- Grundfrequenz (Tonhöhe)
- Intensität

Annahme: Dauer, Grundfrequenz und Intensität unabhängig steuerbar
(separate Komponenten der Prosodiesteuerung)

Dauersteuerung

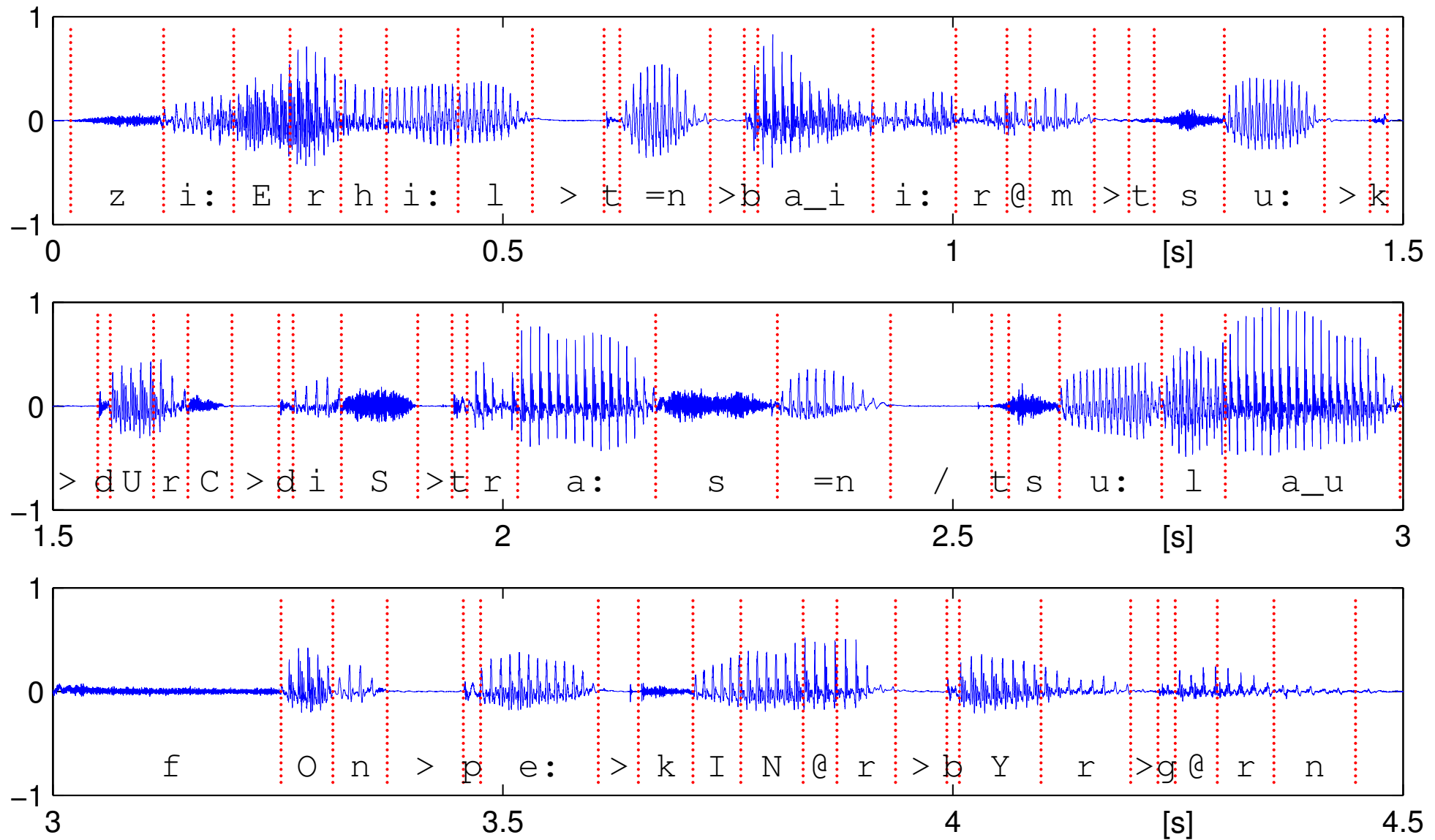
Ziel: Ermitteln der Dauer jedes Lautes
(phonologische Darstellung ist gegeben)

Achtung: Lang- und Kurzvokale müssen unterscheidbar sein
(es sind Phoneme, die z.B. “Wahl” von “Wall” unterscheiden)

Frage: Wie ist vorzugehen?

→ Untersuchung von Beispielsignalen, die so gesprochen sind,
wie die Sprachsynthese “sprechen” sollte!

In Laute segmentiertes Sprachsignal



Segmentierte Sprachsignale

Was kann man mit segmentierten Sprachsignalen machen?

- Dauerwerte für [a] und [a:] untersuchen

>>>

Segmentierte Sprachsignale

Was kann man mit segmentierten Sprachsignalen machen?

- Dauerwerte für [a] und [a:] untersuchen

→ Herausfinden, welche Faktoren die Lautdauer beeinflussen

Vergleich der normierten Histogramme:

H_a : Lautdauerwerte mit Einflussfaktor X

H_b : Lautdauerwerte ohne Einflussfaktor X

Für $H_a \approx H_b$ → Faktor X hat keinen Einfluss auf die Lautdauer

sonst → Faktor X beeinflusst die Lautdauer

Linguistische Einflüsse auf die Lautdauer

linguist. Faktor	diskrete Werte
Lautklasse	Lang- Kurzvokal Diphthong Nasal Frikativ Plosiv andere
Stimmhaftigkeit	stimmhaft stimmlos
Lautposition	Silben-Ansatz -Kern (Nukleus) -Koda
Nachbarlaute	Lautklasse & Stimmhaftigkeit (für linken und rechten Nachbarn)
Silbenakzent	Stärkegrad: 1 2 3 unakzentuiert sonst
Silbengrösse	Anzahl Laute: 1 2 3–4 5–7 >7
Phrasengrösse	Anzahl Silben: 1 2 3–4 5–7 >7
Silbenposition	in Phrase: 1. 2. vorletzte letzte sonst
Phrasengrenze	Stärkegrad: 0 1 2 sonst
Satztyp	Aussagesatz Fragesatz Befehlssatz

Methode zur Dauersteuerung

Grundsatz: Wahrnehmung der **relativen** Lautdauer

Prinzip: Einflüsse müssen multiplikativ wirken

Lautdauer: $d = p_0 \prod_i p_i^{c_i}$ mit p_0 charakteristische Lautdauer
 p_i Einfluss des linguistischen Faktors i
 c_i ling. Faktor i trifft zu (= 1) | trifft nicht zu (= 0)

Frage: Wie kann man die p_i bestimmen ?

Methode zur Dauersteuerung

Grundsatz: Wahrnehmung der **relativen** Lautdauer

Prinzip: Einflüsse müssen multiplikativ wirken

Lautdauer: $d = p_0 \prod_i p_i^{c_i}$ mit p_0 charakteristische Lautdauer
 p_i Einfluss des linguistischen Faktors i
 c_i ling. Faktor i trifft zu (= 1) | trifft nicht zu (= 0)

Regeln:

- RD0:** Laute haben im Mittel eine neutrale Dauer von p_0 ms
(keine weitere Regel trifft zu)
- RD1:** Langvokale werden um den Faktor p_1 verlängert
- RD2:** Laute einer Silbe mit Hauptakzent werden um p_2 verlängert
- RD3:** Laute in unbetonten Silben werden gekürzt ($p_3 < 1$)
- RD4:** Laute der letzten Silbe vor einer Phrasengrenze der Stärke 1
oder vor dem Satzende werden um p_4 verlängert
- ⋮

Tabelle der gemessenen Lautdauern

Laut	Dauer [ms]	Lang- laut	Haupt- akzent	unbetonte Silbe	letzte Silbe	...
z	103.0	0	0	1	0	
i:	77.9	1	0	1	0	
E	62.8	0	0	1	0	
r	56.4	0	0	1	0	
h	50.7	0	1	0	0	
i:	79.4	1	1	0	0	
l	91.2	0	1	0	0	
>	71.0	0	0	1	0	
t	17.6	0	0	1	0	
@	20.0	0	0	1	0	
n	80.4	0	0	1	0	
>	37.7	0	0	1	0	
b	15.2	0	0	1	0	
a_i	127.8	0	0	1	0	
:						

Linearer Ansatz zur Dauersteuerung

Dauer d eines Lautes schätzen als Summe der Einflüsse der zutreff. Faktoren

Wie auf Dauerformel anwendbar? (multiplikative Einflüsse)

→ Dauerformel logarithmieren:

$$\begin{aligned}\log \tilde{d} &= \log p_0 + c_1 \log p_1 + c_2 \log p_2 + c_3 \log p_3 + c_4 \log p_4 + \dots \\ &= [c_0 \ c_1 \ c_2 \ c_3 \ c_4 \ \dots] \cdot [q_0 \ q_1 \ q_2 \ q_3 \ q_3 \ \dots]^t = \mathbf{c} \mathbf{q}^t\end{aligned}$$

→ Einflussgrößen q_i aus gemessenen Lautdauern bestimmen

Bestimmung der Einflussgrößen p_i

Pro Messwert eine Gleichung (Gleichungssystem in Matrizenform)

$$\log \tilde{d} = \begin{bmatrix} \log \tilde{d}(1) \\ \log \tilde{d}(2) \\ \log \tilde{d}(3) \\ \log \tilde{d}(4) \\ \log \tilde{d}(5) \\ \log \tilde{d}(6) \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & \dots \\ 1 & 1 & 0 & 1 & 0 & \dots \\ 1 & 0 & 0 & 1 & 0 & \dots \\ 1 & 0 & 0 & 1 & 0 & \dots \\ 1 & 0 & 1 & 0 & 0 & \dots \\ 1 & 1 & 1 & 0 & 0 & \dots \\ \vdots & & & & & \end{bmatrix} \cdot \begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \\ q_4 \\ \vdots \end{bmatrix} = \begin{bmatrix} c(1) \\ c(2) \\ c(3) \\ c(4) \\ c(5) \\ c(6) \\ \vdots \end{bmatrix} \cdot \begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \\ q_4 \\ \vdots \end{bmatrix} = C q$$

Weil Anzahl Gleichungen $>$ Anzahl Einflussfaktoren

→ Optimierung! Minimierung des Schätzfehlers $E^2 = \sum_n \{\log d(n) - \log \tilde{d}(n)\}^2$

d.h. partielle Ableitungen null setzen

Anwendung des linearen Ansatzes zur Dauersteuerung

Bestimmung der Dauer eines konkreten Lautes:

1. Ermitteln der für den Laut zutreffenden Faktoren: $\rightarrow c$
(analog zu einer Zeile in Messwerttabelle)
2. Schätzen der Dauer des Lautes: $\tilde{d} = p_0 \cdot p_1^{c_1} \cdot p_2^{c_2} \cdot p_3^{c_3} \cdot p_4^{c_4} \cdot \dots$
wobei $p_i = 10^{q_i}$

- Fragen:
- Wie gut schätzt der lineare Ansatz die Lautdauern?
 - Gibt es bessere Ansätze?

>>>

Besserer Ansatz zur Dauersteuerung

Annahme: Dauer der Laute ist nicht zufällig

Folge: Zusammenhang zwischen Einflussfaktoren $c = [c_1 \ c_2 \ c_3 \dots]$ und Lautdauer als (nichtlineare) Transformation darstellbar:

$$d(n) = \Psi\{c(n)\}$$

Frage: Wie kann die unbekannte Transformation Ψ ermittelt werden?
(es sind viele Ein-/Ausgangspaare dieser Transformation bekannt)

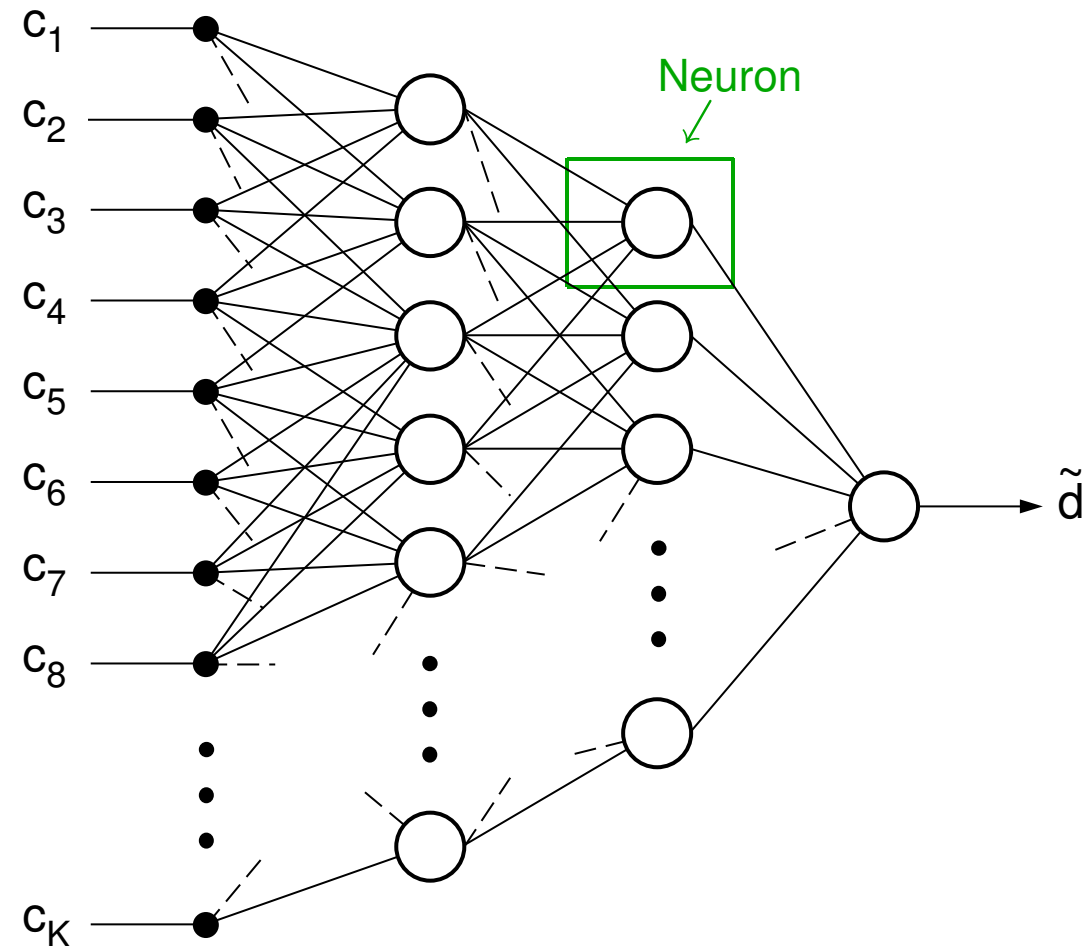
Idee: Transformation mit **neuronalem Netz** approximieren, so dass

$$\tilde{\Psi}\{c(n)\} = \tilde{d}(n) \approx d(n) \quad \forall n$$

Dauersteuerung mit neuronalem Netz

Stärken des neuronalen Netzes:

- im Training beliebige, nichtlineare Transformation ψ lernbar
(falls genügend Trainingsdaten vorhanden)
- gute Verallgemeinerung
(nur wenige aller möglichen c in den Trainingsdaten vorhanden)



Dauersteuerung mit neuronalem Netz

Resultat sehr gut, falls

- alle wichtigen Einflussfaktoren berücksichtigt
- genügend Trainingsdaten vorhanden
- Topologie und Training des neuronalen Netzes passend

Steuerung der Grundfrequenz

Ziel: Verlauf der Grundfrequenz in Funktion der Zeit: $F_0(t)$

>>>

Steuerung der Grundfrequenz

Ziel: Verlauf der Grundfrequenz in Funktion der Zeit: $F_0(t)$ >>>

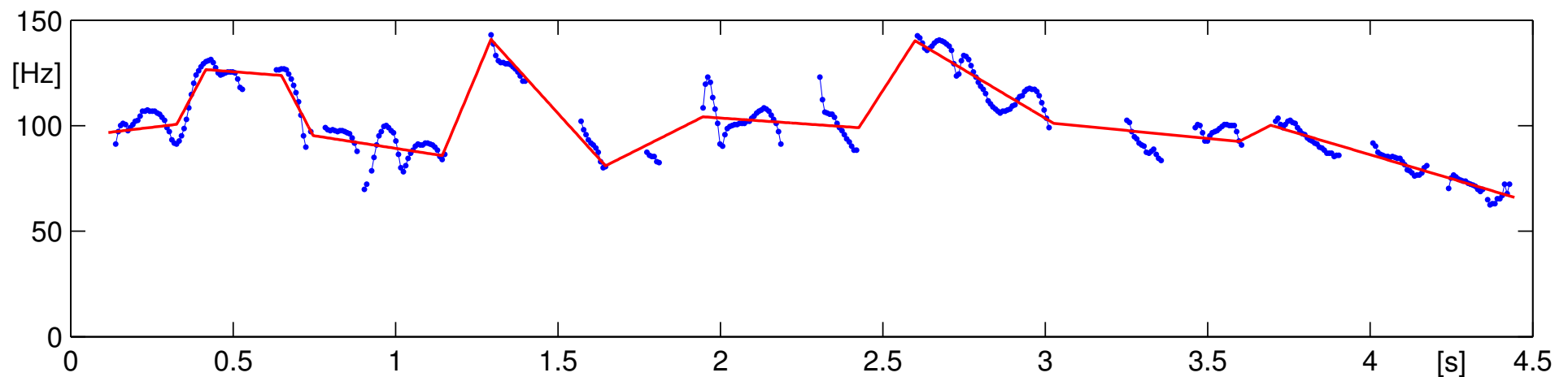
Problem: Grundfrequenz- und Dauersteuerung müssen getrennt sein!
→ Grundfrequenz nicht in Funktion der Zeit formulieren!

Lösung: Grundfrequenz in Funktion des Indexes linguistischer Einheiten
→ $F_0(i)$, wobei $i = 1, 2, 3, \dots$ z.B. der Silbenindex sein kann

Fragen: • Wodurch wird der Grundfrequenzverlauf beeinflusst? >>>
• Wie genau muss der Grundfrequenzverlauf sein?

Approximation des Grundfrequenzverlaufs

stückweise lineare Approximation



LPC-Analyse-Synthese mit originaler F_0

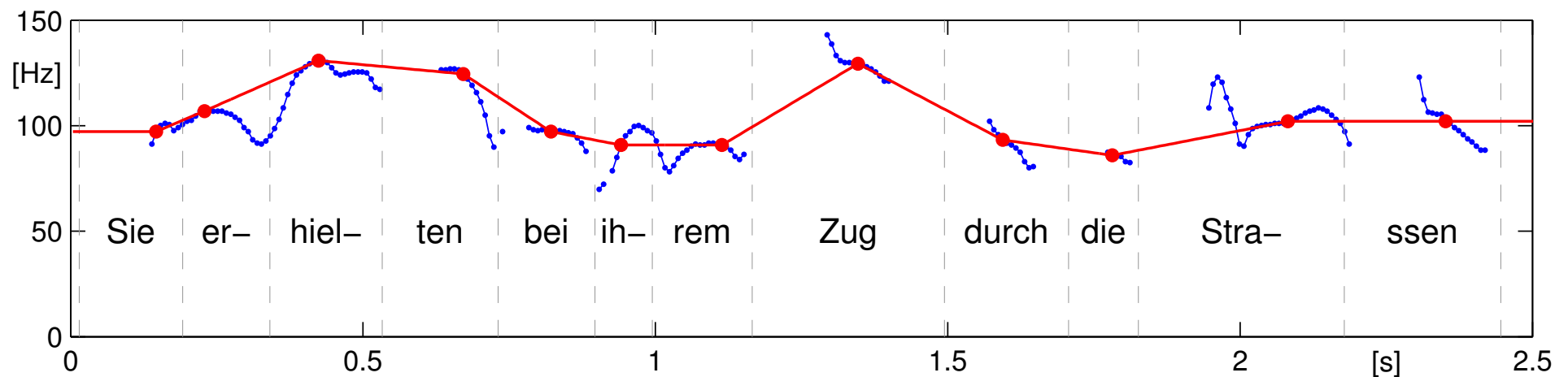


LPC-Analyse-Synthese mit linear approx. F_0



Silbenbezogener Grundfrequenzverlauf

stückweise lineare Approximation zwischen Silbenträgern



LPC-Analyse-Synthese mit originaler F_0

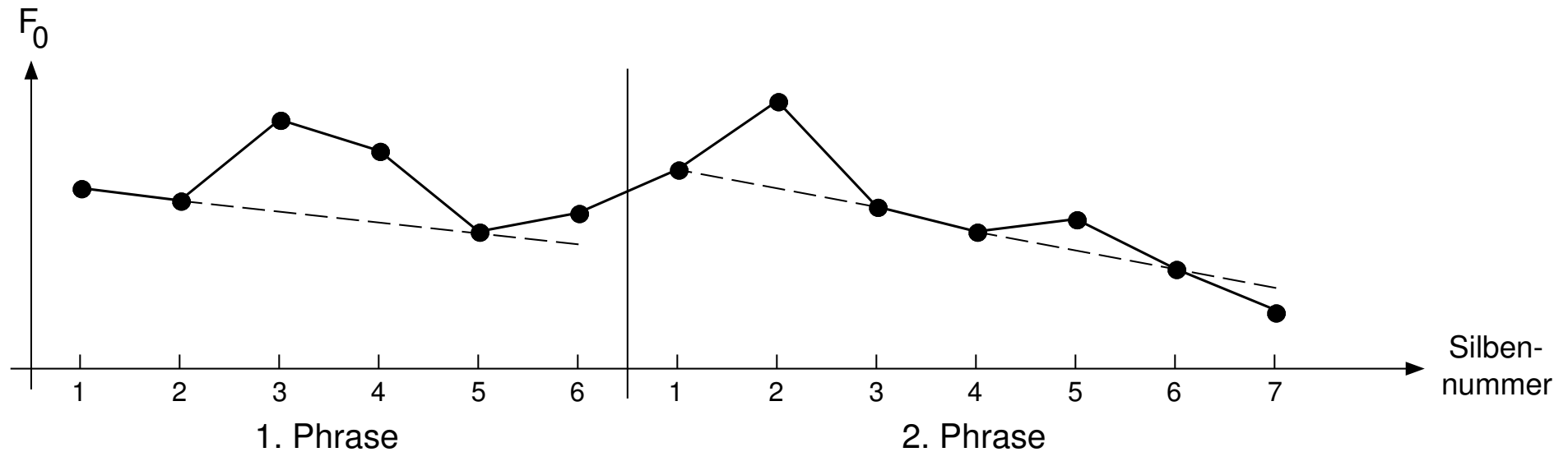


LPC-Analyse-Synthese mit silbensynchr. approx. F_0



Schematisierter Grundfrequenzverlauf

Deklinationsgerade (pro Phrase) & Abweichung der Silbenkerne



Grundfrequenzverlauf einer Phrase in Funktion der Silbennummer:

$$\tilde{F}_0(j) = F_a + \frac{j-1}{J} F_d + F_{\Delta}$$

Merke: F_a , F_d und F_{Δ} sind abhängig von linguistischen Faktoren!

Linguistische Einflussfaktoren als Regeln

- RF1:** Am Anfang einer progredienten Phrase beträgt die Grundfrequenz p_1 .
- RF2:** Am Anfang einer terminalen Phrase beträgt die Grundfrequenz p_2 .
- RF3:** In einer progredienten Phrase ist der Einfluss der Deklination auf die Grundfrequenz der j -ten Silbe $\frac{j-1}{J}p_3$.
- RF4:** In einer terminalen Phrase ist der Einfluss der Deklination auf die Grundfrequenz der j -ten Silbe $\frac{j-1}{J}p_4$.
- RF5:** Hat eine Silbe die Akzentstärke 1, dann ist F_0 gegenüber der Deklinationsgeraden um p_5 erhöht.
- RF6:** Hat eine Silbe die Akzentstärke 2, dann ist F_0 gegenüber der Deklinationsgeraden um p_6 erhöht.
- RF7:** F_0 der letzten Silbe einer Terminalphrase ist gegenüber der Deklinationsgeraden um p_7 erhöht.

⋮

Linearer Ansatz zur Grundfrequenzsteuerung

Mit Regelparameter geschätzte Grundfrequenz für die j -te Silbe:

$$\begin{aligned}\tilde{F}_0(j) &= F_a + \frac{j-1}{J} F_d + F_\Delta \\ &= c_1 p_1 + c_2 p_2 + \frac{j-1}{J} (c_3 p_3 + c_4 p_4) + c_5 p_5 + \dots \\ &= \begin{bmatrix} c_1 & c_2 & \frac{j-1}{J} c_3 & \frac{j-1}{J} c_4 & c_5 & \dots \end{bmatrix} \cdot [p_1 \ p_2 \ p_3 \ p_4 \ p_5 \ \dots]^t = \mathbf{c} \mathbf{p}^t\end{aligned}$$

Bestimmung der Parameter p_i :

- Tabelle mit gemessenen F_0 -Werten und zugeh. Einflussfaktoren
- Optimierung

>>>

Mängel des linearen Ansatzes zur F_0 -Steuerung

- F_0 -Verlauf von Phrasen desselben Typs stets sehr ähnlich
→ Sprechmelodie wirkt stereotyp
(wegen Schematisierung mit Deklinationsgeraden)
- Schlecht für sehr kurze Sätze
(Satz mit nur einer Silbe ergibt konstante F_0)
- Mächtigkeit des linearen Ansatzes ist nicht ausreichend
(Effekt mehrerer Einflussfaktoren ungleich Summe der Einzeleffekte)
- Häufig auftretende Faktoren “überstimmen” die seltenen
(wegen Minimierung des globalen Schätzfehlers)

Besserer Ansatz zur Grundfrequenzsteuerung

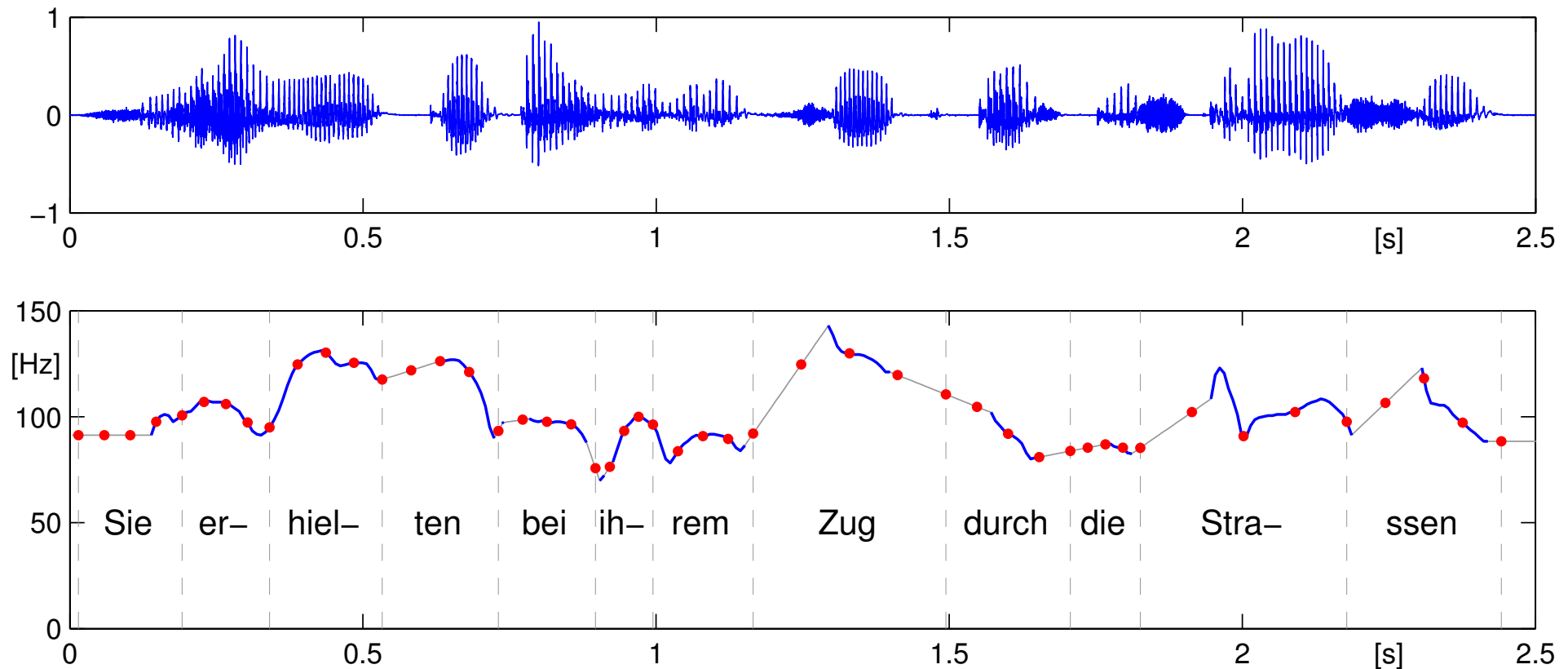
Idee: Neuronales Netz zur Transformation der
Einflussfaktoren $c = [c_1 \ c_2 \ c_3 \dots]$
in die Grundfrequenz F_0 einer Silbe

Frage: Wie ist das Problem mit den kurzen Sätzen zu meistern?

Lösung: Mehrere F_0 -Werte pro Silbe

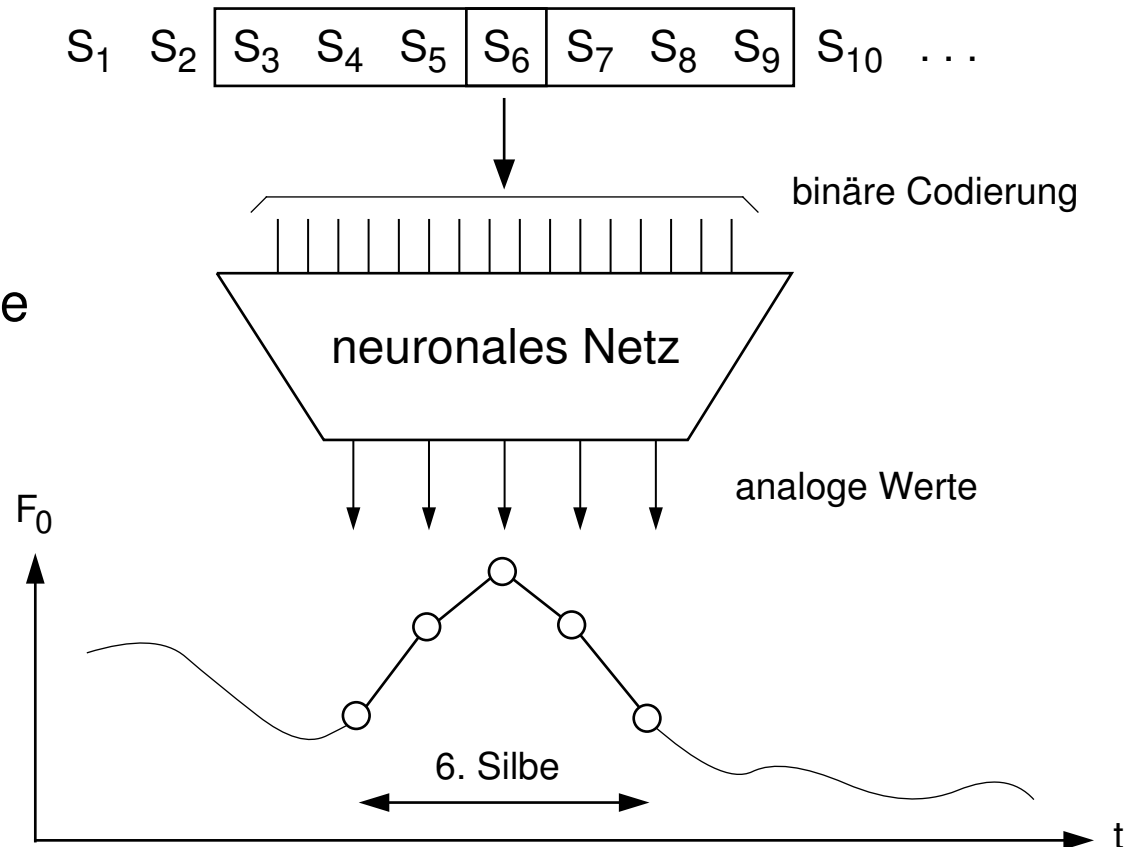
Silbensynchrone Abtastung des interpolierten F_0 -Verlaufs

“Sie erhielten bei ihrem Zug durch die Strassen . . .” (5 Abtastwerte/Silbe)



Grundfrequenzgenerierung mit neuronalem Netz

Neuronales Netz
transformiert
Einflussfaktoren einer Silbe
(und ihrer Nachbarn)
in 5 Stützwerte des
 F_0 -Verlaufs



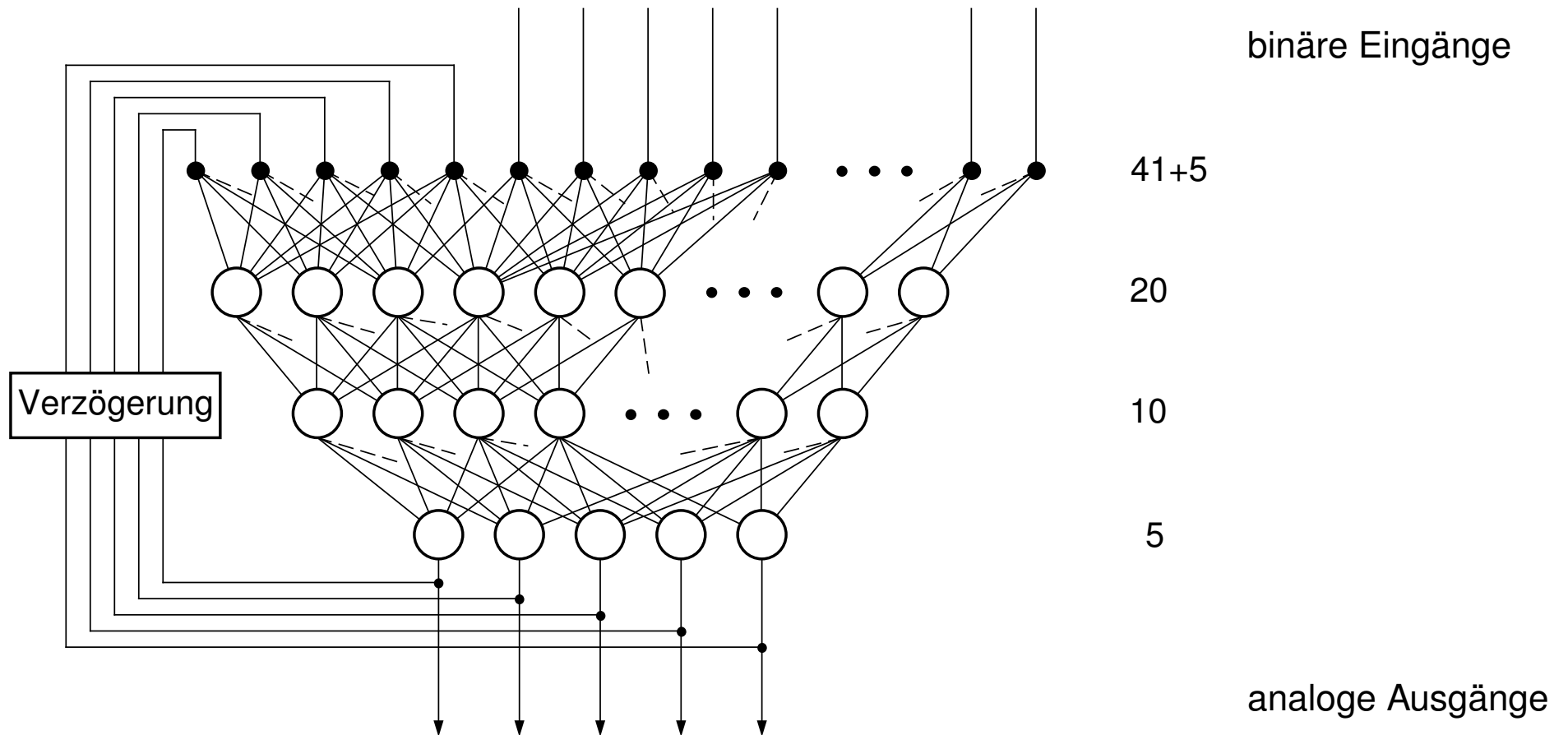
Grundfrequenzgenerierung mit neuronalem Netz

Methode: F_0 wird pro Silbe generiert

Achtung: F_0 darf an Silbengrenzen nicht springen und muss pro Phrase eine angemessene Deklination realisieren!

Folge: Netz muss F_0 einer Nachbarsilbe kennen!

Neuronales Netz für chronologische F_0 -Generierung



Übertragung der Grundfrequenz auf die Laute

Aus F_0 -Steuerung gegeben: $F_0(j, k)$ d.h. K Werte pro Silbe

Für Signalproduktion nötig: F_0 pro Laut

→ Zusammenhang $F_0(j, k) \longleftrightarrow F_0(t)$ ermittelbar über die Dauer der Laute

Sprachsignale mit natürlichen bzw. synthetischen F_0 -Verläufen

- 1: Das bis auf weiteres gültige Lawinenbulletin vom 15. April.
- 2: Am Alpensüdhang ist die Lawinengefahr weiterhin gering.
- 3: Am Alpennordhang, im Wallis und Graubünden ist jedoch immer ...
- 4: Er führt theologische, rechtliche und praktische Gründe ins Feld.
- 5: Die Caritas Schweiz sucht Freiwillige für Einsätze in ...
- 6: Chadli ist der erste algerische Staatschef, der die USA besucht.
- 7: Im Mittelpunkt seiner Gespräche stehen Rüstungs- und ...



Sprachsignale mit natürlichen bzw. synthetischen F_0 -Verläufen

1: Das bis auf weiteres gültige Lawinenbulletin vom 15. April.



2: Am Alpensüdhang ist die Lawinengefahr weiterhin gering.



3: Am Alpennordhang, im Wallis und Graubünden ist jedoch immer ...



4: Er führt theologische, rechtliche und praktische Gründe ins Feld.



5: Die Caritas Schweiz sucht Freiwillige für Einsätze in ...



6: Chadli ist der erste algerische Staatschef, der die USA besucht.







7: Im Mittelpunkt seiner Gespräche stehen Rüstungs- und ...



Steuerung der Intensität

Sprachsynthesysteme mit Verkettungsansatz haben
gewöhnlich **keine Intensitätssteuerung**

Kurze Begründung:

- Falls Diphone aus subjektiv gleich lauten Sprachsignalen,
ist das Fehlen der Intensitätssteuerung ist nicht störend
(fehlende Dauer-  oder F_0 -Steuerung  stört) 
- Nur eine sehr gute Intensitätssteuerung verschlechtert das Signal nicht!
Sie muss u.a. die gehörte Lautheit berücksichtigen 

(ausführliche Begründung: Buch Abschnitt 9.3.3)

Zusammenfassung

Für die Sprachsynthese mittels Verkettung von Segmenten gilt:

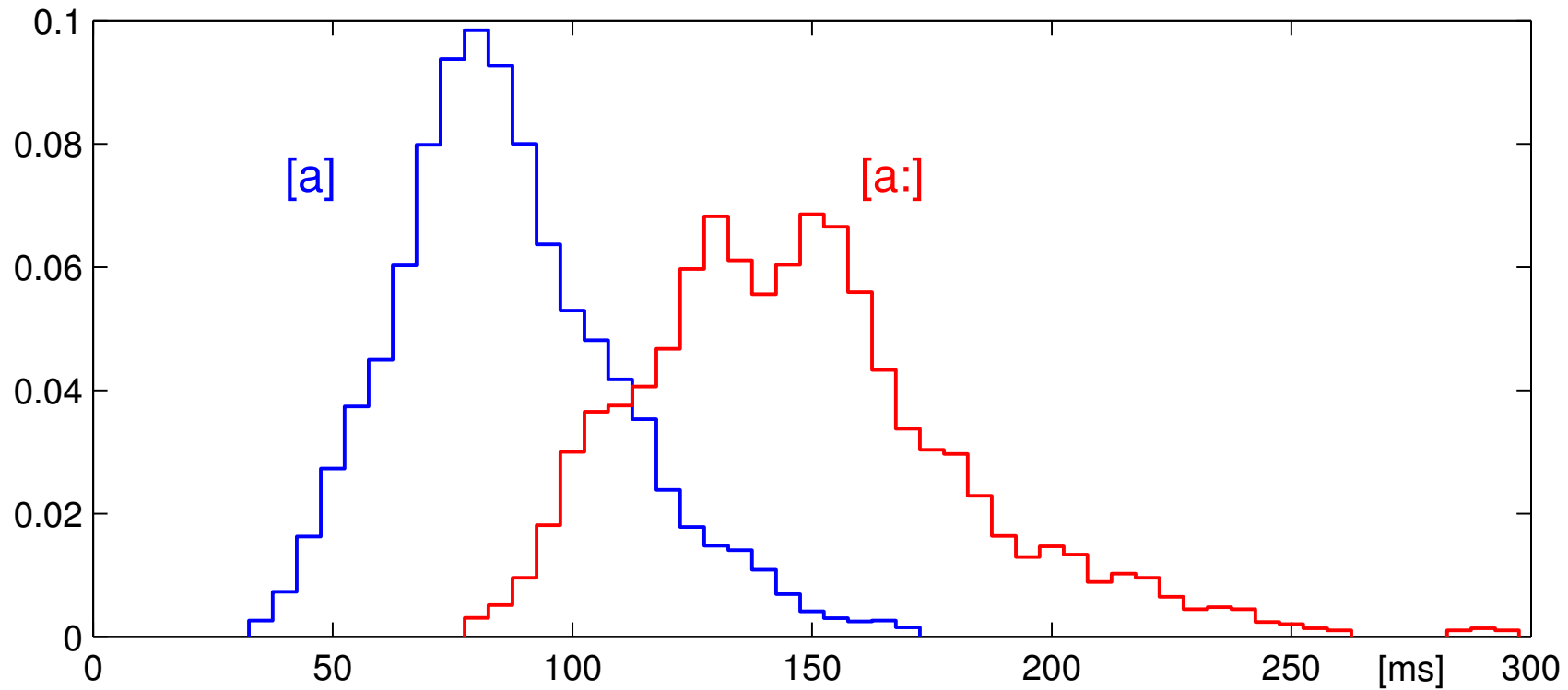
- Dauer- und Grundfrequenzsteuerung sind wichtig
(Intensitätssteuerung ist nebensächlich)
- viele linguistische und ausserlinguistische Einflussfaktoren
- nicht-linearer Ansatz zu Steuerung besser (z.B. neuronales Netz)

Thema der nächsten Lektion:

Fourier-Analyse-Synthese von Sprachsignalen

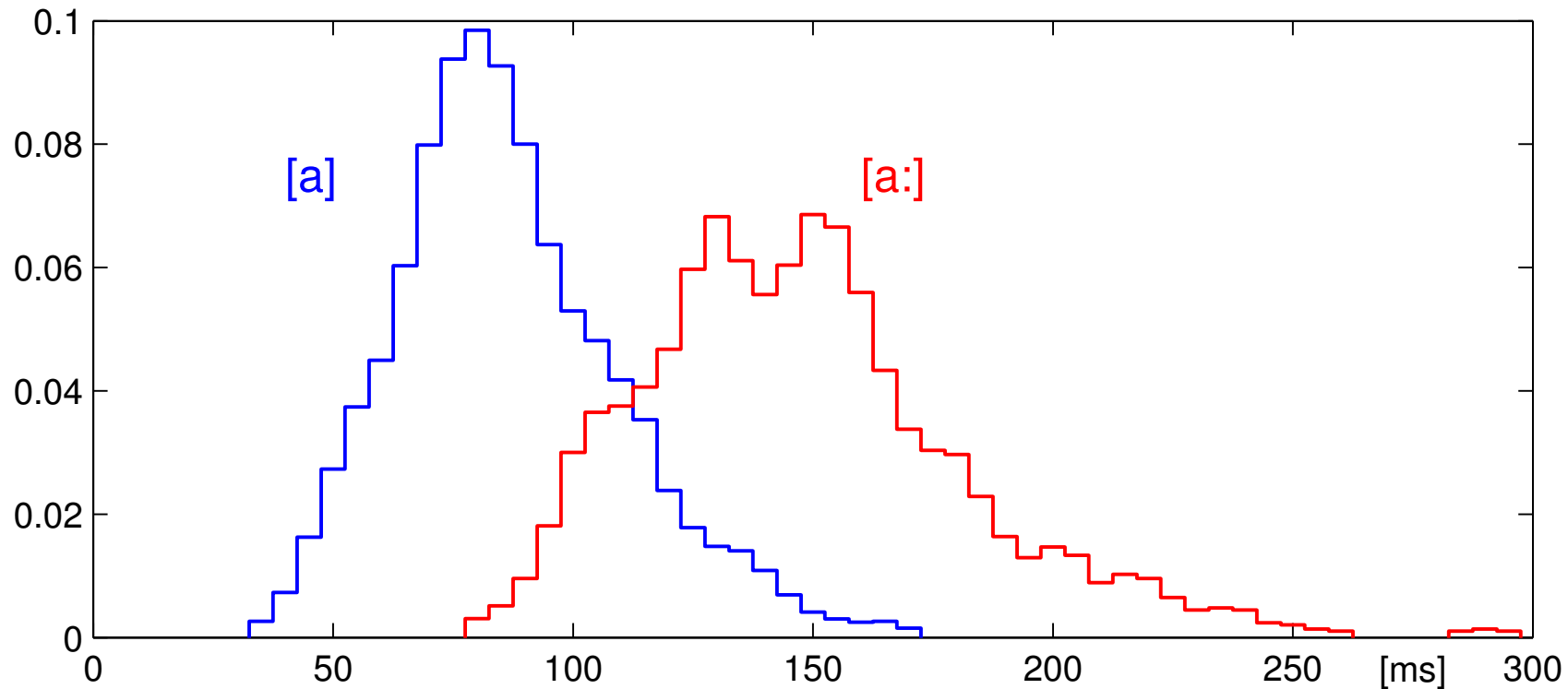
Zur Übersicht der Vorlesung *Sprachverarbeitung I* >>>

Normierte Lautdauer-Histogramme



Was kann man aus diesen gemessenen Lautauern schliessen ?

Normierte Lautdauer-Histogramme



→ [a] und [a:] sind anhand der Dauer nicht eindeutig unterscheidbar!

Aber: Der Faktor “Langvokal” beeinflusst die Lautdauer!

<<<

Mängel des linearen Ansatzes zur Dauersteuerung

Linearer Ansatz für gewisse Lautdauern schlecht

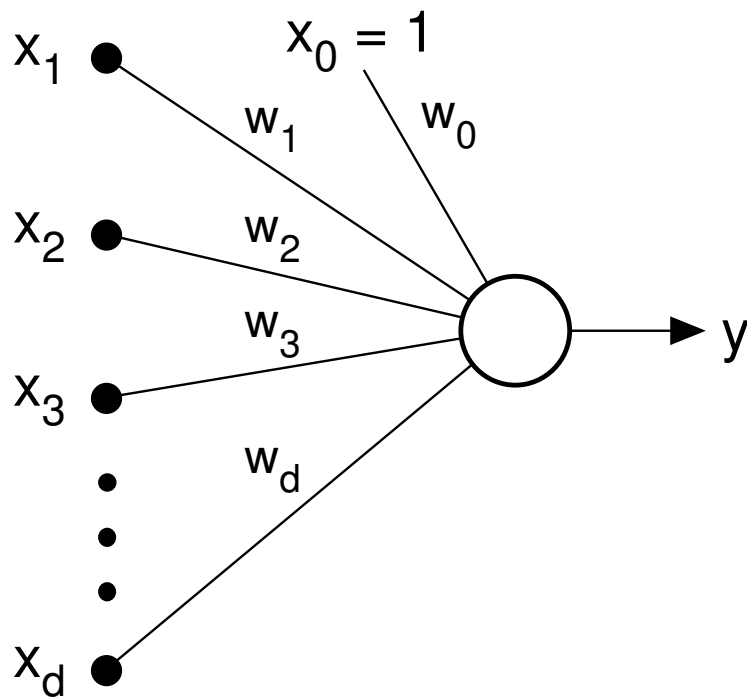
(auch wenn alle möglichen Einflussfaktoren berücksichtigt werden)

Wichtigste Gründe:

- Kombination mehrer Faktoren ist nicht linear
(Effekt mehrerer Einflussfaktoren ungleich Summe der Einzeleffekte)
- Häufig auftretende Faktoren überstimmen die seltenen
(wegen Minimierung des globalen Schätzfehlers)

<<<

Neuron mit d Eingängen und einem Ausgang



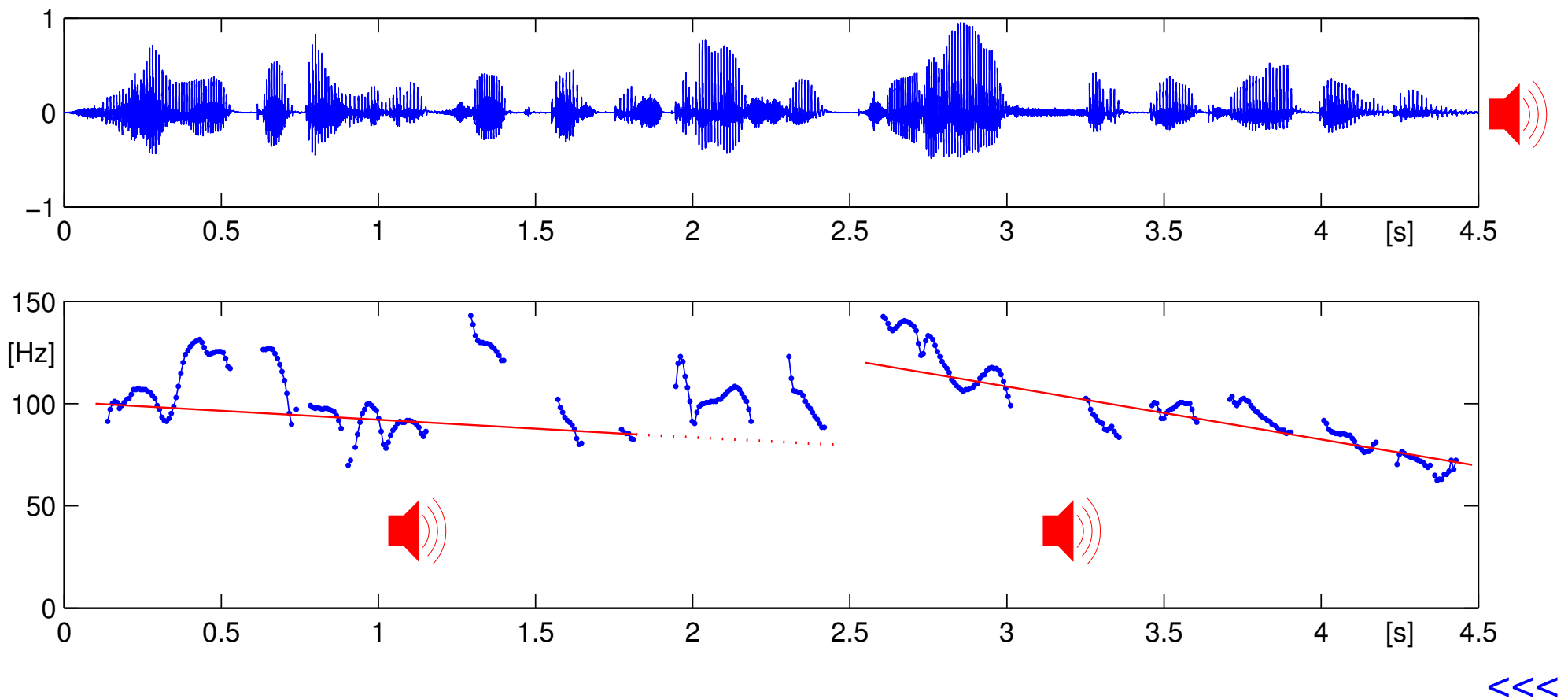
$$y = f(z) = f \left(\sum_{i=0}^d w_i x_i \right)$$

$f(z)$ ist eine nicht-lineare Funktion

<<<

Grundfrequenzverlauf eines natürlichen Sprachsignals

“Sie erhielten bei ihrem Zug durch die Strassen Zulauf von Pekingern.”



Einflüsse auf die Grundfrequenz

linguistischer Faktor	diskrete Werte
Satztyp:	Aussage Befehl Frage
Phrasentyp:	initial progredient terminal
Phrasengrenze:	Stärkegrad: 1 2 3 sonst
Silbenakzent:	Stärkegrad: 1 2 3 unakzentuiert sonst
Silbentyp:	Silbenkern (Nucleus): Lang- Kurzvokal Diphthong offener geschlossener Laut
	Anfangskonsonanten: stimmhaft stimmlos keine
	Endkonsonanten: stimmhaft stimmlos keine
Silbenlage:	vor nach dem Phrasenhauptakzent

(ausserlinguistische Einflüsse: Deklination, Männer-/Frauenstimme etc.)

<<<

Erstellen der Tabelle mit F_0 -Werten und zugehör. Einflussfaktoren

Gegeben: Sammlung von Sprachsignalen

1. Schritt: Phonologische Darstellung der Sätze ermitteln

(P) zi:- Er-[1]hi:l-t@n- #{2} (P) ba_i- i:-r@m- [1]tsu:k- #{2}
(P) dUrC- di- [1]Stra:-s@n- #{1} (P) [1]tsu:-la_uf- #{2}
(T) fOn- [2]pe:-kI-N@r- [1]bYr-g@rn.

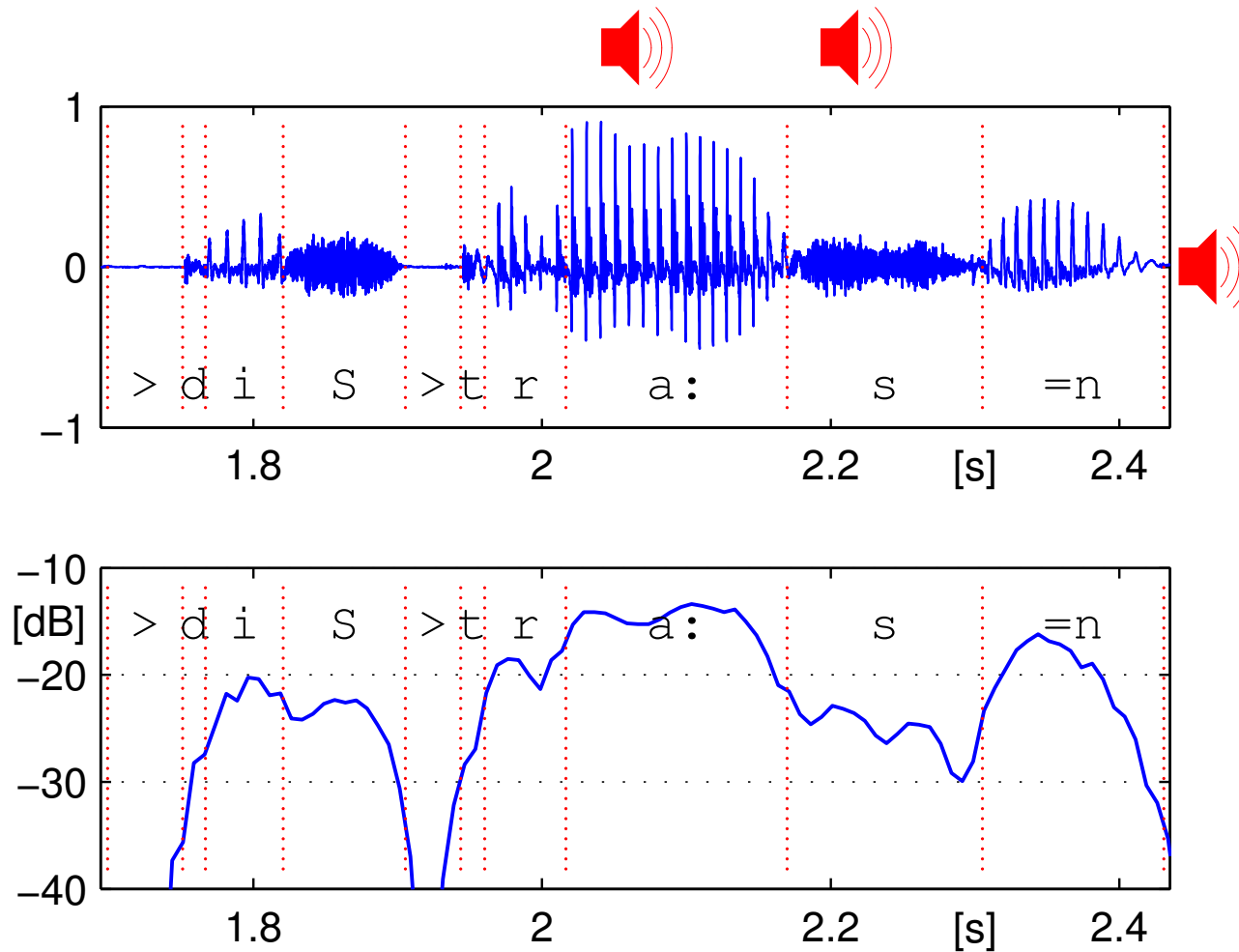
2. Schritt: F_0 -Wert jedes Silbenkerns messen

3. Schritt: F_0 -Wert und Einflussfaktoren pro Silbe als Zeile in Tabelle eintragen

F_0	Phrasentyp		Silben-	Akzent	Akzent	...
	prograd.	terminal	nummer	1	2	
98.5	1	0	1	0	0	
104.9	1	0	2	0	0	
132.8	1	0	3	1	0	
125.0	1	0	4	0	0	
90.1	1	0	5	0	0	
90.2	1	0	6	0	0	
89.8	1	0	7	0	0	
136.7	1	0	8	1	0	
95.1	1	0	9	0	0	
87.3	1	0	10	0	0	
109.4	1	0	11	1	0	
110.7	1	0	12	0	0	
140.6	0	1	1	1	0	
112.3	0	1	2	0	0	
95.7	0	1	3	0	0	
97.5	0	1	4	0	1	
100.9	0	1	5	0	0	
89.1	0	1	6	0	0	
89.8	0	1	7	1	0	
75.0	0	1	8	0	0	

<<<

Gemessene Signalleistung vs. wahrgenommene Lautheit



<<<

