

Sprachsynthese: Prosodie-Modellierung

Uwe Reichel
Institut für Phonetik und Sprachverarbeitung
Ludwig-Maximilians-Universität München
reichelu@phonetik.uni-muenchen.de

10. Januar 2017

Inhalt

- Prosodische Struktur
 - Phrasengrenzen
 - Akzente
- Intonation
 - Tonsequenzmodell
 - Fujisaki-Modell
- Segmentdauern
 - Klatt-Modell

Prosodische Struktur

Phonetische Korrelate

- **Phrasengrenzen**
 - Pausen
 - *Prefinal lengthening*
 - *Pitch Reset*
- **Akzente**
 - F0-Bewegung
 - Längung
 - erhöhte Intensität

Vorhersage von Phrasengrenzen: Interpunktion, Wortarten

Interpunktion

- Satzzeichen → Phrasengrenze
- v.a. bei Lesesprache adäquat

Wortart: Chink-Chunk

- Liberman et al. (1992)
- Einteilung der Wortarten danach, ob sie überwiegend phraseninitial auftreten (*Chinks*: Determiner, Konjunktionen, Präpositionen ...) oder nicht (*Chunks*: Nomen, Verbpartikeln, ...)
- prosodische Phrase: *CHINK* CHUNK**

Vorhersage von Phrasengrenzen: Syntax

Syntaktische vs. prosodische Phrasen

- Prosodische Struktur i.d.R. **flacher** als syntaktische Struktur

*This is [the cat that caught [the rat that stole [the
cheese]_{NP}]_{NP}]_{NP}.*

*[This is the cat]_{IP} [that caught the rat]_{IP} [that stole
the cheese]_{IP}.¹*

- Suche nach einer prosodisch motivierten syntaktischen Struktur

¹aus Nespors & Vogel, 1986; NP: Nominalphrase, IP: Intonationsphrase

Vorhersage von Phrasengrenzen: Syntax

Performance-Struktur (Gee et al., 1983; Bachenko et al., 1990)

- empirisch abgeleitet von gemittelten Pausendauern zwischen benachbarten Wörtern
- ϕ -Phrase: Segmentierung des Satzes hinter jedem Inhaltswort, das den Kopf einer syntaktischen Konstituente bildet
- Ausnahmen: attributive Adjektive
- **Beispiel:**

[John] [asked] [*the strange young man*] [*to be quick*] [*on the task*]

Vorhersage von Phrasengrenzen: Syntax

Chunk-Parser (Abney, 1991)

- **Chunks:** *major head*-fähige Inhaltswörter + zugehörige Funktionswörter + dazwischenliegende Inhaltswörter
- attributive Adjektive sind nicht *major head*-fähig

- **Beispiel:**

[**John**] [**asked**] [*the strange young man*] *and*
[**nodded**]

- Eingliederung unverknüpfter Wörter (**orphan nodes**; '*and*' im obigen Beispiel) in den nachfolgenden Chunk → ϕ -Phrasen

Vorhersage von Phrasengrenzen: Statistische Ansätze

Statistischer Ansatz (Taylor&Black, 1998)

$$\begin{aligned}\hat{G} &= \arg \max_G [P(G|W)] \\ &= \arg \max_G [P(W|G) \cdot P(G)]\end{aligned}\quad (1)$$

- analog zu *Noisy-Channel-Modell* (vgl. *POS-Folien*)
- G : (binäre) Sequenz von Grenzlabels (ein Label je Wort; “Grenze folgt”, “folgt nicht”)
- W : POS-Sequenz
- Suche nach der Grenzlabelsequenz, die der beobachteten POS-Sequenz am wahrscheinlichsten zugrundeliegt

Vorhersage von Akzenten: Fokus

Fokus

- **Informationszentrum** eines Satzes
- z.B. neue Information, Kontrastierung
- **weiter vs. enger Fokus**:
 - *Was gibt's Neues? [Connie hat ihrem Sohn die **Haare** geschnitten]. (Bem.: **All-New-Satz, neutraler Akzent**)*
 - *Was hat Connie getan? Connie hat [ihrem Sohn die **Haare** geschnitten].*
 - *Was hat Connie mit ihrem Sohn gemacht? Connie hat ihrem Sohn [die **Haare** geschnitten].*
 - *Was hat Connie ihrem Sohn geschnitten? Connie hat ihrem Sohn [die **Haare**] geschnitten.*
- hier bleibt **Fokusexponent** (Akzentträger) unverändert

Vorhersage von Akzenten

Wortarten

- Funktionswörter sind i.d.R. unakzentuiert

Syntaktisch-phonologische Ansätze

- **metrische Phonologie** (Lieberman, 1975; Lieberman et al., 1977)
- Vorhersage des **neutralen Akzents** (bei weitestmöglichem Fokus)

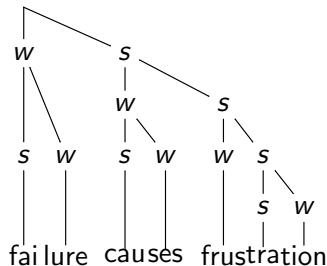
geg. (deutsche) Konstituente $[AB]_C$:

Nucleus Stress Rule (NSR): Ist C eine Phrase, so ist B *strong* s.

Compound Stress Rule (CSR): Ist C ein Wort oder Teil eines Wortes, so ist B *strong*, wenn es sich weiter verzweigt, ansonsten ist A *strong* und B *weak* w

Vorhersage von Akzenten: Syntax

- Akzentuierungsverhältnisse in Form **metrischer Bäume** oder **metrischer Gitter**



Vorhersage von Akzenten: Semantik

Semantische Ansätze

- Existenz eines **neutralen Akzents**, der sich anhand der Syntax vorhersagen lässt (vgl. *NSR*), wird bestritten (Bolinger, 1972):

Er versteht es, Erklärungen zu ignorieren.

*Er versteht es, **Erklärungen** zu geben.*

- Akzentuierung eines Wortes durch sein **semantisches Gewicht** (= relative Vorhersagbarkeit aus dem Kontext)
- Fülle an Einflußfaktoren → '*Accent is predictable (if you're a mind reader)*'

Vorhersage von Akzenten: Semantik

Statistische Modellierung des semantischen Gewichts

- **n-Gramm-Wahrscheinlichkeiten** eines Worts w_i (z.B. Pan et al., 2000)
- **Bezug:** je geringer die n-Gramm-Wahrscheinlichkeit eines Worts, desto weniger vorhersagbar, desto wahrscheinlicher akzentuiert
- $P(w_i)$: **globale** (kontextunabhängige) Vorhersagbarkeit
- $P(w_i|w_{i-1})$: **lokale** (kontextabhängige) Vorhersagbarkeit

Vorhersage von Akzenten: Diskurs

Neue vs. gegebene Information

- **neue** Information akzentuiert, **gegebene** Information nicht
- **gegebene Information:**
 - im bisherigen Diskursverlauf bereits übermittelt
 - zum von Sprecher und Hörer geteilten Weltwissen gehörig
 - aus dem situativen Kontext erschließbar.

Vorhersage von Akzenten: Diskurs

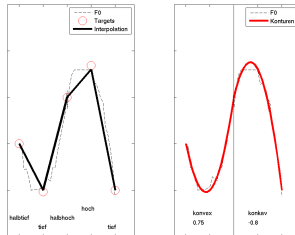
Verfahren nach Hirschberg (1993)

- *given*: globaler und lokaler Hintergrund
- **global**: grundsätzliches Thema, repräsentiert durch Liste der Inhaltswörter des ersten Satzes; über gesamten Text gültig
- **lokal**: *queue* von Hintergrundbereichen (Inhaltswörter eines Satzes), die durch *push*- und *pop*-Operationen laufend aktualisiert wird
- Operationen durch Interpunktion und Diskursmarker gesteuert; an Absätzen wird der lokale Hintergrund komplett geleert
- Wörter, die sich bereits im globalen oder lokalen Hintergrund befinden, werden **deakzentuiert**

Intonationsmodelle

Dichotomien

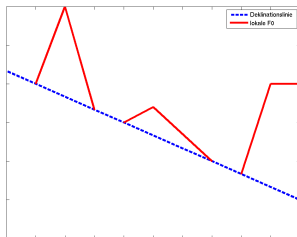
- 1 **Tonbasiert vs. konturbasiert**
- 2 **Symbolisch vs. parametrisch**



Links: F0-Verlauf als Abfolge symbolisch etikettierter Ton-Targets. **Rechts:** F0-Verlauf als Abfolge von Konturen mit symbolischen Etiketten bzw. parametrischen Krümmungskoeffizienten der Stilisierungsparabeln.

Intonationsmodelle

3 Einschichtig vs. Superpositional

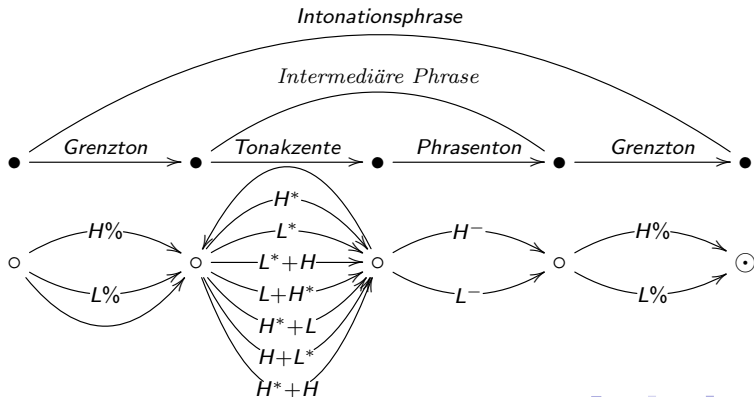


Superpositionale Darstellung des F0-Verlaufs als Überlagerung von globaler Deklinationslinie und lokalen F0-Bewegungen.

Tonsequenzmodell

Tonbasiert

- Intonation als Abfolge von Tönen



Tonsequenzmodell

- **elementare Töne:** H, L
- **komplexe Töne:** $H + L^*$ etc.
- * verknüpft Ton mit der akzentuierten Silbe
- **Grenztöne:**
 - ‘%’ am Rand von Intonationsphrasen (*progredienter* vs. *finaler* F0-Verlauf)
 - ‘—’ am Ende von intermediären Phrasen

Tonsequenzmodell

Vorhersage der Töne

- **kompositionales Modell** nach Pierrehumbert&Hirschberg (1990)
- Informationsstatus \rightarrow Tonakzent:
 - neue Information, Hervorhebung $\rightarrow H^*, L + H^*$
 - gegebene Information, Inferierbarkeit $\rightarrow L^*, H + L^*$
- Orientierung der aktuellen Intonationsphrase im Diskurs \rightarrow Grenztöne
 - final $\rightarrow LL\%$; progredient $\rightarrow LH\%$

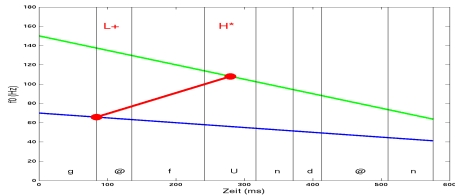
Tonsequenzmodell

Symbolische Beschreibung

→ zusätzliche Modelle zur Überführung der Tonsymbole in F0-Werte nötig

- **regelbasiert** (Anderson, 1984; Jilka , 1999):
 - Ermittlung von F0-Zielwerten relativ zum Silbennukleus und relativ zu Topline und Baseline
 - **Faktoren:** Akzenttyp, metrische Prominenz der assoziierten Silbe, Position innerhalb der Phrase, Phrasenlänge, vorangehende F0-Werten

Tonsequenzmodell



F0-Kontur für $L + H^*$: L auf Grundlinie zu Beginn des Nukleus der präakzentuierten Silbe; H^* auf Toplinie in der Mitte des Nukleus der akzentuierten Silbe; Beispiel nach Schröder&Trouvain (2003).

Tonsequenzmodell

- **statistisch** (Black, 1996)
 - drei F0-Werte y je tonal markierter Silbe
 - Berechnung mittels Regressionsanalyse:

$$y = c_0 + \sum_i c_i \cdot p_i \quad (2)$$

- Prädiktoren p_i : Tonlabel, Grenzlabel, Wortbetonung, Position der Silbe in der Intonationsphrase
- binäre Codierung kategorialer Prädiktoren

Tonsequenzmodell

Einschichtig

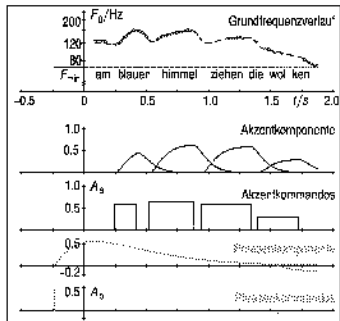
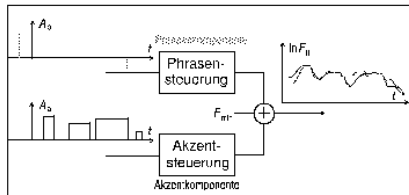
- Modellierung globaler Phänomene als Abfolge lokaler Ereignisse
- Deklination repräsentiert als Sequenz von *Downsteps*
- keine Vorausplanung modellierbar (Tonsequenz in Form eines endlichen Automaten)

Fujisaki-Modell

Aufbau

- Fujisaki (1987)
- konturbasiert, superpositional, parametrisch
- Superposition der Outputs von **Phrasenkomponente** $C_p(t)$, **Akzentkomponente** $C_a(t)$ und **Baseline-F0** F_{min}
- $C_p(t)$ und $C_a(t)$ als **kritisch gedämpfte Systeme** realisiert (d.h. sie schwingen bei der Rückkehr in die Ruhelage nicht darüber hinaus)
- Anregung durch **Phrasenkommandos** A_p (Impuls), bzw. **Akzentkommandos** A_a (Rechteckfunktion)

Fujisaki-Modell



Fujisaki-Modell

Phrasenkomponente

- globale Intonationskontur
- positives A_p markiert den Beginn einer Intonationsphrase (*pitch reset*) oder am Ende Frageintonation, bzw. Nicht-Finalität der Phrase
- negatives A_p am Ende der Phrase markiert Finalität (*final lowering*)

Akzentkomponente

- lokale F₀-Bewegungen auf akzentuierten Silben
- A_a -Amplitude nimmt im Verlauf der Phrase ab (im Zuge der Deklination fallende Topline)

Fujisaki-Modell

Stilisierung

$$\ln F_0(t) = \ln F_{\min} + \sum_i A_{pi} C_p(t - T_{pi}) + \sum_j A_{aj} [C_a(t - T_{1j}) - C_a(t - T_{2j})] \quad (3)$$

$$C_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & : t \geq 0 \\ 0 & : t < 0 \end{cases} \quad C_a(t) = \begin{cases} 1 - (1 + \beta t) e^{-\beta t} & : t \geq 0 \\ 0 & : t < 0 \end{cases}$$

- T_p : Zeitpunkt des Phrasenkommandos,
- T_1, T_2 : Start- und Endzeitpunkt des Akzentkommandos,
- A_p, A_a : Amplituden der Kommandos,
- α, β : Dämpfungsfaktoren des Phrasen- und Akzentsystems, die die Dauer der F0-Bewegungen mitbestimmen.

Fujisaki-Modell

Schätzung der Parameter: Analyse durch Synthese

- **Analyse** der F0-Kontur **durch Synthese** mit dem Fujisaki-Modell
- Mixdorff (2002):
 - Hochpassfilterung der geglätteten und interpolierten Kontur zur Trennung von Phrasen- (nieder-) und Akzentkontur (hochfrequent)
 - getrennte Anpassung der Parameter an jeweilige Kontur (Akzentkontur für Akzentkomponente, Phrasenkontur für Phrasenkomponente) mit Gradientenverfahren
 - weitere Feinanpassung der Parameter an komplette Kontur

Fujisaki-Modell

Vorhersage der Parameterwerte

- **Möbius (1993, 1995)**
 - regelbasiert
 - **Faktoren:** u.a. Satzmodus, Position in Intonationsphrase, Wortart
- **Mixdorff (1998)**
 - regelbasiert
 - Textsegmentierung in **Intonemsegmente** nach Isacenko (1964)
 - $I \downarrow$: **Informationsintonem:** fallend, Abschluss einer Informationseinheit
 - $N \uparrow$: **Nonterminalitäts-Intonem:** steigend, Nichtabgeschlossenheit der Äußerung
 - $C \uparrow$: **Kontaktintonem:** steigend, Kontaktaufnahme mit dem Hörer beispielsweise (z.B. Frage)

Fujisaki-Modell

Probleme bei der Interpretation

- Kein eindeutiger Zusammenhang zwischen Parameterwerten und Verlauf der generierten Kontur (unterschiedliche Parameterbelegungen können zur selben Kontur führen) → textbasierte Vorhersage erschwert
- Modell kann jede Kontur mit Nullabweichung beschreiben, wenn Akzentkommandos nur eng genug aufeinanderfolgen; linguistisch dann nicht mehr interpretierbar

Einflussfaktoren auf Segmentdauern

Prosodische Struktur

- Akzente: Segmentlängung
- präfinale Längung an Phrasengrenzen

Lautkontext

- **intrinsische Lautdauern:**
bei hohen Vokalen kürzer als bei tiefen
bei stimmhaften Konsonanten kürzer als bei stimmlosen
- **ko-intrinsische Lautdauern:**
Vokaldauer kürzer vor stimmlosen Obstruenten als vor stimmhaften

Dauer-Modellierung

Klatt (1979)

$$D = m \cdot D_{min} + \prod_i f_i \cdot (D_{inh} - m \cdot D_{min}) + d \quad (4)$$

- **Parameter:**

D : aktuelle Lautdauer

D_{inh} : inhärente Lautdauer

D_{min} : minimale Lautdauer (abhängig von Kompressionsfähigkeit)

m, f_i, d : Faktoren, deren Werte über Regeln zu bestimmen sind (Default 1)

Dauer-Modellierung

- **Faktoren:**
 - Lautkontext
 - Wortbetonung, Akzent
 - Position in Silbe, Wort, Intonationsphrase
- **Regelbeispiele (experimentell ermittelt):**
 - Nichtfinale Kürzung: Der Silbenkern jeder Silbe in nichtfinaler Stellung wird verkürzt; $f_3 := 0.6$.
 - Kürzung in nicht-initialer Stellung. Konsonanten, die nicht am Wortanfang stehen, werden verkürzt: $f_6 := 0.85$.