Intonation in der Sprachsynthese

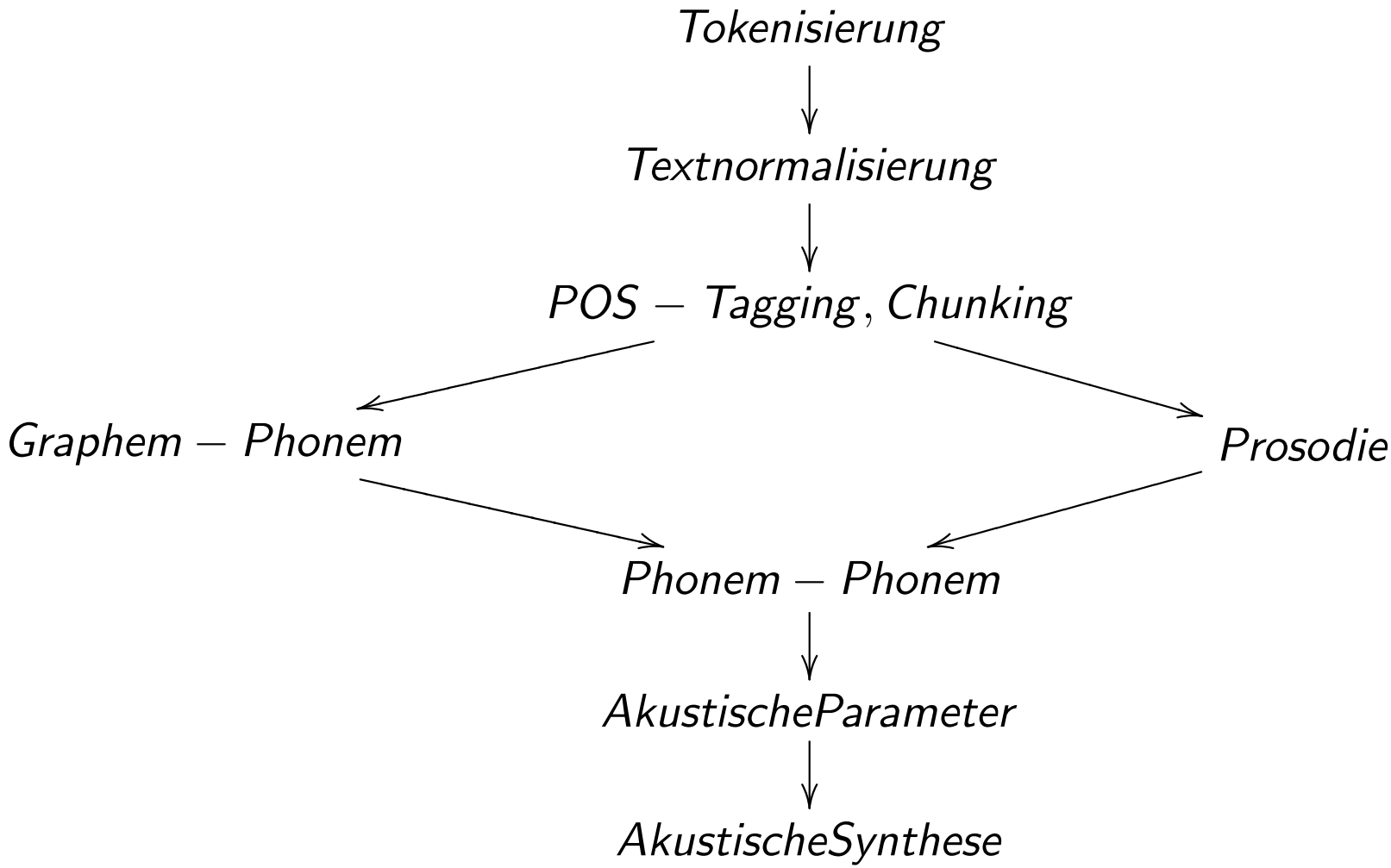
Uwe Reichel
Institut für Phonetik und Sprachverarbeitung
Ludwig-Maximilians-Universität München
reichelu@phonetik.uni-muenchen.de

1. Dezember 2010

Das MARY TTS-System

- **TTS:** Text-to-Speech
- **MARY:**
  - **M**odular **A**rchitecture for **R**esearch on Speech S**y**nthesis
  - entwickelt am DFKI, Saarbrücken
  - Download, Dokumentation: *http://mary.dfki.de*
  - Anwendung über Webserver: *http://marytts:59125*
  - Stand der Folien: Schröder, M. & Trouvain, J. (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *J. Speech Technology, 6*, pp. 365–377.
  - mittlerweile Erweiterung hinsichtlich emotionaler Synthese

MARY-TTS-Module



Tokenisierung, Textnormalisierung

Tokenisierung

- Zerlegung des Texts in Wörter und Satzzeichen
- Regelbasierte Disambiguierung des Punkts (Satzende vs. Ordinalzahl, Abkürzung, usw.)

Textnormalisierung

- Expansion von Zahlen (Jahreszahl vs. Telefonnummer usw.)
- kontextabhängige Flektion von Ordinalzahlen
- Table-Lookup: Expansion von Abkürzungen, Akronymen

POS-Tagging

Allgemeine Aufgabenstellung

- Schätzung der wahrscheinlichsten Wortart-Sequenz $\hat{G} = g_1 \ldots g_n$, gegeben die beobachtete Wortfolge $W = w_1 \ldots w_n$

$$\hat{G} = \arg\max_G \Big[P(G|W) \Big]$$

- Umformung unter Zuhilfename des **Satzes von Bayes** und vereinfachender Annahmen:

$$\hat{G} = \arg\max_G \left[\frac{P(G)P(W|G)}{P(W)} \right]$$

$$= \arg\max_G \left[\prod_{i=1}^{n} P(g_i|g_{\text{vorgänger}}) P(w_i|g_i) \right]$$

POS-Tagging

TNT-Tagger

- Brants (2000)
- Wenn w_i unbekannt (**Out-of-Vocabulary OOV**): Verwendung der w_i-Suffixe, die im Deutschen Aufschluss über die Wortart geben können
- *Umgeh***ung**, *Blau***wal**, *farb***ig**

Chunking

- Flache syntaktische Analyse als Grundlage für prosodische Phrasierung
- Parser von Skut&Brants (1998)
- Grenzen von Nominal- und Präpositionalphrasen

$$[\textit{Der Ball}]_{NP}\ \textit{blieb}\ [\textit{auf der Torlinie}]_{PP}\ \textit{liegen.}$$

Graphem-Phonem-Konvertierung

Lexika

- G2P-Lexikon für **Simplex-Formen**
- G2P-Lexikon für **gebundene Morpheme** (Affixe, usw.)

Konvertierung

- morphologische Zerlegung $\longrightarrow$ Simplex-Formen + gebundene Morpheme
- Lexikon-Lookup
- bei **OOVs**: **regelbasierte** G2P-Konvertierung, Silbifizierung, Wortbetonungszuweisung (Kompositumstruktur, betonte Affixe, usw.)

Prosodische Struktur

Prosodische Grenzen

- 6 Grenzstärken
- an Interpunktion
- zwischen *Vorfeld* und *linker Verbklammer*

$$[\textit{die Frau}]_{VF} \mid [\textit{ruft}]_{LK}\ \textit{ihren Hund}$$

- vor satzverbindenden Konjunktionen
- wahlweise (in Abhängigkeit des gewünschten Sprechstils) an Chunk-Grenzen

$$\textit{er half} \mid [\textit{dem Mann}]_{NP} \mid [\textit{in den Mantel}]_{PP}$$

Prosodische Struktur

Akzente

- einige POS stets akzentuiert, z.B. Substantive und Adjektive
- weitere POS hinsichtlich Akzentuierbarkeit geordnet: Vollverben $>$ Modalverben $>$ Adverben
- Vorgehen:
  - Akzentuiere in einer prosodischen Phrase alle Substantive und Adjektive
  - falls nicht vorhanden, suche nach akzentuierbarem Material in oben gegebener POS-Reihenfolge

Der Hund | liegt | auf der grünen Bank

Tonakzente, Phrasen-, Grenztöne

- GTOBI-Inventar
- Tonzuweisung in Abhängigkeit des Satztyps (Deklarativsatz, W-Frage, Interrogativsatz, Entscheidungsfrage, Exklamativsatz)

Mögliche Erweiterungen (gemäß kompositionalem Modell nach Pierrehumbert&Hirschberg, 1990):

- Informationsstatus $\longrightarrow$ Tonakzent:
  - neue Information, Hervorhebung $\longrightarrow$ H^*, $L + H^*$
  - gegebene Information, Inferierbarkeit $\longrightarrow$ L^*, $H + L^*$
- Orientierung der aktuellen Intonationsphrase im Diskurs $\longrightarrow$ Grenztöne
  - final $\longrightarrow$ LL%; progredient $\longrightarrow$ LH%

F0-Konvertierung

Regelbasierte F0-Vorhersage: Positionierung der Targets

- **zeitlich** relativ zum Silbennukleus
- in ihrer **Frequenz** relativ zu Deklinationsgrundlinie und Toplinie

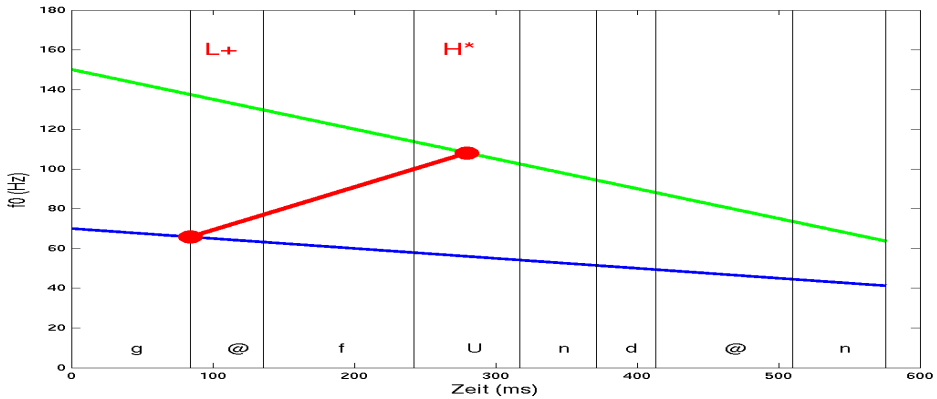


Abbildung: F0-Kontur für $L + H^*$: L auf Grundlinie zu Beginn des Nukleus der präakzentuierten Silbe; H^* auf Toplinie in der Mitte des Nukleus der akzentuierten Silbe; Beispiel nach Schröder&Trouvain (2003).

Dauer-Modellierung

Klatt-Modell (Klatt, 1979)

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

- **Parameter:**
  D: aktuelle Lautdauer
  D_{inh}, D_{min}: inhärente und minimale Lautdauer
  m, f_i, d: Faktoren, deren Werte über Regeln zu bestimmen sind (Default 1)
- **Faktoren:** Lautkontext; Wortbetonung, Akzent; Position in Silbe, Wort, Intonationsphrase

Phonem-Phonem-Konvertierung

- Regelbasierte Assimilationsoperationen
- Lautreduktionen in unbetonten Silben

Unit-Selection

Zur Auswahl in MARY

- Unit Selection
- HMM-Synthese

Im Folgenden Vorstellung des konkatenativen Unit-Selection-Ansatzes (am Beispiel von Diphonen)

Konkatenative Synthese

- **Konkatenative Synthese:** Verkettung von akustischen Segmenten

Diphon

- Segment von der Mitte eines Phons bis zur Mitte des folgenden Phons
- Berücksichtigung lokaler **koartikulatorischer** Effekte
- minimale **Inventargröße:** (Anzahl der Phoneme)^2 – (Anzahl phonotaktisch nicht erlaubter Kombinationen)

Konkatenative Synthese

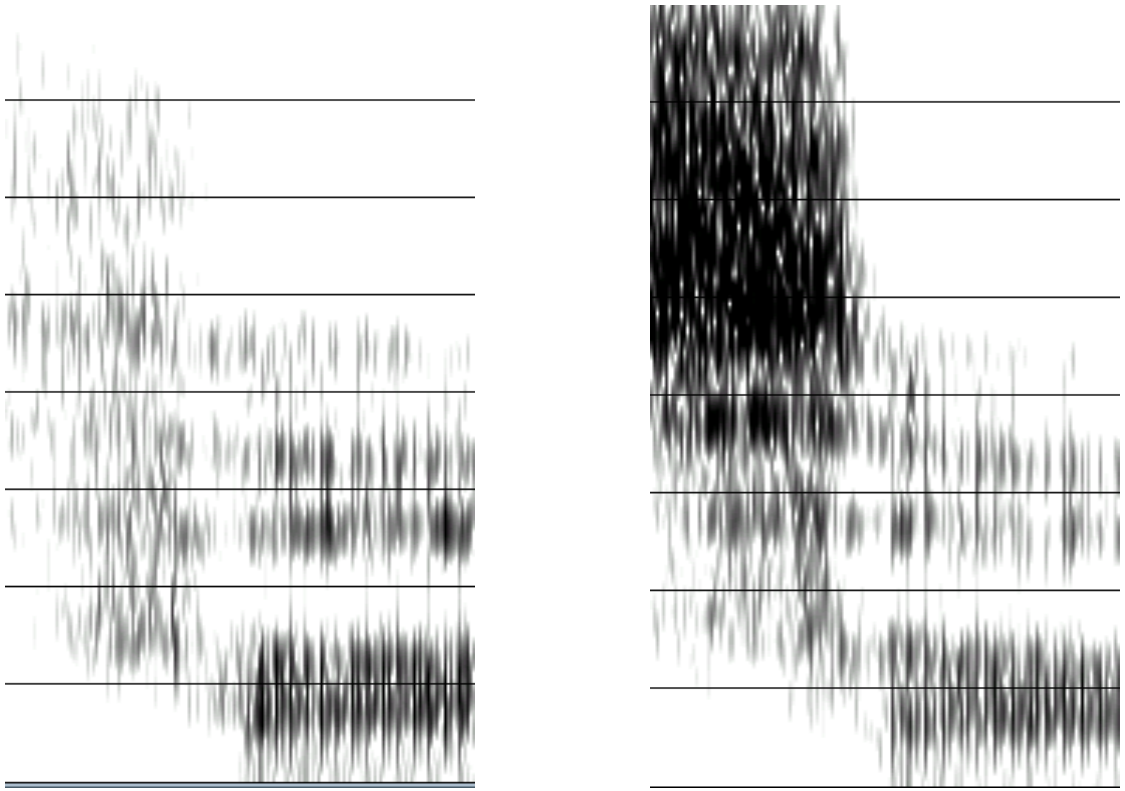


Abbildung: Diphone /fa/ und /sa/: unterschiedliche Formanttransitionen.

Konkatenative Synthese

2 Philosophien

- **Klassische Diphonsynthese**
  - **Datenbank:** geringe Menge gespeicherter Units (z.B. jedes Diphon 2x +/– phrasenfinal)
  - **Synthese: Signalmanipulation** bei Verkettung
- **Eigentliche Unit-Selection-Synthese**
  - **Datenbank:** große Menge gespeicherter Units (Diphone in vielen verschiedenen Kontexten, +/–akzentuiert, +/– phrasenfinal, unterschiedliches Sprechtempo, unterschiedliche emotionale Markierung, . . .)
  - **Synthese: kontextabhängige Auswahl** der geeigneten Unit statt Signalmanipulation

Diphon-Synthese: Signalmanipulation

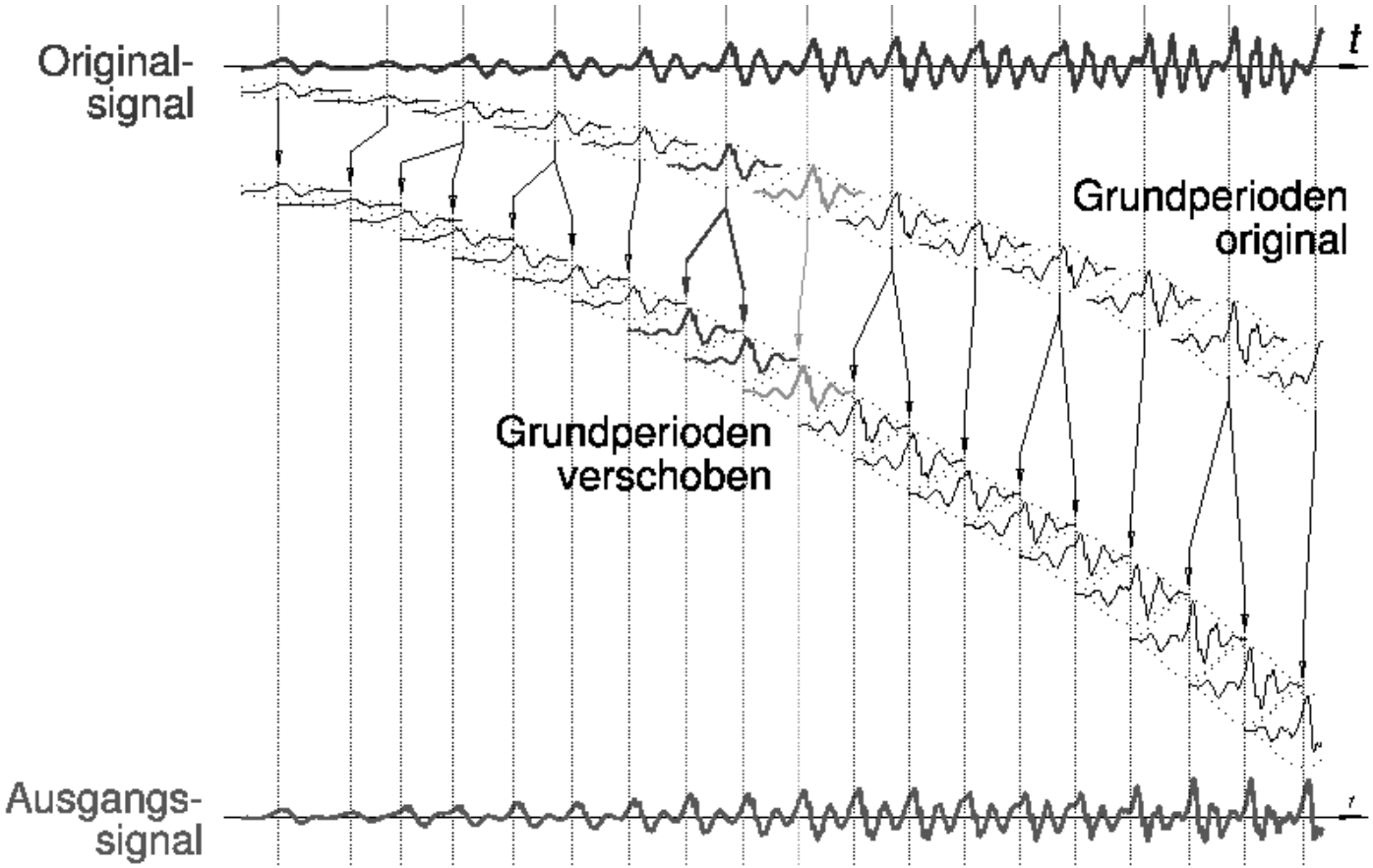
Klassischer Diphonsynthese: Signalmanipulation mit TD-PSOLA

- **TD:** *Time-Domain*, d.h. keine Überführung in Spektralbereich nötig
- **PS:** *pitch-synchron*, d.h. Verfahren operiert auf Einheiten der Größe einer glottalen Schwingungsperiode
- **OLA:** *overlap and add*, d.h. Einheiten werden überlagert und addiert

Diphon-Synthese: Signalmanipulation

- **Fensterung** der Einheiten: Multiplikation der Signalauschnitte mit einem Gewichtsfenster zur Abschwächung der Signalränder
- **Dauer-Manipulation:** Wiederholung von Kopien einer Periode
- **F0-Manipulation:** Verschiebung der Einheiten gegeneinander ($\longrightarrow$ Erhöhung) oder auseinander ($\longrightarrow$ Absenkung). Auffüllen mit/Löschen von Perioden zur Aufrechterhaltung der Dauer
- **Intensität:** Aufaddieren von Kopien einer Periode

Diphon-Synthese: Signalmanipulation



aus Hess (2004)

Unit-Selection

Unit-Selection: Kontextabhängige Auswahl der Units

- **Statt Signalmanipulation Suche nach der besten Sequenz $\hat{U}$ aus gespeicherten Unit-Varianten**
- basierend auf der Minimierung von **Target-** (T) und **Join-Kosten** (J)

$$\hat{U} = \arg\min_{U} \sum_{i} \left[J(u_{i-1}, u_i) + T(u_i, s_i) \right] \quad (1)$$

- s_i: durch die vorgeschalteten Text- und Prosodie-Module vorgegebenen Zielspezifikationen
- u_i: gespeicherte Unit

Unit-Selection

Target-Kosten $T(u_i, s_i)$

- Abstand des Exemplars u_i zu den Zielvorgaben s_i
- u_i, s_i als **Merkmalsvektoren** repräsentiert mit Angaben zu:
  - Identität der Unit
  - Unit-Kontext
  - prosodische Spezifikationen
  - F0-Kontur
  - Dauer
  - Intensität

Unit-Selection

- **Beispiel:**
  - s_i = [*/u:d/*, *+akz*, *–phrasenfinal*, *120-110-100*, *80*], d.h.
  - Ziel ist ein /u:d/-Diphon in akzentuierter und nicht-phrasenfinaler Position mit der F0-Kontur 120-110-100 Hz und der Dauer 80 ms

Unit-Selection

Join-Kosten $J(u_{i-1}, u_i)$

- Diskontinuitäten zwischen aufeinanderfolgenden Units u_{i-1} und u_i
- Features:
  - Mel-Cepstral-Distanz an der Konkatenationsstelle
  - absolute F0-Distanz
  - absolute Log-Energiedistanz