Unit-Selection-Synthese

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

9. Januar 2017

Inhalt

- Datengetriebenes Vorgehen vs. Signalmanipulation
- Klassische Diphon-Synthese
  - Korpuserstellung
  - Synthese
  - Erweiterungen
- Unit-Selection-Synthese
  - Datenbank
  - Synthese
- Evaluierung
  - Verständlichkeit
  - Natürlichkeit

Datengetriebenes Vorgehen vs. Signalmanipulation

Datengetrieben: Unit-Selection-Ansatz

- Speicherung großer Mengen an Sprachsignalen
- keine oder sehr wenig Manipulation der Signale bei ihrer Verknüpfung
- **Pro:** höhere Natürlichkeit der Synthese
- **Kontra:** großer Aufwand zur Gewinnung des akustischen Materials

Datengetriebenes Vorgehen vs. Signalmanipulation

Signalmanipulation: klassischer Diphon-Ansatz

- Speicherung einer geringen Menge an Sprachsignalen
- Manipulation der Sprachsignale bei ihrer Verknüpfung
- **Pro:** höhere Flexibilität, weniger Aufwand bei der Datengewinnung
- **Kontra:** geringere Natürlichkeit der Synthese

Klassische Diphonsynthese

Einheit: Diphon

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

Klassische Diphonsynthese

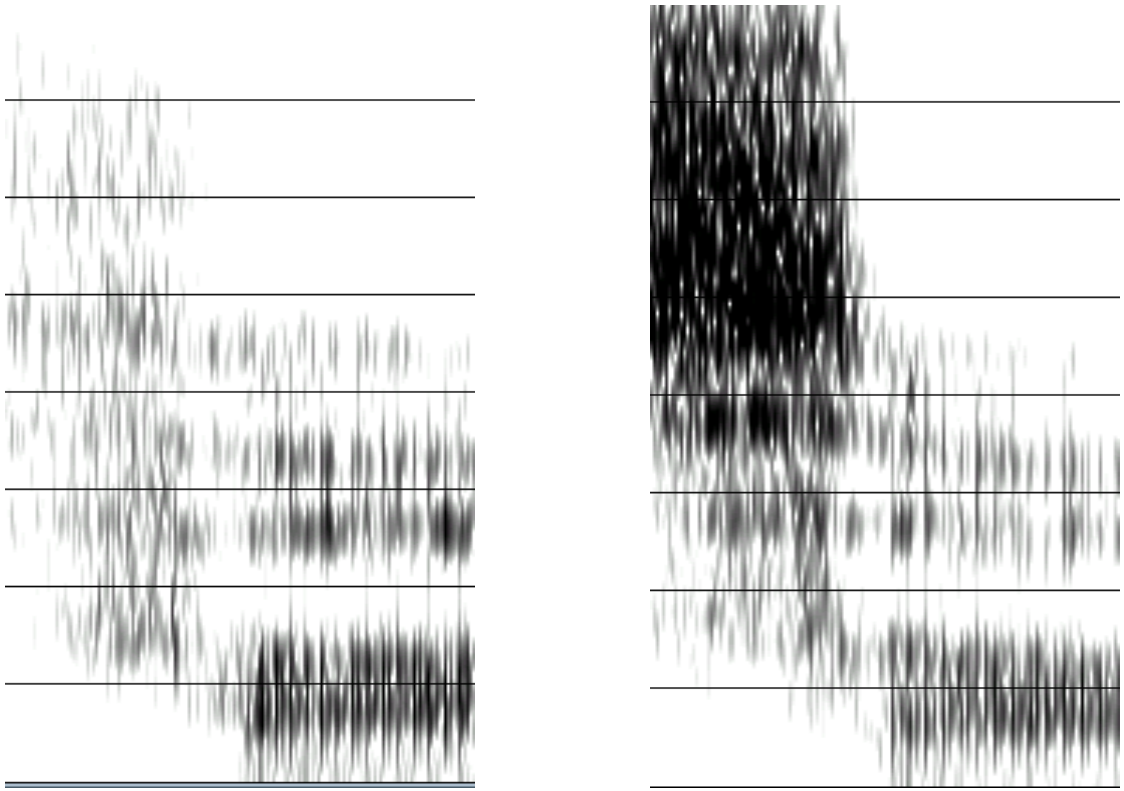


Abbildung : Diphone /fa/ und /sa/: unterschiedliche Formanttransitionen.

Klassische Diphonsynthese: Diphondatenbank

Erstellung der Diphondatenbank (voice)

- Ermittlung des nötigen Diphon-Inventars
- Einbettung der Diphone in einen **Trägersatz** $\longrightarrow$ **prosodisch homogene Realisierung der Diphone**
- Rekrutierung eines Sprechers (*voice talent*)
- Aufnahme des Sprechers beim Lesen der eingebetteten Diphone
- Evtl. mit elektroglottographischer Aufnahme
- Segmentierung, Pitch-Markierung

Klassische Diphonsynthese: Diphondatenbank

Pitch-Markierung (= Epoch Detection)

- **Epoche:** Ereignis im glottalen Schwingungszyklus
- z.B. Verschluss der Stimmlippen $\longrightarrow$ Führungsamplitude
- nötig für Signalmanipulation bei Konkatenation

Mittels Elektroglottographie EGG

- während der Aufnahme des Sprechers
- transglottaler Stromfluss
- Messung der glottalen Impedanz, die vom Abduktionsgrad der Glottis abhängt
- **Problem:** Messung der Stimmlippenschwingung bei nicht vollständiger Adduktion

Klassische Diphonsynthese: Diphondatenbank

Segmentierung

- Automatische Vorsegmentierung mittels **Forced Alignment** (z.B. durch MAUS; Schiel, 2004)
- **Forced Alignment:** Abbildung des Signals auf eine bereits bekannte Phonemfolge
- manuelle Nachsegmentierung

Klassische Diphonsynthese: Resynthese

Ablauf

- Auswahl der zur G2P-Vorgabe passenden Diphone
- **Resynthese:** Manipulation von Segmentdauer, Grundfrequenz, Intensität
- **iGgs zur Vollsynthese** (z.B. **Formantsynthese**) sind die zu manipulierenden Signale **bereits gegeben**

Manipulation von F0 und Dauer: TD-PSOLA

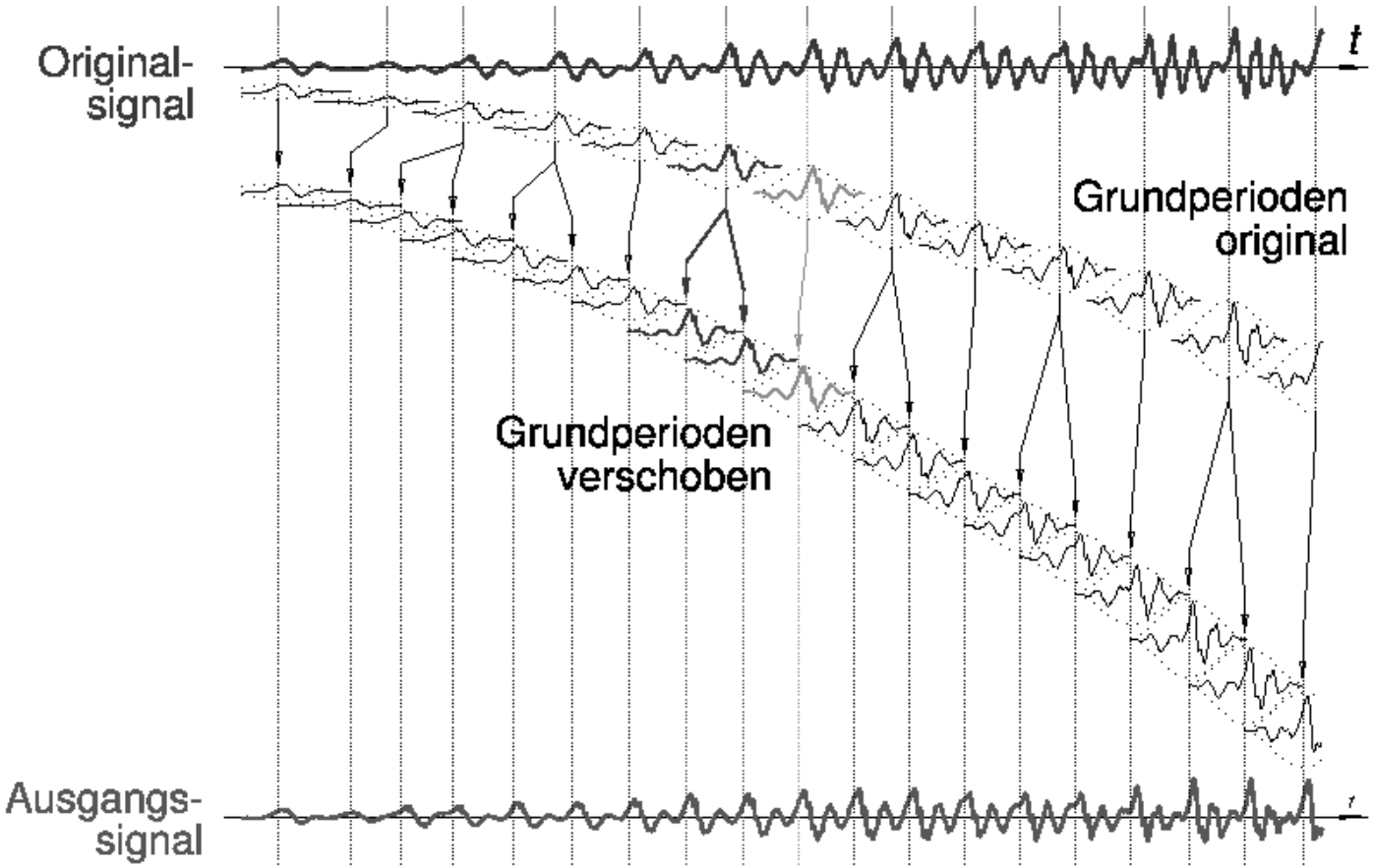
Signalmanipulation: TD-PSOLA

- **TD:** *Time-Domain*, d.h. keine Überführung in Spektralbereich
- **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

Manipulation von F0 und Dauer: TD-PSOLA

- **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

Manipulation von F0 und Dauer: TD-PSOLA



aus Hess (2004)

Unit-Selection-Synthese

Unterschiede zur klassischen Diphon-Synthese

- große Datenbank, keine oder geringe Signalmanipulation
- ermöglicht höhere Natürlichkeit der Synthese
- **Units:** variable Größe (z.B. Diphone); je größer die Einheiten, desto größer das benötigte Inventar

Unit-Selection: Datenbank

- Für jede Unit Aufnahme von **mehreren Exemplaren:**
  - +/–akzentuiert, +/– phrasenfinal, unterschiedliches Sprechtempo, unterschiedliche emotionale Markierung, ...
- Extrahierung der akustischen Charakteristika (s.u.)

Unit-Selection: Verkettung

- **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: Verkettung

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: Verkettung

- **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
- $T(u_i, s_i)$ als Kombination von **Teilkosten**
  - eine Teilkostenfunktion für jedes der betrachteten Merkmale j: $T_j(u_{ij}, s_{ij})$
  - gewichtete Summe voneinander unabhängiger Teilkosten:

$$T(u_i, s_i) = \sum_j w_j T_j(u_{ij}, s_{ij}) \quad (2)$$

Unit-Selection: Verkettung

- **Teilkosten T_j numerischer Features:**
  - Korrelation zwischen F0 der Unit u_i und der F0-Zielkontur in s_i
  - Mittlere absolute Distanz zwischen F0 in u_i und Zielkontur
  - Absolute Distanz zwischen u_i-Dauer und Zieldauer
- **binäre Teilkosten T_j kategorialer Features:**
  - gleicher Unit-Kontext von u_i und s_i: 0, sonst 1
  - Akzentangaben gleich: 0, sonst 1

Unit-Selection: Verkettung

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

- Diskontinuitäten zwischen aufeinanderfolgenden Units u_{i-1} und u_i
- ebenfalls als gewichtete Summe von unabhängigen Teilkosten modellierbar:

$$J(u_{i-1}, u_i) = \sum_j v_j J_j(u_{i-1j}, u_{ij}) \quad (3)$$

- Teilkosten J_j (nach Hunt&Black, 1996):
  - Cepstral-Distanz an der Konkatenationsstelle
  - absolute F0-Distanz
  - absolute Log-Energiedistanz

Unit-Selection: Verkettung

Ermittlung der besten Unit-Sequenz $\hat{U}$

- analog zu statistischen POS-Taggern, Alignment (vgl. POS-, G2P-Folien)
- **HMM-Modellierung**
  - Beobachtungen: Targets $\{s\}$
  - Zustände: gespeicherte Units $\{u\}$
  - Transitionswahrscheinlichkeiten $\rightarrow$ Join-Kosten
  - Emissionswahrscheinlichkeiten $\rightarrow$ Target-Kosten
- $\hat{U}$ mittels **Viterbi**: Finden des Pfades durch die Trellis, auf dem die minimalen Kosten anfallen

Evaluierung

Verständlichkeit

- Verwendung semantisch nicht vorhersagbarer Sätze (SUS)
- Erzeugung eines SUS: zufälliges Auffüllen eines syntaktisch wohlgeformten POS-Templates mit Wörtern der entsprechenden Wortarten
- **Reimtest**:
  - Diskriminierbarkeit von Konsonanten
  - Paare sich reimender Wörter, die sich jeweils in einem distinktiven phonologischen Merkmal unterscheiden
  - **ABX-Test**: Präsentation eines Worts aus einem Paar (in SUS) + Aufgabe, zu beurteilen, welches

Evaluierung

Natürlichkeit

- **Mean-Opinion-Score (MOS)**: Qualitätsurteile auf einer Skala von 1–5