Sprachsynthese – Überblick

Uwe Reichel (Änderungen von F. Schiel 2016)
Institut für Phonetik und Sprachverarbeitung
Ludwig-Maximilians-Universität München
reichelu—schiel@phonetik.uni-muenchen.de

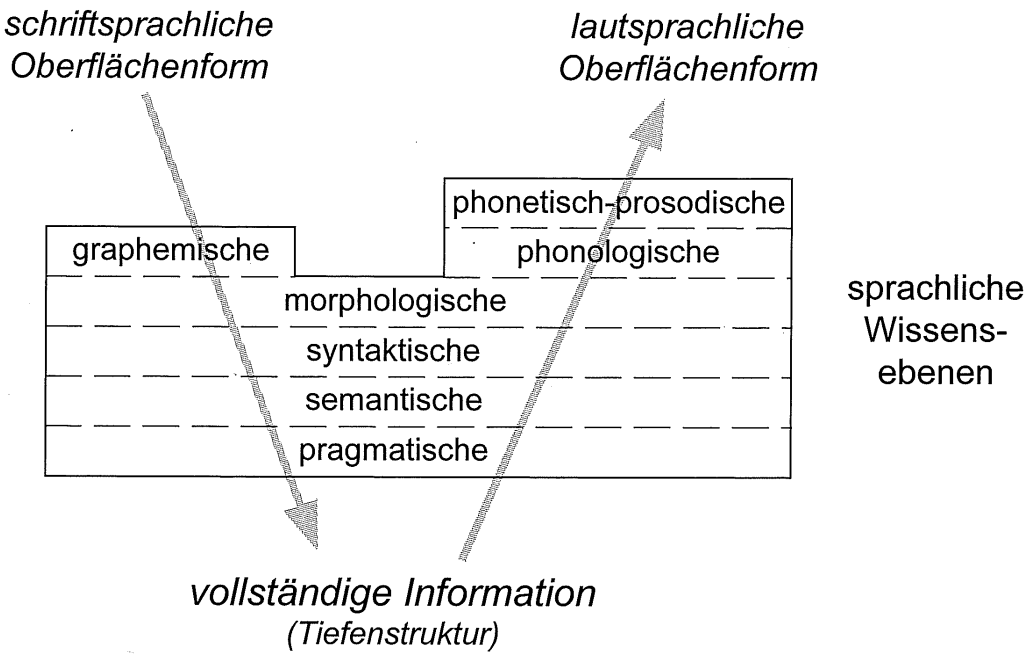
25. Oktober 2022

Inhalt

Was ist Sprachsynthese?

Was ist bedeutet Sprachsynthese und warum ist sie so schwierig?

Pfister 2008. S. 196: Oberfläche $\longrightarrow$ Oberfläche



Typologie von Synthesesystemen

Reichweite

- **Concept-to-Speech CTS:**
  Intention $\longrightarrow$ *Text* $\longrightarrow$ *akustisches Signal*
- **Text-to-Speech TTS**: *Text* $\longrightarrow$ *akustisches Signal*

Signalgenerierung

- **Konkatenative Synthese:**
  Verkettung gespeicherter phonetischer Segmente
- **Formantsynthese:** direkte akustische Generierung
- **Vocoder:** steuerbarer 'Voice-Encoder'
- **Artikulatorische Synthese:** Echtes akustisches Modell
- **Statistische Modelle: HMM-Synthese, Deep NN**

Architektur

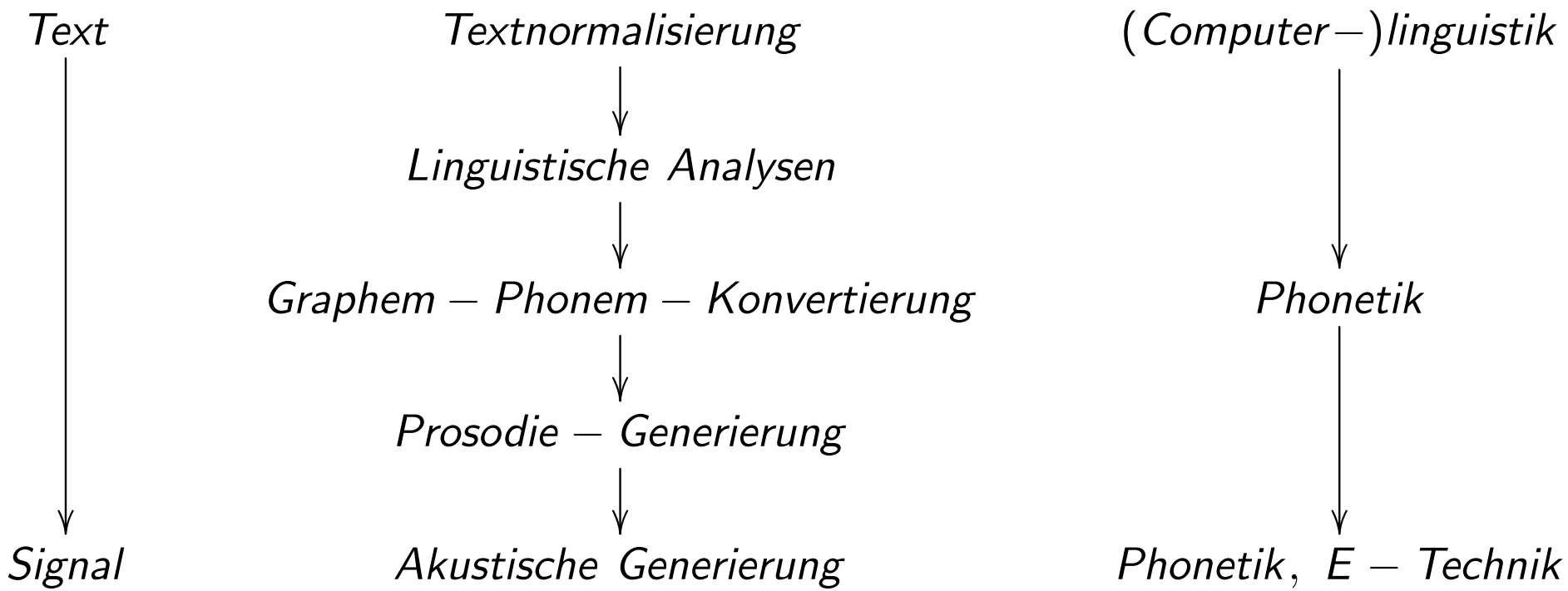


Abbildung: Architektur eines TTS-Systems.

Architektur: Textebene

Textnormalisierung

- **Satzsegmentierung** (Disambiguierung der Interpunktion)
- **Expansion von Non-Standard-Words** (Titel, Abkürzungen, Ziffernfolgen etc.)
- **Beispiel:**
  - *am 2.* $\longrightarrow$ *zweit***en**
  - *der 2.* $\longrightarrow$ *zweit***e**
  - *1980 Studenten* $\longrightarrow$ *eintausendneunhundertachtzig*
  - *im Jahr 1980* $\longrightarrow$ *neunzehnhundertachtzig*

Architektur: Textebene

Linguistische Analysen

- Zuweisung der **Wortarten** (*Part-of-Speech-Tagging*)
- ggf. **syntaktische** Analyse
- ggf. **semantische** Analyse, Fokuslokalisierung
- Grundlage für **prosodische Struktur:** Phrasierung, Akzente

Architektur: Textebene

Graphem-Phonem-Konvertierung

- Überführung der Buchstabenfolge in eine Transkription
- **Schwierigkeiten:** Eigennamen, fremdsprachliches Material
- erweiterbar durch Phonem-Phonem-Konvertierung
  (kanonische Aussprache $\longrightarrow$ Spontansprache)

Architektur: Textebene

Graphem-Phonem-Konvertierung

(Nicht ganz Ernst zu nehmendes) Beispiel Englisch:
The word *'ghoti'* could be pronounced like *'fish'*
(Charles Ollier, 1788–1859)
'gh' like /f/ in *'enough'*
'o' like /I/ in *'women'* (plural)
'ti' like /esch/ in *'information*

Architektur: Prosodie

Prosodiemodellierung

- Ermittlung der **prosodischen Struktur**: Phrasengrenzen, Akzente
- Modellierung von **Pausen, Grundfrequenzverlauf, Lautsegmentdauern, sowie Intensität**

Architektur: Signalebene

Konkatenative Synthese

- Verkettung von phonetischen Einheiten aus einer Datenbank
- **Diphonsynthese**: Einheiten sind Diphone
- **Unit selection**: variable Länge der Einheiten
- Wiedergabe der Einheit als Signal oder mit Vocoder

Direkte Generierung des akustischen Signals

- Quellsignal + Filterung
- Z.B. **Formantsynthese**: Kombination mehrerer Filter (Resonatoren)

Architektur: Signalebene

Artikulatorische Synthese

- Lautproduktionsmodell
- Abbildung der Produktion auf ein akustisches Signal
- (bis jetzt nur wissenschaftliche Systeme)

HMM-Synthese

- Verwendung statistischer *Hidden-Markov*-Modelle zur Steuerung eines Vocoders
- Suche nach dem im jeweiligen Kontext wahrscheinlichsten Wert eines akustischen Parameters

Anwendungen

- Ansagen (Bahnhöfe, U-Bahn, etc.)
- Dialogsysteme (Siri, Alexa, telephone banking, etc.)
- Unterstützung bei körperlichen Beeinträchtigungen
  - Sprechersatz bei Stummheit
  - Vorlesen von Webseiten etc. bei Sehbehinderten
- Geräte Output, Nav.system, Spielzeuge, Computerspiele

Wissenschaftliche Systeme

- VocalTractLab (P. Birkholz, TU Dresden)
  www.vocaltractlab.de
- FESTIVAL (IMS Stuttgart)
  festvox.org/festival
- CHATR (N. Campell, ATR Kyoto)
  winnie.kuis.kyoto-u.ac.jp/members/ian/chatr-doc/chatr_5.html

Synthesebeispiele

Umfassende Kollektion von Felix Burghardt:
http://ttssamples.syntheticspeech.de/