Einführung in die Spracherkennung & Sprachsynthese

Lucia D. Krisnawati

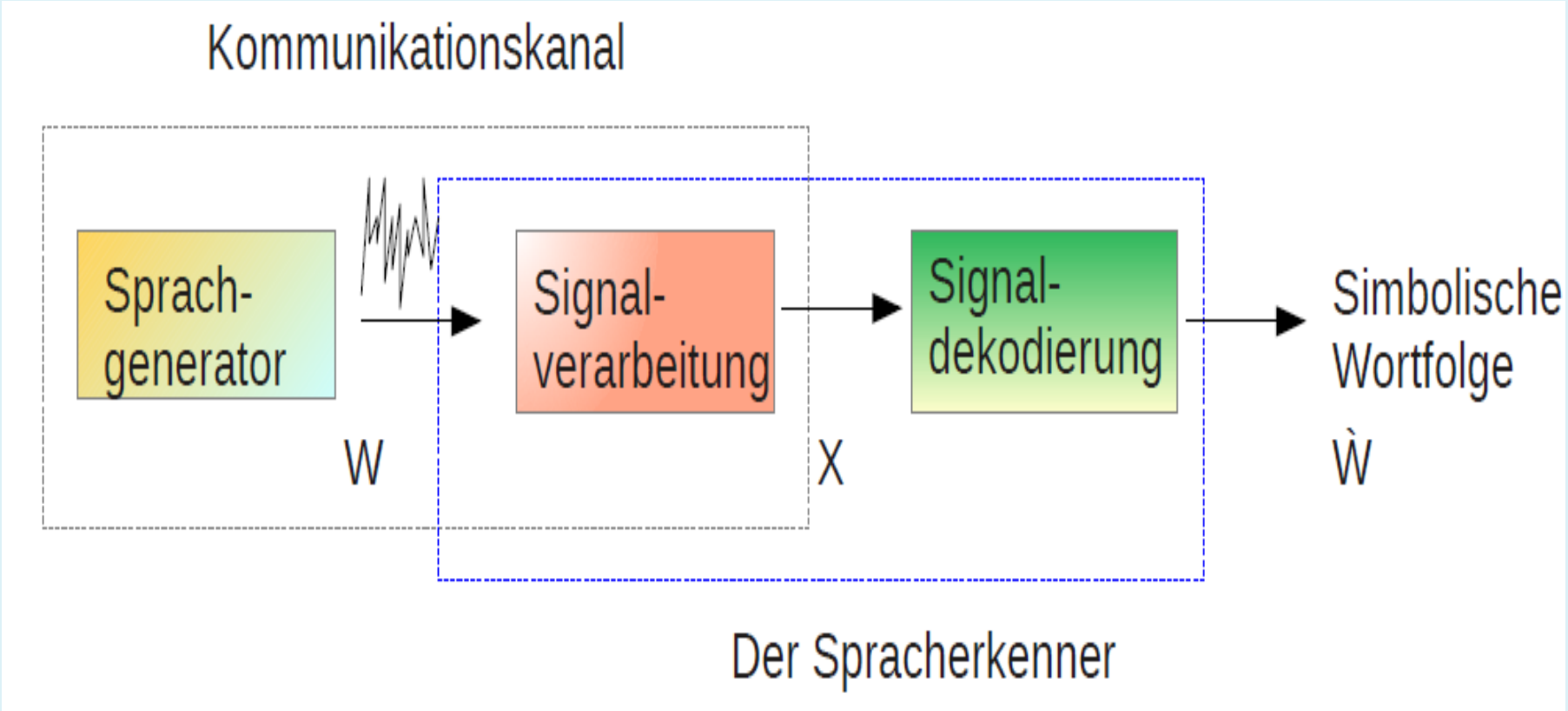
Vorlesungsinhalt

- Spracherkennung:
  - Einleitung
  - Die Architektur der Spracherkennung
  - Die Spracherkennungskomponente
  - Sie Anwendungen & Demo
- Sprachsynthese:
  - Die Architektur der Sprachsynthese
  - Die Komponente der Sprachsynthese
  - Die Anwendungen & Demo

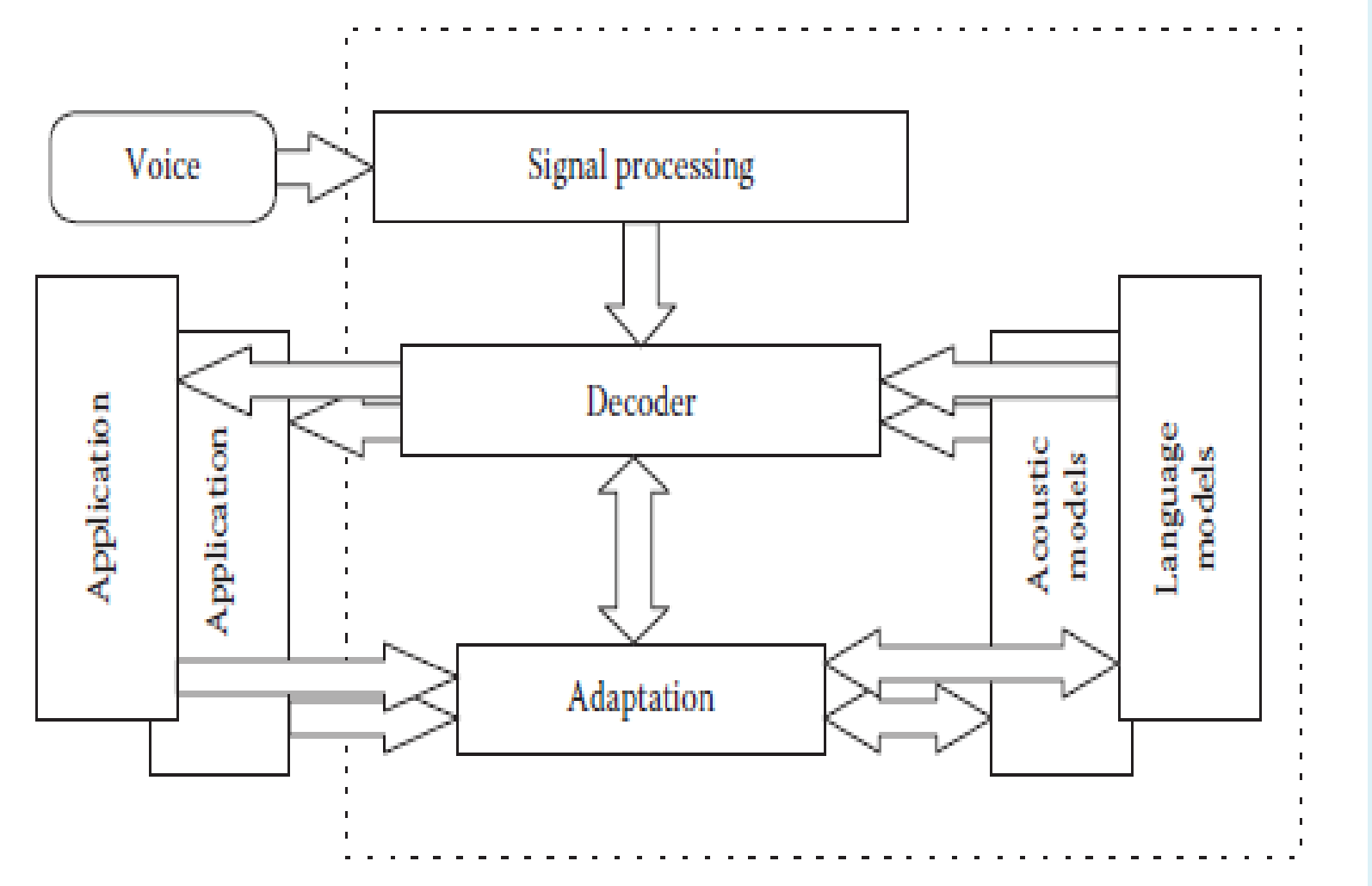
Was ist automatische Spracherkennung (ASR) ?

Ziel der ASR

- Ziel der ASR ist die symbolische Darstellung einer sprachlichen Äußerung, die als akustisches Signal vorliegt.



Die Architektur der ASR

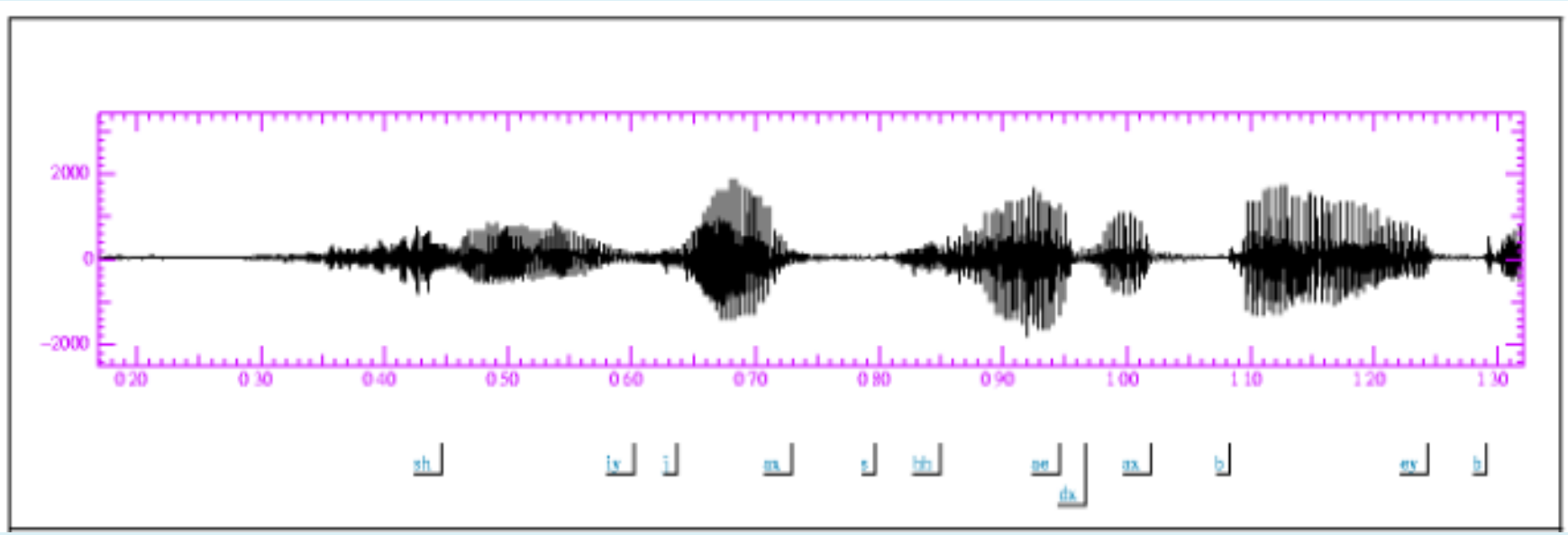


Die Architektur der ASR

- Die Komponenten der Spracherkennung:
  - Signal Verarbeitung
  - Dekodierung, die 2 Modelle ausnutzt:
    - Akustische Modellierung
    - Sprachmodellierung
  - Adaptation

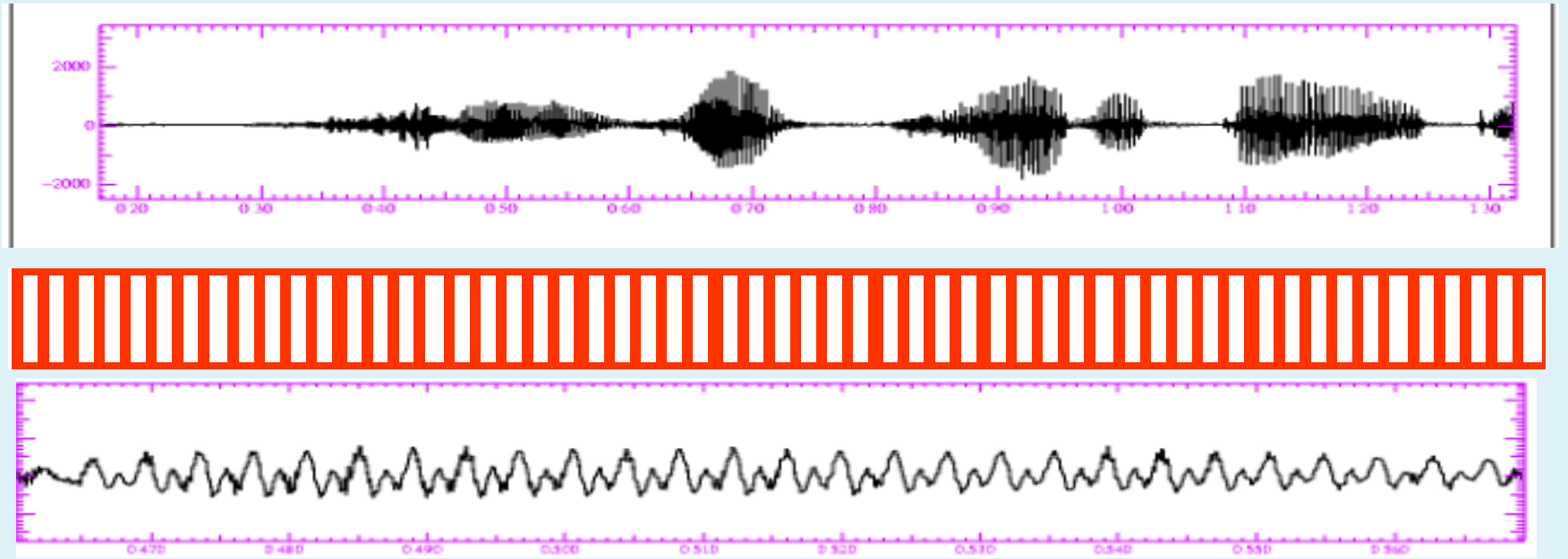
Signalverarbeitung

- 2 Hauptmerkmale eines akustischen Signals:
  - Die Frequenz : wie oft wiederholt sich eine Schallwelle in einem Durchlauf pro Sekunde
  - Die Amplitude: zeigt die Höhe des Luftdrucks
- Wie interpretieren wir eine Schallwelle?



[2]

Signalverarbeitung



- Die Schallwelle wird im Fenstersegment etwa 10,15,20 ms geschnitten.
- Das Fenstersegment wird in Spektrumsmerkmale transformiert.
- Spektrumsmerkmal ist eine Repräsentation der Schallwelle bezüglich der Frequenzverteilung.

Signalverarbeitung

- Die Phase der Merkmalsextraktion besteht aus:
  - Signalabtastung
  - Quantisierung
  - Phonerkennung
- Signalabtastung:
  - Das kontinuierliche Signal wird an Äquidistante Zeitpunkten abgetastet.
  - Nach dem Shanon'schen Abtasttheorem reichen 8kHz & 16kHz aus um die Hauptspektralanteile abtasten zu können.
  - Mindesten gibt es 2 Abtastungen in einem Ablauf, eine für positive Werte und die andere für negative Werte.
  - Die Speicherung erfolgt für Diktieranwendungen auf 16 bit & für Telefonanwendungen auf 8 bit.

Signalverarbeitung

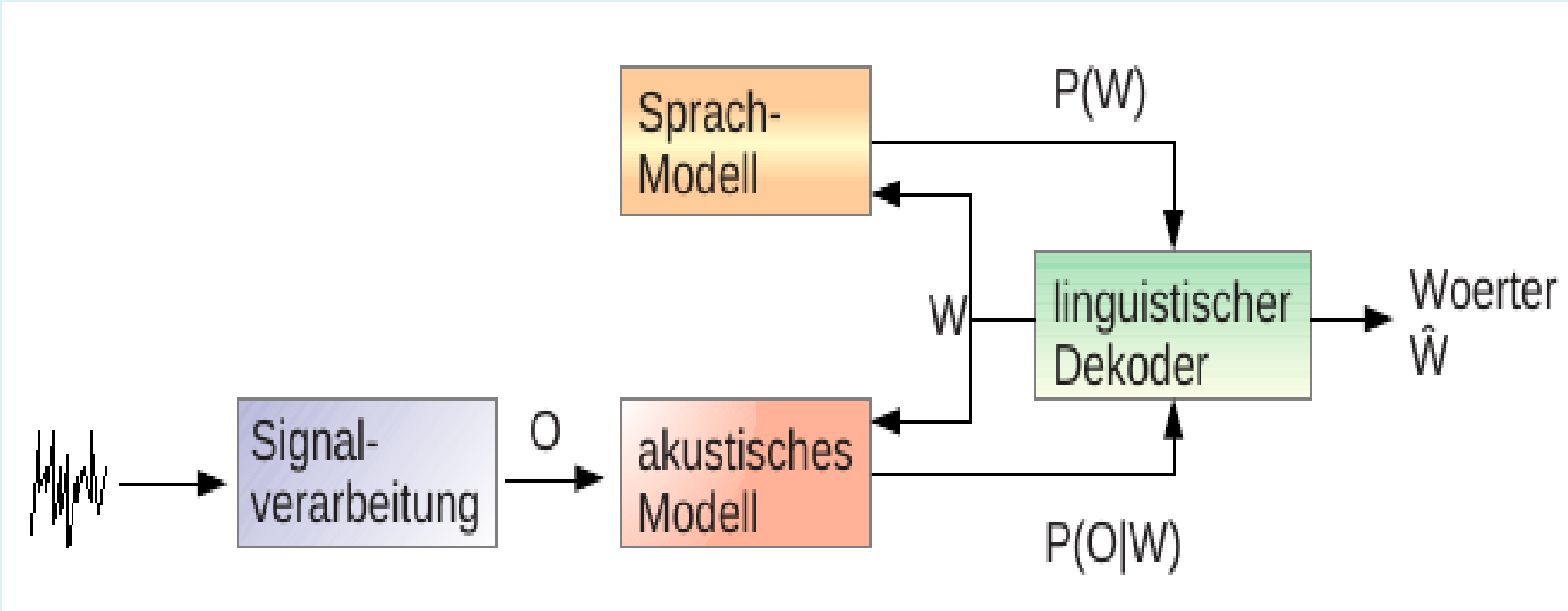
- Quantisierung:
  - Das abgetastete Signal wird quantisiert.
  - Quantisierung ist ein Umwandlungsprozess einer real-bewerteten Zahl zu einem Integer.
  - Resultat ist eine Diskrete Folge von Signalwerte

[4]

Signalverarbeitung

- Phonerkennung :
  - Benutzt statistische Ansätze wie Neuronale Netze oder Gaussianisches Modell, um das individuelle Phon zu erkennen, z.B. [p], [b].
  - Die Ausgabe ist ein wahrscheinlichster Merkmalsvektor eines Phons in jedem Fenstersegment.
  - Wie berechnet man eine Wahrscheinlichkeit der Merkmalsvektoren?
    - Erst wird die Signalwerte in einem berechenbar diskreten Symbol „geclustert“.
    - Die Wahrscheinlichkeit von gegebenen Cluster wird durch ihre Häufigkeiten im Trainingsdata berechnet.

Dekodierung



- Ziel der Dekodierung ist die Entdeckung einer Wortfolge, deren akustische & Sprachmodelle ganz ähnlich zu der Eingabe in der Form einer Merkmalvektorfolge sind.

Dekodierung

- Wahrscheinlichkeitsmodell
  - Umformulierung mit Bayes :

$$\hat{W} = \underset{W \in \mathcal{L}}{argmax} \frac{P(O|W)P(W)}{P(O)}$$

  - übrig bleibt:

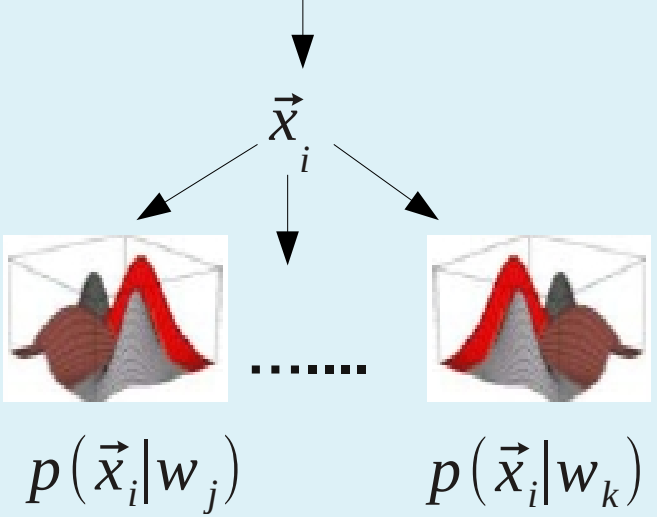
$$\hat{W} = \underset{W \in \mathcal{L}}{argmax}\, P(O|W)P(W)$$

  - P(O|W): welche akustische Beobachtungen sind bei Äußerung einer Wortfolge W zu erwarten → **akustisches Modell.**
  - P(W): Welche Wortfolgen sind im Anwendungskontext zu erwarten → **Sprachmodell**

Akustisches Modell

- Berechnung der Merkmalsvektoren

$$P(O|W)=\prod_{i=0}^{N} P(\vec{x}_i|w_i)$$



- Jede phonetische Einheit wird z.B. von Gaussdictung modelliert:

$$P(\vec{x}|w)=\sum_{j=0}^{M} g_j N(\vec{x}|\mu_j,\Sigma_j)$$

- Wobei g_j, μ_j, & Σ_j für die Gewichtung, Mittelwert & Kovarianz stehen

Sprachmodell

- ASR-System schränkt viele möglichen Wortkombinationen beim Sprachmodell ein:
  - Finite-state network
  - Deterministisch, sequenzielle Hemmung (Wortpaar).
  - Probabilistisch, sequenzielle Hemmung (Bigramm, Trigramm)
  - Bi-/Trigramme sind das dominierende Sprachmodell für ASR:

$$P(W_i|w_{i-2},w_{i-1})=\frac{C(w_{i-2},w_{i-1},w_i)}{C(w_{i-2},w_{i-1})}$$

    wobei C(x) steht für die Anzahl des Vorkommens von x in Trainingsdata.

Sprachmodell

- zB. in Training data: „John read her book. I read a different book. John read a book“
- Die Benutzung der Bigramme, um 3. Satz zu berechnen:

$$P(John|\langle s\rangle)=\frac{C(\langle s\rangle,John)}{C(\langle s\rangle)}=\frac{2}{3} \qquad P(read|John)=\frac{C(John,read)}{C(John)}=\frac{2}{2}$$

$$P(a|read)=\frac{C(read,a)}{C(read)}=\frac{2}{3} \qquad P(book|a)=\frac{C(a,book)}{C(a)}=\frac{1}{2}$$

$$P(\langle s\rangle|book)=\frac{C(book,\langle s\rangle)}{C(book)}=\frac{2}{3}$$

- P(john,read,a,book) = P(John|<s>)P(read| John)P(a|read)P(book|a)P(</s>|book) ≈ 0.148

Adaptation

- Adaptation bezieht sich auf die Fähigkeit, um sich lernen zu können.
- Hintergrund: Mensch-geliefertes Wissen ist sehr schnell veraltet, z.B. Aussprachewörterbuch
- Selbstlernende Fähigkeit wird durch Generalisierung realisiert:
  - kleine Testdaten, die ganz ähnlich zu Trainingsdaten
  - Trainingsdatenpartition
- Die Adaptation ist noch eine Herausforderung in heutzutage ASR

ASR Anwendungen & Demo

- Samsung's Smartphone :
  - Voice Talk, Voice command, voice search

/mounts/Users/student/lucia/Documents/lucias-file/CL1/lecture/Samsung Galaxy S II Voice Talk App Demo using Justin Bieber - Glenn Ong of Glich's Life.mp4

- Android vs Iphone
  - Google Now vs Siri

/mounts/Users/student/lucia/Documents/lucias-file/CL1/lecture/Google Now vs. Siri: The results speak for themselves.mp4

ASR Anwendungen

- Sprach-gesteuerter Software im Auto, z.B. BMW

- Front-end Technology in Dialogsysteme:
  - Deutsche Bahn, (0800-1-507090 oder 0241-604020)
  - Sparda-Bank

Probleme der ASR

- Phonologische Variationen:
  - Lokale und globale Kontexte
- Individuelle Variationen:
  - Anatomie & sozial-linguistischer Faktor
- Umgebungsfaktor:
  - Störungen, Geräusch
- Variationen der Sprachbenutzung
  - Syntax, Semantik, Diskurs
- Real-Welt Probleme
  - Störung des Redeflusses, neue Wortschätze
- ……

Sprachsynthese

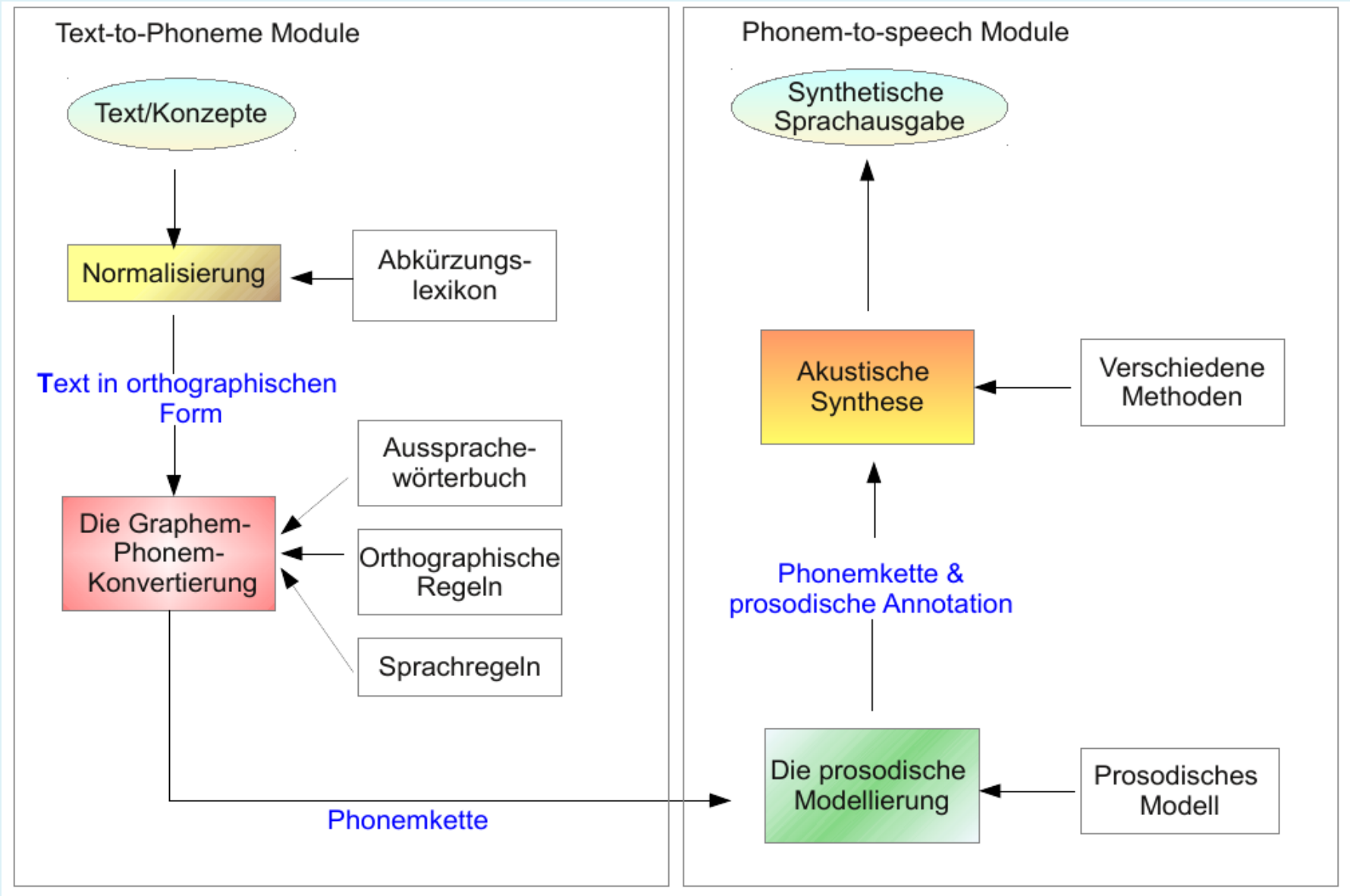
Was ist eine Sprachsynthese?

Voice
as a Service

Sprachsynthese

- Sprachsynthese wird auch als 'Text-to-Speech' (TTS) genannt.
- TTS ist ein System, das das Text als Eingabe in einer synthetischen Sprache konvertiert
- Eingabetyp:
  - Text-zu-Sprache vs. Konzept-zu-Sprache (CTS)
  - In CTS wird die Inhalt der Gespräch/Nachricht durch internen Repräsentationen ermittelt.
    - z.B. Datenbank aus Anfragesystem

Die Architektur der TTS



Komponente der TTS

- Komponente einer TTS:
  - Textnormalisierung
  - Graphem-Phonem-Konvertierung
  - Prosodische Modellierung
  - Akustische Synthese

Textnormalisierung

- Jedes Wort & Text, das besondere Aussprache hat, soll ins Lexikon gespeichert werden:
  - Abkürzungen (Dr., Str. usw)
  - Akronyme (UNESCO, ADAC, BaFög)
  - Spezielle Symbole (&, %)
  - Schreibkonventionen ($5 Million, 12°C)
  - Zahlen (Telefon, Datum)
    - 1995 1,995
    - 199 5236 

Graphem-Phonem-Konvertierung

- Benutzung einer Aussprachewörterbuch
  - Enthält Lemma und ihre Aussprache
  - Viele Ausnahme sollen auch im Lexikon/Wörterbuch vorhanden sein
- Lexikon oder Regeln?
  - Es gibt kein Problem bei der Wortsuche und Speicherung → Sehr schnell
  - Aber benötigt Regeln für die unbekannten Wörter
- Manche Wörter haben mehrfachen Aussprachen
  - Freie Variation (either, economics)
  - Konditionierte Variation (schwache Forme, the, adult)
  - Homograph (does, content)

Graphem-Phonem-Konvertierung

- Die Homograph-Disambiguierung benötigt eine syntaktische Analyse
  - He makes a record of everything they record.
  - She is getting number as she stays in room number 124.
- Syntaktische Analyse ist auch benötigt, um die prosodische Merkmale zu bestimmen.

Prosodische Modellierung

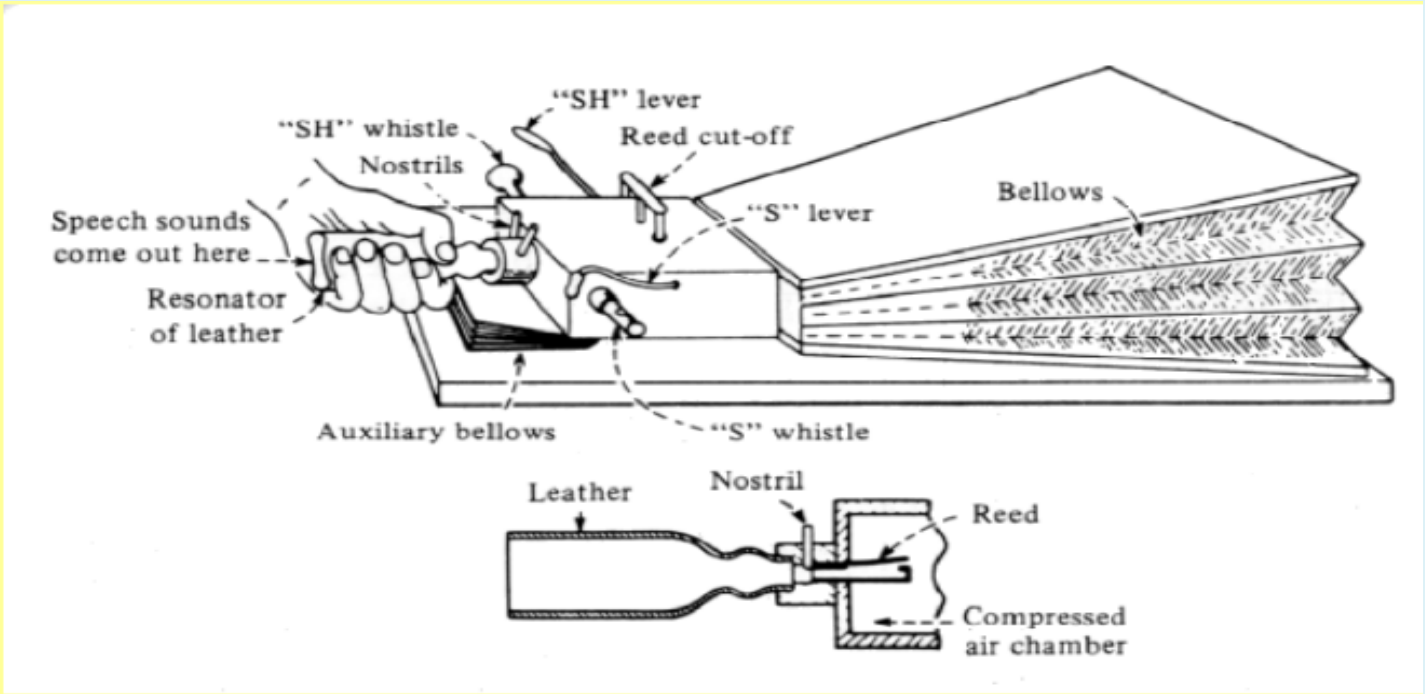
- Die Tonhöhe (pitch), Lange, Lautheit
- Die Satzmelodie (Tonhöhe):
  - Wird benötigt, um den monotonen Klang zu vermeiden
  - Sehr eng verbunden mit der Syntax (Frage, Aussage)
  - Wird benötigt, um den Satz zu thematisieren

Akustische Synthese

- Die alternativen Methoden:
  - Artikulatorische Synthese
  - Formantsynthese
  - Konkatenative Synthese
    - Diphon Synthese
    - Unit Selection Synthese

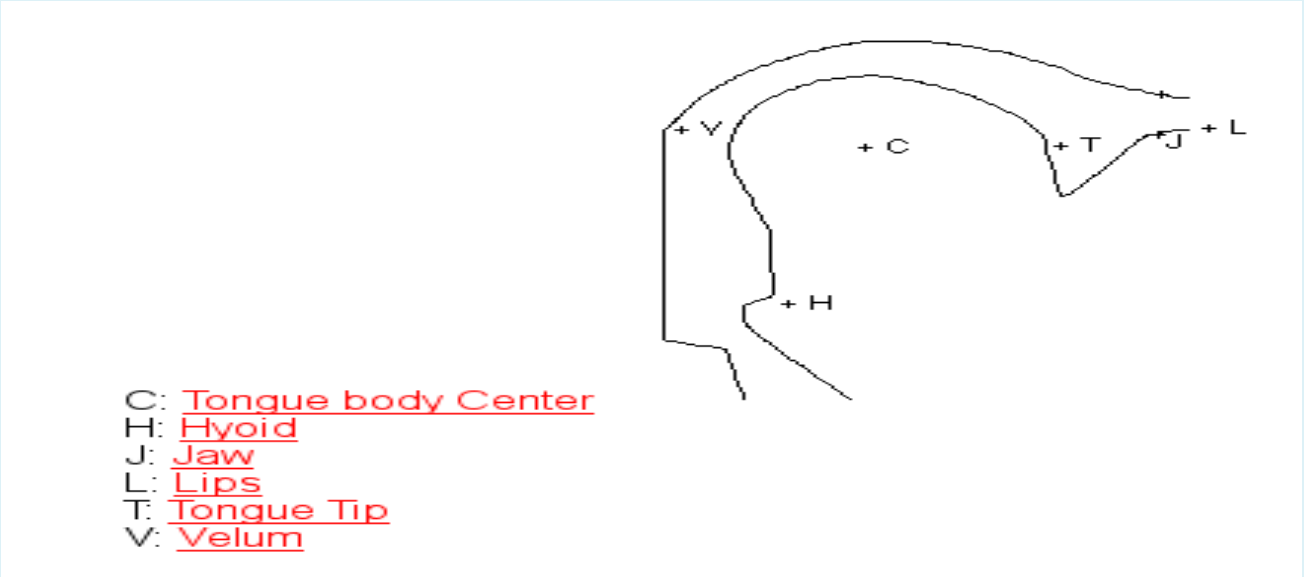
Die artikulatorische Synthese

- Physikalisches Modell für Simulation der Artikulation von Sprachlauten.
  - Man versucht die menschliche Sprachproduktion nachzubilden.
- Wolfgang v. Kempelen (1734-1804) benutzte die Bälge, Röhre, & Blätter.



Die artikulatorische Synthese

- Die moderne Version simuliert die Effekte der artikulatorischen Positionen, z.B. Model des Schwingenden Lippen, Model des Vokaltrakts



- Zu viel Arbeit ohne gutes Ergebnis
- Beispiel: /mounts/Users/student/lucia/Documents/lucias-file/CL1/lecture/koeln_artikulator_demo.mp3

Die Formantsynthese

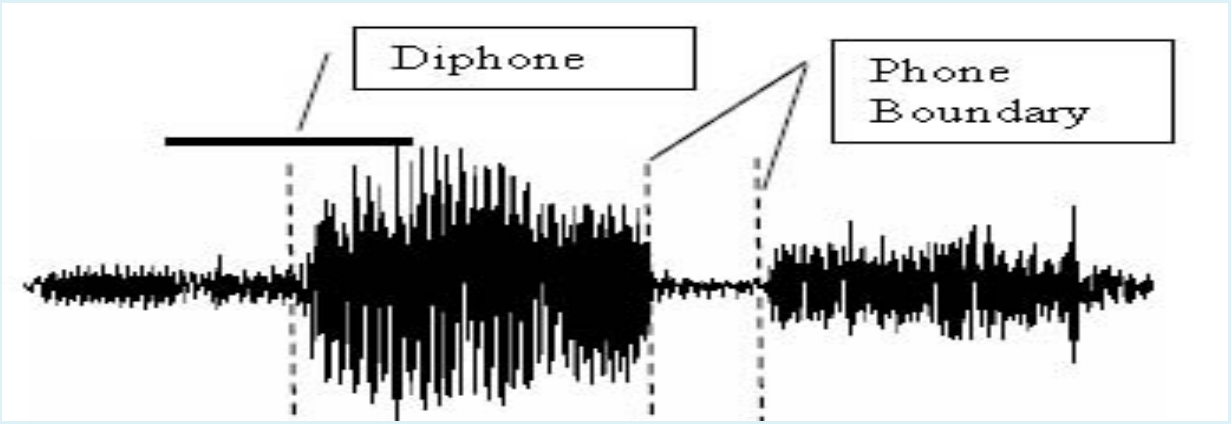
- Formant: die Konzentration akustischer Energie in einem bestimmten Frequenzbereich.
- Formantsynthese versucht die relevanten Merkmale von der akustischen Signalen nachzubilden:
  - Amplitude & Formantfrekuenz
  - Die Resonanz & Geräusch, z.B für Nasale, Laterale, Frikative, usw.
- Die werte der akustischen Parameter werden von den Regeln der Aussprachebezeichnung abgeleitet.
- Das Ergebnis ist verständlich aber klangt unnatürlich.
- Beispiel: Eloquence
  - /mounts/Users/student/lucia/Documents/lucias-file/CL1/lecture/eloqdemo.mp3

Konkatenative Synthese

- Basis ist eine Sprachdatenbank
    - Sprache wird aufgenommen, segmentiert und annotiert.
- Die Verkettung der Sprachsignale erfolgt durch Auswahl der am besten zueinander passenden Teilsignale.
- Das Segment könnte in Form von:
    - einem Phonem, Wort, eine Phrase, oder
    - Kombination von Phonem, Wort & Phrase
    - Oder etwas anders, das intelligenter ist, z.B Diphon, Unit.

Diphon Synthese

- Das Sprachsignal wird durch Verkettung von Diphonen erzeugt, Die Prosodie-Anpassung durch Signalmanipulation.
- Diphon:
  - Nachbar-Laut Kombinationen oder
  - Das Fragment von des Sprachsignals, das quer durch Phonemgrenze schneidet.
- Sehr wichtig für natürlicher Klang



Konkatenative Synthese

- Die Eingabe sind phonetische Repräsentationen & prosodische Merkmale.
- Die Lange, Tonhöhe, & Lautheit des Diphonsegments können digital manipuliert.
- Beispiele:
  - Konkatenative Synthese: First Byte

    /mounts/Users/student/lucia/Documents/lucias-file/CL1/lecture/firstbyte_demo.mp3
  - Diphon Synthese: Babeltech's Babil /mounts/Users/student/lucia/Documents/lucias-file/CL1/lecture/diphon.mp3

Unit Selection Synthese

- Aus einer sehr großen Datenbasis werden die am besten passenden Sprachteile (units) miteinander verkettet.
  - Unit → Phonen, Diphone, Halbsilben, Morpheme
- Die unit (Segmente) werden mit einem Verzeichnis von einer Reihe akustischer & phonetischer Eigenschaften (Grundfrequenzverlauf, Dauer, Nachbarn) gespeichert.
- Eine Reihe von möglichst großen Segmenten werden durch Suchalgorithmen, z.B. gewichtete Entscheidungsbäume, für die Synthese bestimmt.

Unit Selection Synthese

- Die sehr bekannte Variation ist non-uniform unit selection Synthese.
- In Non-uniform Unit Selection Synthese erfolgt die Verkettung der Sprachteile durch die unterschiedliche Länge
- Beispiel: cerevoice

  /mounts/Users/student/lucia/Documents/lucias-file/CL1/lecture/aristech_s1-Alex.mp3

Anwendungen der Sprachsynthese

- Ansagen in:
  - Telefon- & Navigationssystemen
  - Öffentlicher Verkehr (Ubahn, Bus)
- Dialogsysteme (incl, IVR)
- Voice-reading App:
  - Emails Aloud, translation reader (Google)
  - Voice Aloud reader (Android):
    - Webseiten, Nachricten, emails, SMS, PDF files
- App 4 Diffable

Spracherkennung & sprachsynthese Demo

Literatur

- [1] Huang, X., & Deng, L., (2010). An Overview of Modern Speech Recognition. In Indurkya, & Damerau (Eds.), *Handbook of Natural Language Processing. Chapman & Hall/CRC. Kap 15.*
- [2] Jurafsky, D., & Martin, H. (2000). *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, & Speech Recognition*. New jersay: Prentice Hall. Kap 4 & 7.
- [3] Thakur, B.M., Chettri, B, & Shah, K.B. (2012). Current Trends, Frameworks & Techniques Used in Speech Synthesis – A Survey. In. *International Journal of Soft Computing and Engineering (*IJSCE), Vol 2(2), pp. 442-446.
- [4] Willet, D. (2000). *Beitraege zur statistischen Modellierung und effiziente Dekodierung in der automatischen Spracherkennung*. Dissertation.
- http://ttssamples.syntheticspeech.de/deutsch/