Erkennung und Erzeugung gesprochener Sprache

8.11.2005

Sebastian Pado

Einführung

- Viele computerlinguistische Anwendungen gehen von **textbasierter** Ein/Ausgabe aus
  - Wieso?

- Nicht überall möglich oder sinnvoll
  - Beispiele?

- Übersetzungen
  - Gesprochen » geschrieben: **Spracherkennung**
  - Geschrieben » gesprochen: **Sprachsynthese**

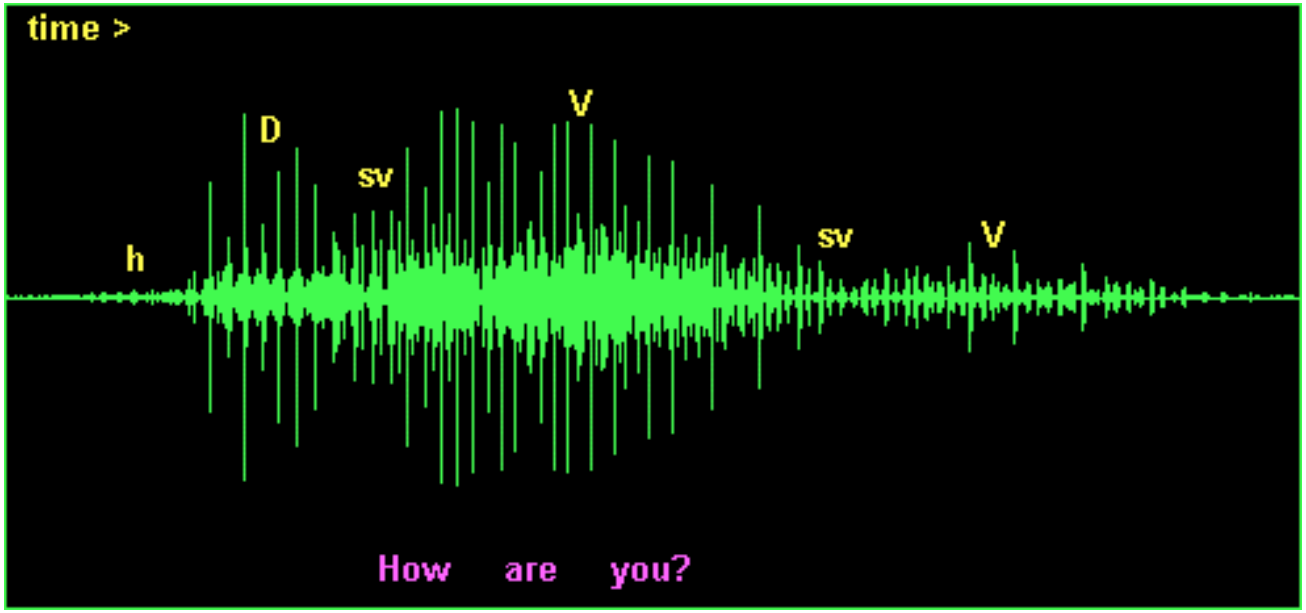
Spracherkennung und -synthese: Wozu?

- Sprachsynthese
  - Vorlesesysteme
  - Durchsagen (zB im Bahnhof)
- Spracherkennung
  - Diktiersysteme (zB Röntgendiagnosen)
  - Computersteuerung
- Kombinationen: Dialogsysteme
  - Auskunftssysteme (zB Fahrplanauskunft)
  - Telefonbedienung im Auto
  - Liftbedienung

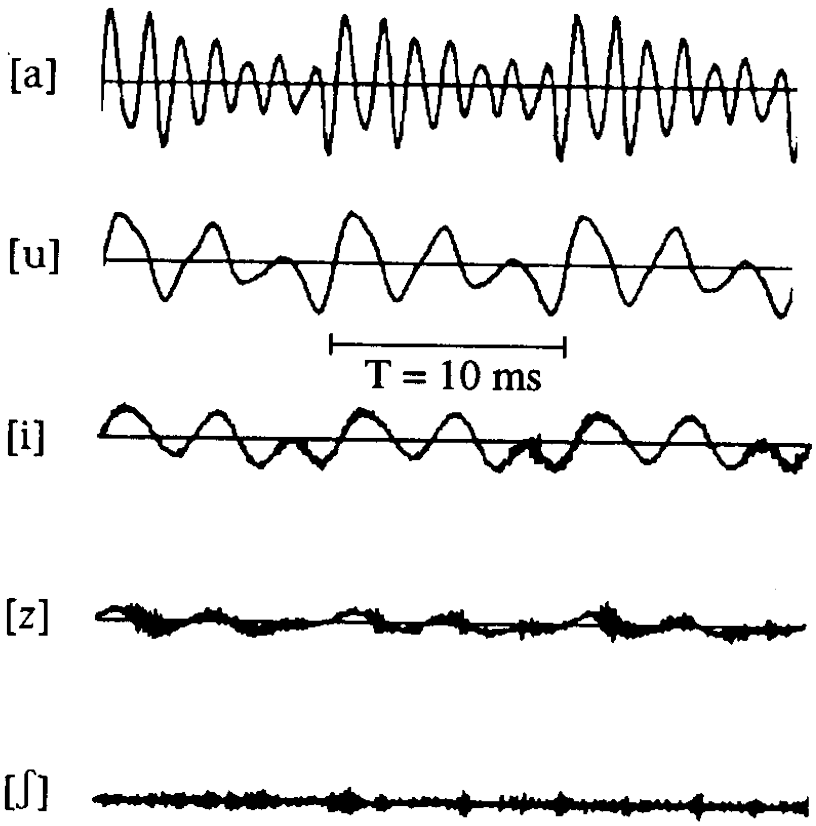
Teil 1: Spracherkennung

Spracherkennung

- Grundaufgabe
  - Gesucht: Wortkette (Äußerung)
  - Gegeben: Kontinuierliches Schallsignal (Mikrophon)
    - **Oszillogramm**: Zeit-Energie-Diagramm

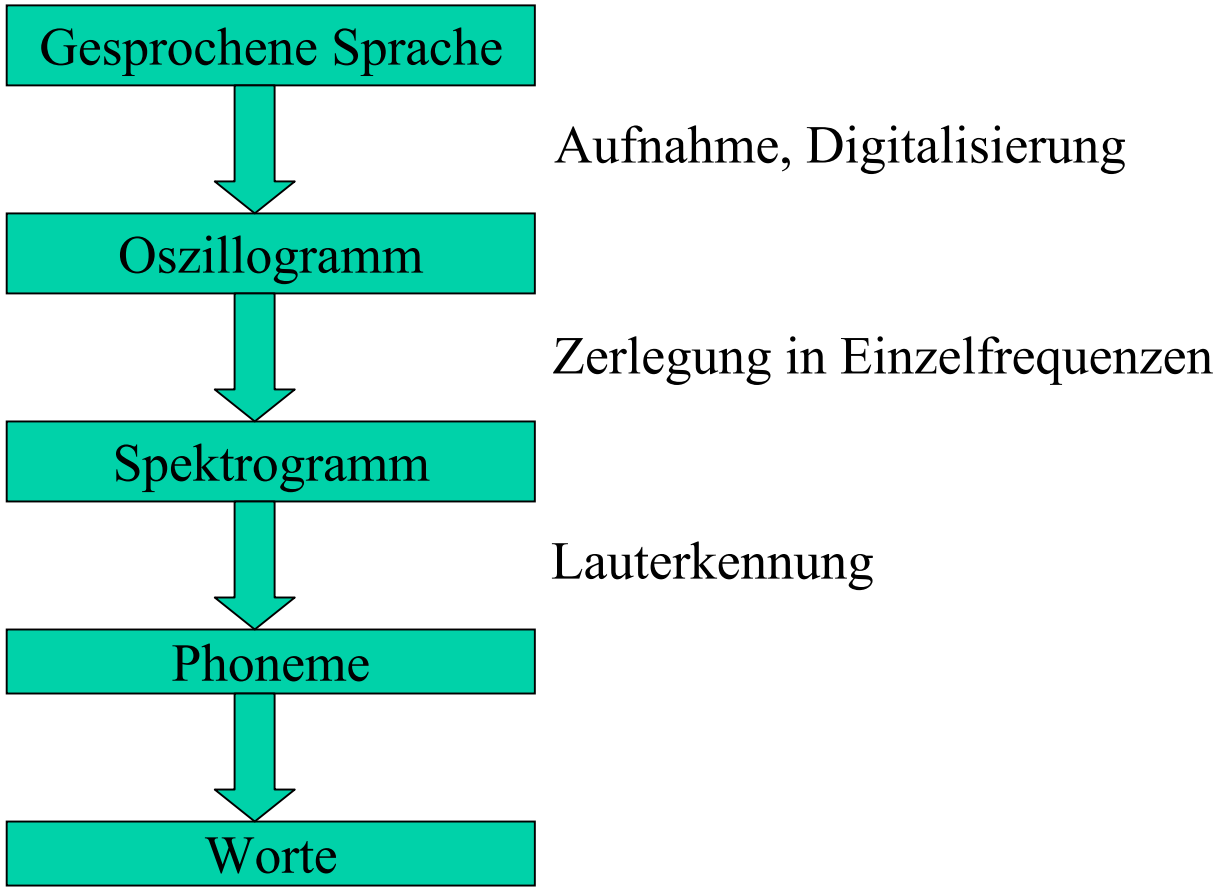


Einzelne Laute als Oszillogramme

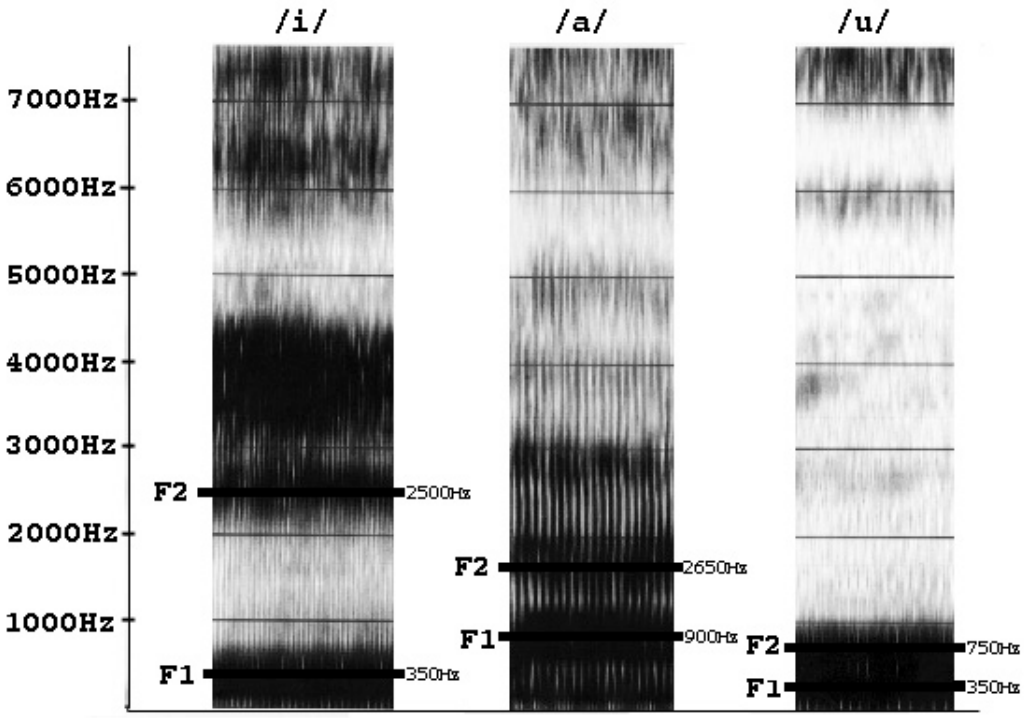


- Laute charakterisiert durch Kombination von Schwingungen verschiedener Frequenzen
- Im Oszillogramm **schwer erkennbar** (Überlagerung)
- Daher: Geschicktere Repräsentation durch Komponentenanalyse (Fourier-Transformation)
- Ergebnis: Zeit-Frequenz-Diagramm (**Spektrogramm**)

Spracherkennung: (Vereinfachtes) Schema

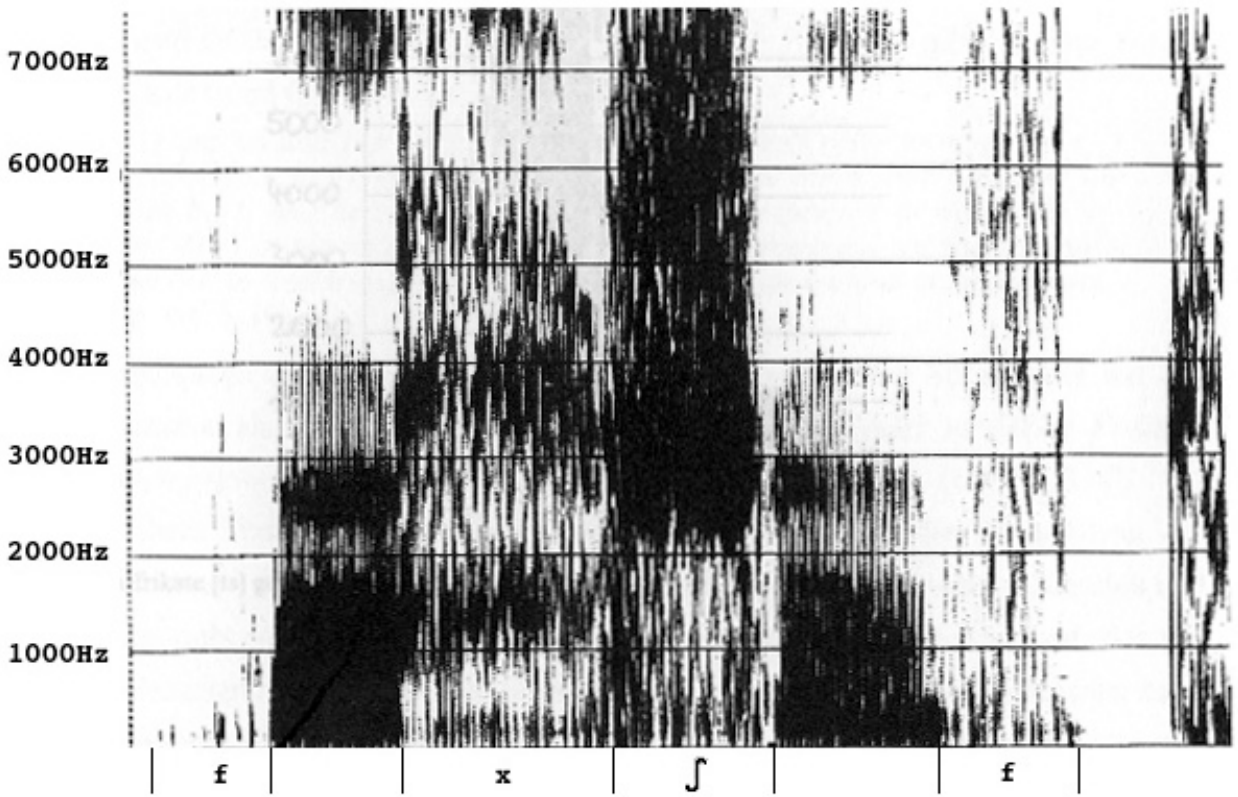


Spektrogramm für die Vokale i,a,u



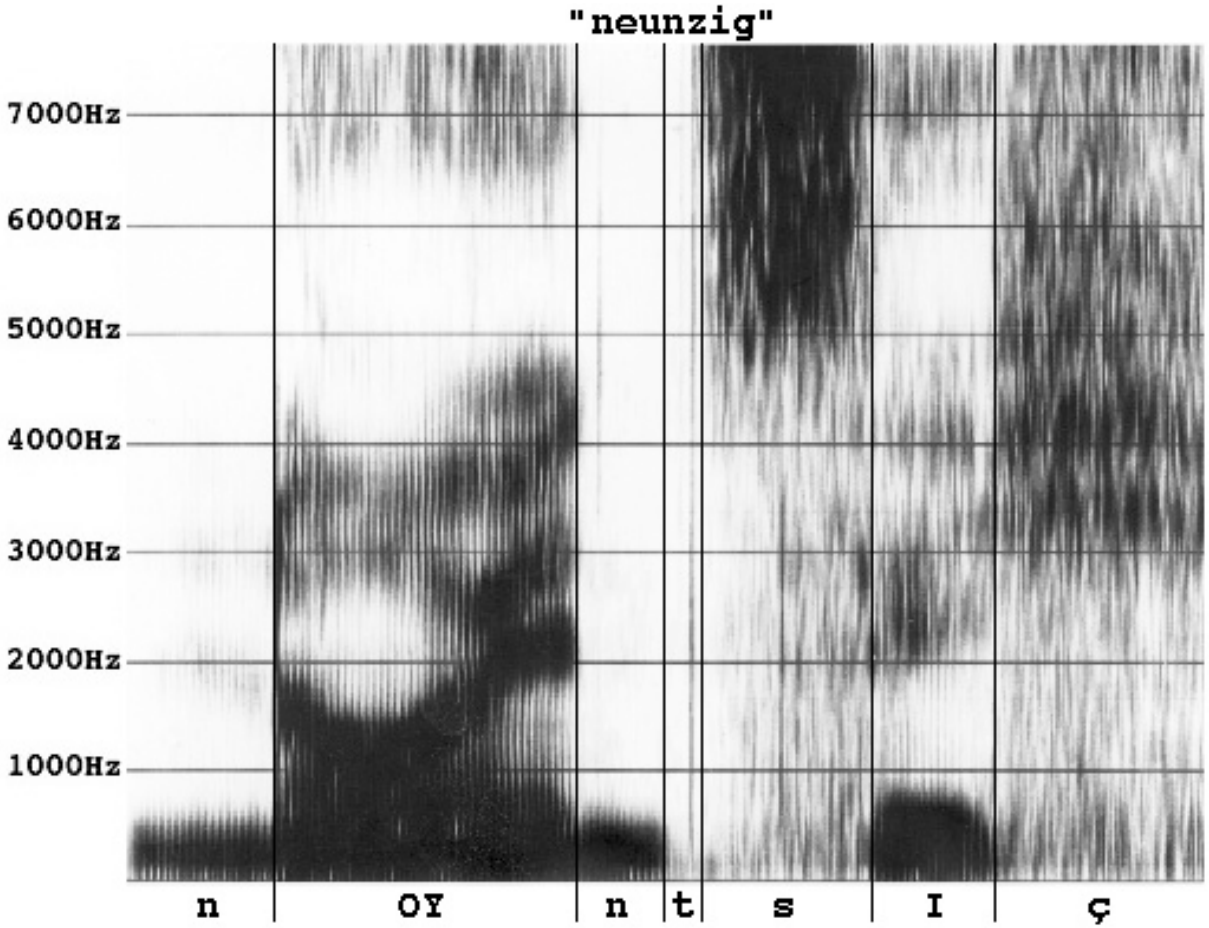
- Färbung entspricht Schallenergie in bestimmtem Frequenzbereich
- Die Formanten (Obertöne) F1 und F2 charaktierisieren Vokale
- Der Verlauf des Basisformanten F0 (hier nicht sichtbar) gibt die Intonation (Sprachmelodie) der Äußerung wieder.

Spektrogramm für einige Konsonanten

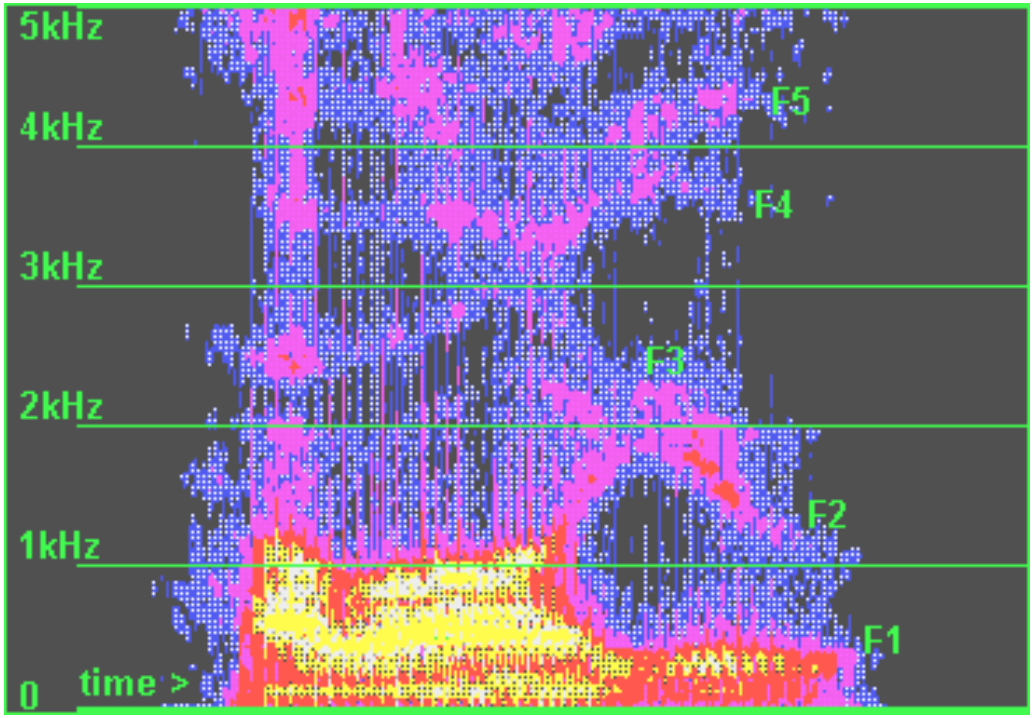


Frikative: f und ch-Laut („ach“-Laut); Sibillant: „sch“-Laut

Spektrogramm für ein deutsches Wort



Ein buntes Spektrogramm



... für den englischen Satz „How are you?“

Naives Modell zur Lauterkennung

- Schritt 1: Identifikation einzelner Spektrogramm-Schnipsel = Laute (Segmentierung)
  - Finde “Übergänge” in Spektrogramm
- Schritt 2: Vergleiche Spektrogramm-Schnipsel mit Datenbank “idealer” Laute (Identifikation)
  - Identifiziere passende Phoneme
- Schritt 3: Setze orthographische Realisierungen der Phoneme hintereinander
  - Ergibt die entsprechenden Wörter

Funktioniert leider nicht!

Problem: Kontinuität des Signals

Gesprochene Sprache läßt sich schwer unterteilen

- Die Laute eines Wortes lassen sich schwer abgrenzen
  - Wo hört Laut 1 auf, wo fängt Laut 2 an?
  - Schlimmer noch ist Koartikulation: Laute beeinflussen sich gegenseitig.
    - In Lautfolgen wie [am], [um], [an] kann man nicht den Vokal vom Nasal trennen: Vokal hat Nasal-Qualität und umgekehrt.
    - /k/ wird verschieden realisiert in Koffer, Kind, Kabel
- Wörter sind nur in der Orthografie sauber getrennt.
  - In der gesprochenen Sprache gibt es zwischen Wörtern meistens keine Pause
  - Pausen kommen in spontaner Sprache auch innerhalb von Wörtern vor

Problem: Varianz des Signals

Sprachexterne Einflüsse verändern das Signal

- Raumakustik, Entfernung
- Medium: Face-to-Face, Telefon, Handy
- Mikrophonqualität und -charakteristik
- Störgeräusche („Rauschen“, „Noise“)

Problem: Varianz der Realisierung

Gleicher Laut wird nicht immer gleich ausgesprochen

- Verschiedene Dialekte
- Verschiedene Sprecher
- Unterschiedliche Sprechgeschwindigkeit
- Physischer und emotionaler Zustand des Sprechers
- Kontext, in dem ein Laut/Wort auftritt
  - z.B. Auslautverhärtung
    - gab = [ga:p]
  - z.B. umgangssprachliche Aussprache
    - haben = [haben, haben, ham, han, ...]
  - z.B. Reduktion von Funktionswörtern
    - für = [fa], wegen = [we]

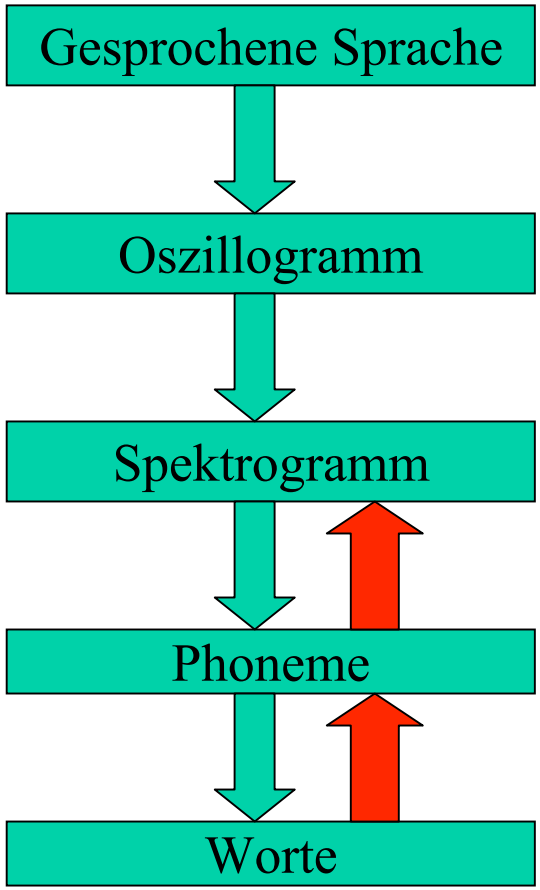
Beispiel: Dialekte

Tony Blair repeatedly passed up opportunities to put a brake on the rush to war in Iraq. (The Guardian, 07.11.05)

- British / English 
- American 
- Indian 
- Tschörmen 

Quelle: Nuance RealSpeak Demo

Lauterkennung: Wissen hilft



Sprachmodell: Welche Ausgaben sind wahrscheinlicher, welche weniger wahrscheinlich?

- Finde optimale Ausgabe gegeben die Eingabe und das Sprachmodell
  - Wissen über Wörter (Lexikon)
  - Wissen über zulässige Lautkombinationen
  - Usw.

- Hilft bei der Segmentierung:
  - "Ich esse" vs. "I chesse"

- Hilft bei der Identifikation von Phonemen:
  - [ç] vs. [ʃ] in Straße

Lauterkennung, Option 1: regelbasiert

- Idee: Bestimme Laute und wende Regeln an
  - Wenn Lautkette mit [çt] oder [ʃ t] beginnt, ist Wort st-
  - [habn] = haben
  - [fa] = für
  - Usw.
- Probleme
  - Manche Regeln sind optional (Korrektheitsproblem): Führt zu falschen Analysen
    - [t] am Wortende kann d sein, muss aber nicht: [bat] = Bad oder bat
  - Vollständiger Regelsatz unmöglich zu erhalten (Abdeckungsproblem)
    - Idiosynkratische Abweichungen
    - Gesprochene Sprache sehr flexibel „hamwa“, „isses“

Regelbasiertes Verfahren in der Praxis nicht möglich

Lauterkennung, Option 2: statistisch

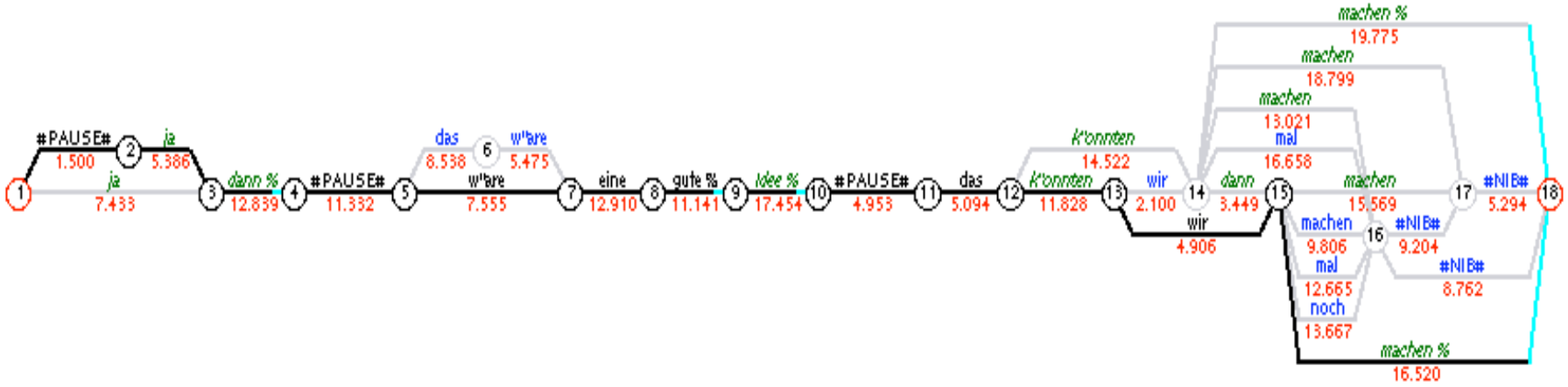
- Idee: Regelmäßigkeiten werden nicht explizit repräsentiert, sondern automatisch aus Datenquelle gelernt
- Grundlage des Lernens: annotierte Spektrogramme
  - Phonetiker ergänzen eine grosse Anzahl Spektrogramme mit Phonemen
- Maschinelles Lernen:
  - Daten in sehr kurze Zeitscheiben (zB 50 ms) einteilen
  - Intensität von ca. 25 Frequenzbändern bestimmen (Formanten!)
  - Lernen eines Modells der Korrelation zwischen Merkmalen und Phonemen (Hidden Markov Model, HMM)
    - Mit welchem Phonem kamen bestimmte Merkmalskombinationen am häufigsten vor?
    - In welchem Kontext?
- Anwendung auf neue Daten
  - Eingabe: Merkmale der Zeitscheibe (und des Kontextes)
  - Ausgabe: wahrscheinlichstes Phonem

Erkennerausgaben

- Die „beste Kette“ (oder die n besten Ketten), ggf. mit „Konfidenzwert“ (einem Maß für die Verlässlichkeit der Hypothese).

- Alternativ: Ein Worthypothesengraph: Auf der Zeitachse werden die „geratenen“ Wörter mit ihrem zugehörigen Zeitintervall und einem Wahrscheinlich-keitswert abgetragen.

Ein Worthypothesengraph (WHG)



Quelle: Verbmobil, Terminvereinbarungsdialoge:

„*Ja, das wäre eine gute Idee. Das könnten wir dann machen*“

Stand der Spracherkennungstechnik

- Maß für die Erkennerperformanz: Wortfehlerrate (wieviele Wörter der „besten Kette“ wurden falsch verstanden/gar nicht verstanden/hinzuphantasiert?)
- Wortfehlerrate hängt von der verfügbaren Verarbeitungszeit und verschiedenen externen Faktoren ab.
- Bei gängigen Systemen kann man mit Echtzeitverhalten (Verarbeitungszeit ≤ Sprechzeit) und einer Wortfehlerrate in der Größenordnung von deutlich unter 10 % rechnen.

Erkennerperformanz ist abhängig von:

- Sprechmodus: Einzelwort, kontinuierlich, spontan
- Sprecherbindung: abhängig, unabhängig, adaptiv
- Größe des Lexikons:

  LIFT: ca. 150 Wortformen

  Verbmobil: ca. 10000 Wortformen

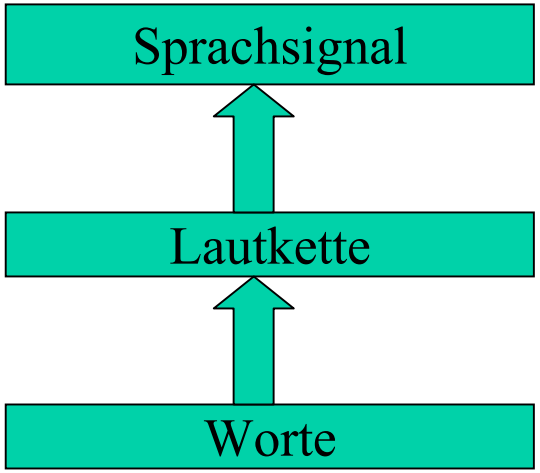
  Diktiersysteme: ab 50000 Wortformen
- Perplexität: Maß für die Uniformität der Eingabe

  beschränkte Domäne, gesteuerter Dialog: niedrige Perplexität

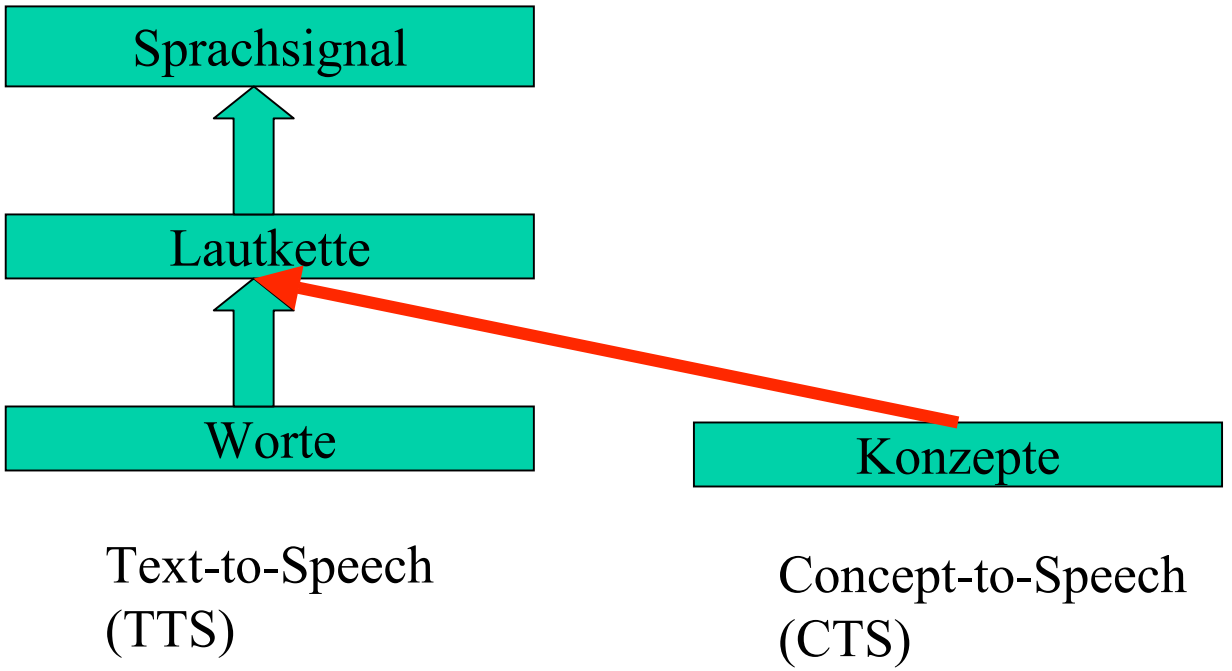
  keine Domänenbeschränkung, freie Rede: hohe Perplexität
- Eingabequalität
- Verarbeitungszeit: online, offline

Teil 2: Sprachsynthese

Einfacher als Spracherkennung?



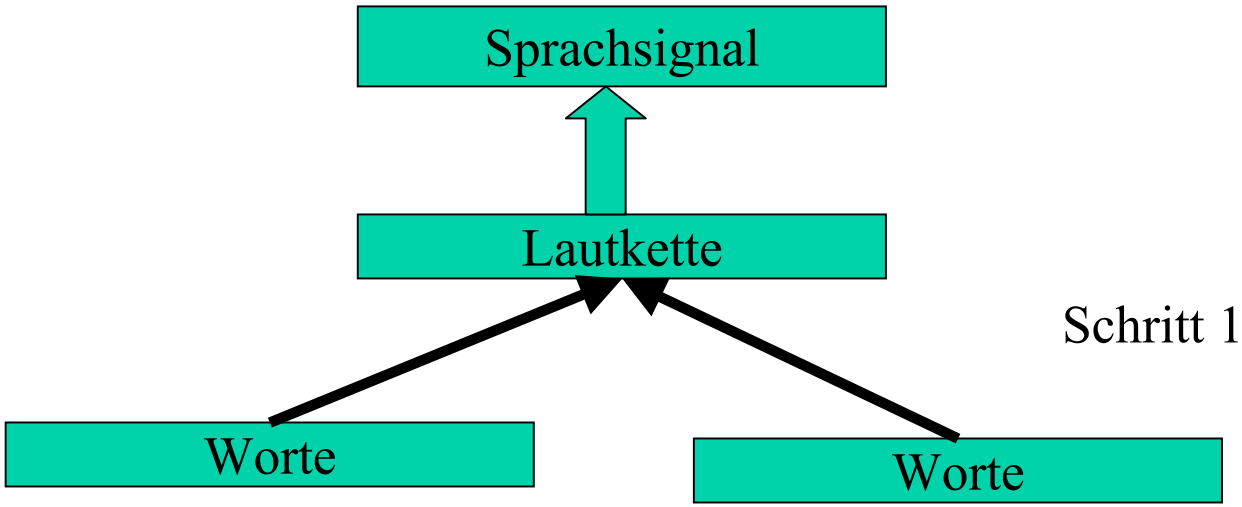
Einfacher als Spracherkennung?



TTS vs. CTS

- Text-to-speech (TTS): Lese Text vor
  - Vorlesesoftware für Blinde
- Concept-to-speech (CTS): Verbalisiere Sachverhalt (z.B. Datenbank)
  - Dialogsystem (Fahrplanauskunft)
- Vorteil / Probleme?

Sprachsynthese: segmentale Information



Schritt 1: Lauterzeugung

Text-to-Speech
(TTS)

Concept-to-Speech
(CTS)

Lauterzeugung: Zwei alternative Techniken

Wortkonkatenation: Wörter bzw. Äußerungen werden mit Sprechern voraufgenommen und geschnitten. Wörter werden bei der Synthese zusammengehängt, bzw. in ein Äußerungsmuster eingehängt und prosodisch modifiziert.

Problem: Satz-Prosodie, Erweiterbarkeit

```
Der Flug | Lufthansa | LH | 4 | 7 | 5 | 2 |
  um | 11 Uhr | 35 | aus | Dresden | ist um |
  2 Stunden | und 25 Minuten | verspätet.
```

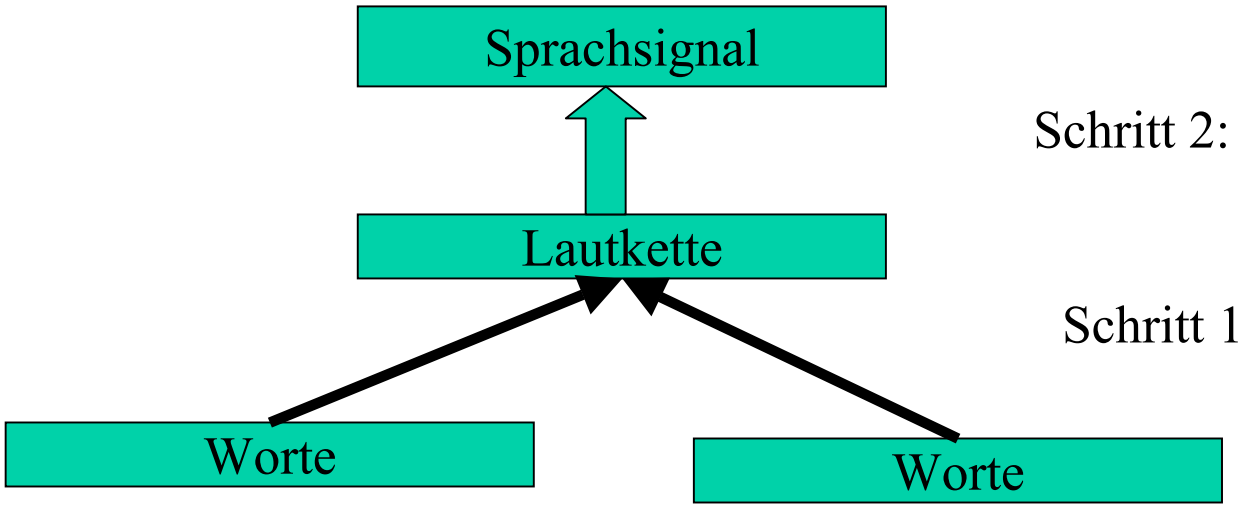
Synthese: Problem: Vollsynthese von Wörtern aus einzelnen Lauten ist unmöglich, weil eine scharfe Begrenzung zwischen Lauten (Koartikulation!)

Diphon-Synthese (vereinfacht): Alle möglichen Kombinationen von zwei Lauten werden voraufgenommen, geschnitten , ...

Probleme der Lauterzeugung

- Orthographie (bei TTS)
  - Keine direkte Korrespondenz Schriftbild - Lautbild
- Generierung eines Satzes aus semantischer Repräsentation (bei CTS)
- Fremdsprachige Wörter, Zahlausdrücke, Sonderzeichen, Abkürzungen, Akronyme..
  - USA vs. UNO

Sprachsynthese: suprasegmentale Information



Prosodie

- Das Sprachsignal enthält zusätzlich zur „segmentalen Struktur“, d.h. Information über die Abfolge von Lauten, die Wörter identifizieren, „**suprasegementale**“/„prosodische“ Information:
  - Sprechgeschwindigkeit, Rhythmus, Pausen
  - Akzent
  - Intonation
- Funktionen prosodischer Information:
  - Gliederung des Satzes
  - Satzmodus (Aussage, Frage, Befehl)
  - Text- und Dialogkohärenz, **Informationsstruktur**

Informationsstruktur

- **Jeder** Mann liebt eine Frau.
- Jeder **Mann** liebt eine Frau.
- Jeder Mann **liebt** eine Frau.
- Jeder Mann liebt **eine** Frau.
- Jeder Mann liebt eine **Frau**.

Richtige Rhythmik und Intonation setzen sehr viel linguistisches und Weltwissen voraus

Sprachsynthese und –erkennung: Zum Ausprobieren

- Sprachsynthese:
  - Nuance (war: ScanSoft (war: Rhetorical Systems))
    http://www.nuance.com/realspeak/demo/default.asp
  - Logox:
    http://www.logox.de/cgi-bin/speechform.cgi
- Dialogsysteme:
  - Deutsche Bahn (Philips)
    0241 - 60 40 20
  - Fränki Kino
    09131 - 6161116
  - IBM Staumelder
    06221 - 593129