

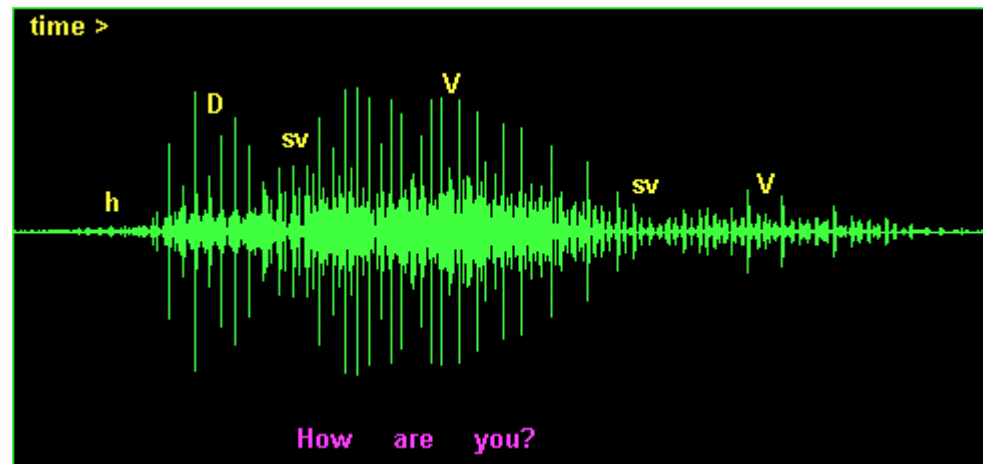
Erkennung und Erzeugung gesprochener Sprache

18.11.2003

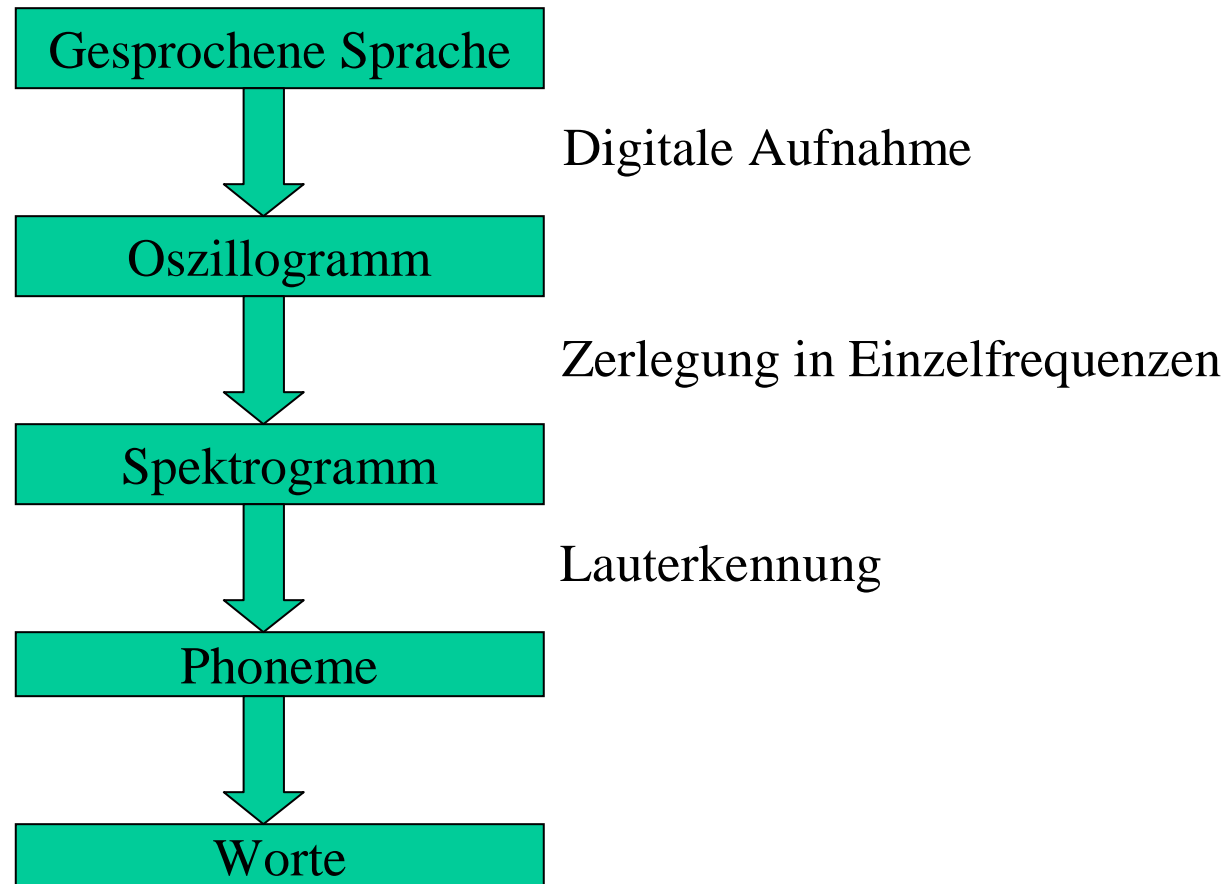
Sebastian Pado

Spracherkennung

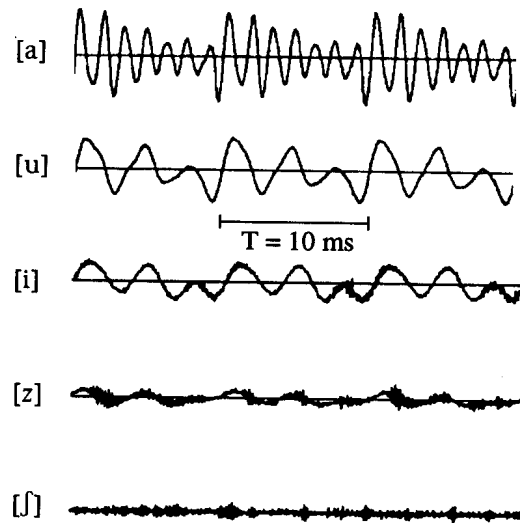
- Die Grundaufgabe der Spracherkennung: Gegeben ist ein kontinuierliches Schallsignal. Welche Kette von Wörtern wurde vom Sprecher geäußert?
- Beispiel: Das Oszillogramm für eine Äußerung von „How are you“



Spracherkennung

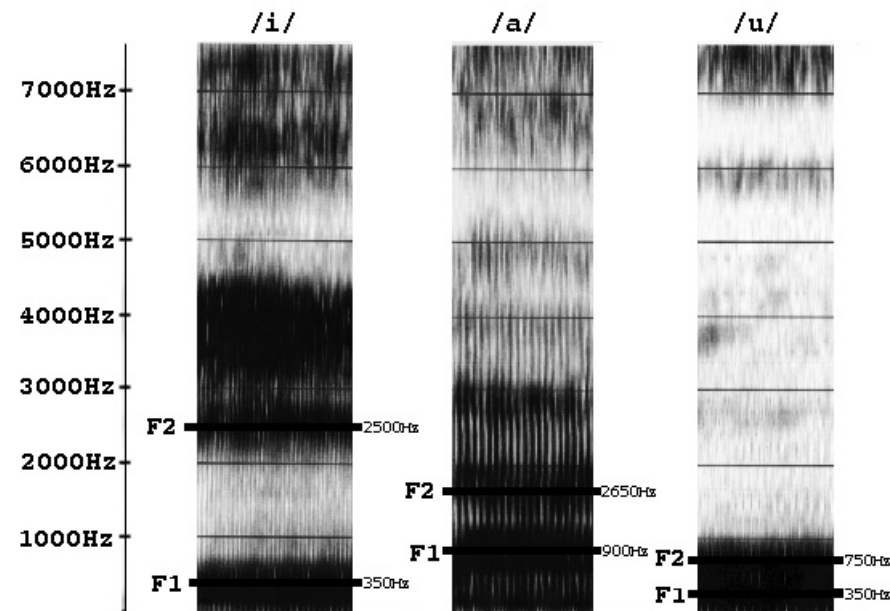


Ein paar Oszillogramme



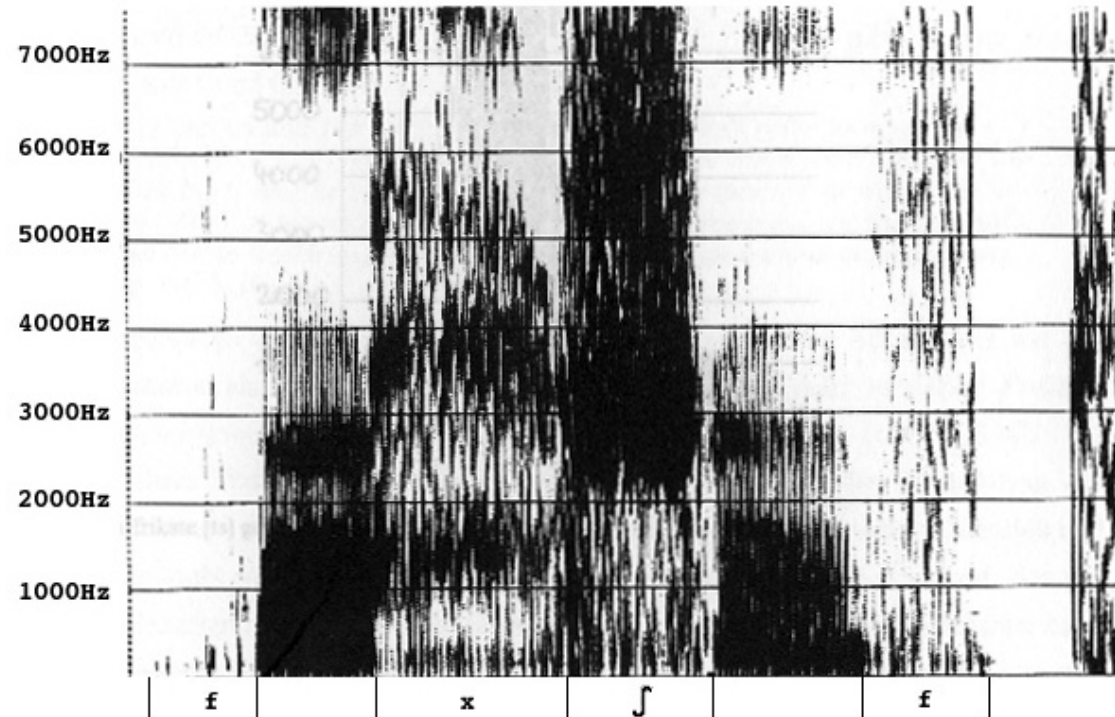
Anmerkung zur Notation: Eckige Klammern werden zur Bezeichnung von Lauten/phonetischen Einheiten verwendet, Schrägstriche zur Bezeichnung von Phonemen (funktionalen Einheiten der Phonologie). Die Notation geht, wie die Beispiele zeigen, manchmal durcheinander. Zum grundsätzlichen Unterschied s. Einführung in die Sprachwissenschaft.

Spektrogramm für die Vokale i,a,u



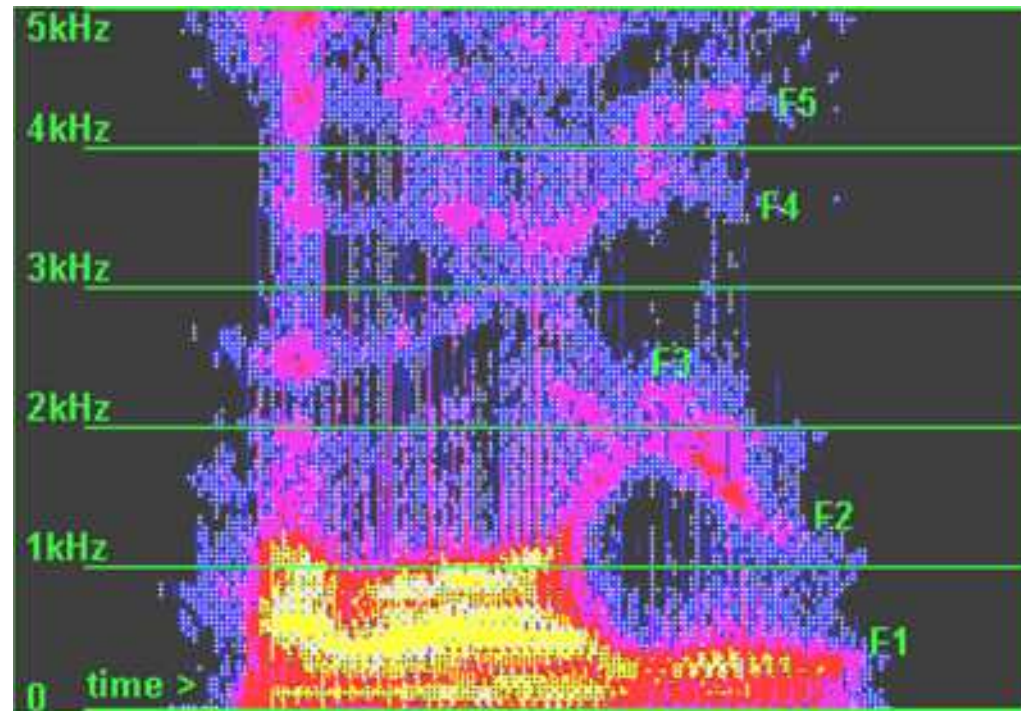
- Dunkle Färbung: große Schallenergie in einem bestimmten Frequenzbereich.
- Die **Formanten** (Obertöne) F1 und F2 sind für die charakteristische Vokalqualität verantwortlich sind.
- Der Verlauf des **Basisformanten** F0 (hier nicht sichtbar) gibt die Intonation der Äußerung wieder.

Spektrogramm für einige Konsonanten



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

Ein buntes Spektrogramm



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

Lauterkennung

Aufgabe: Übersetze Spektrogramm in Äußerung

Problem: Kontinuität des Signals

Gesprochene Sprache lässt sich schwer unterteilen

- Die **Laute** eines Wortes lassen sich nicht gegeneinander abgrenzen (**Koartikulation**)
 - Man kann in Lautfolgen wie [am], [um], [an] nicht den Vokal vom Nasal trennen: Man hört mit dem Vokal die Nasal-Qualität mit und umgekehrt.
- **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

Gleicher Laut hört sich nicht immer gleich an

- Raumakustik, Entfernung
- Medium: Face-to-Face, Telefon, Handy
- Mikrophon-Qualität und -Charakteristik
- Störgeräusche („Rauschen“, „Noise“)

Problem: Varianz der Realisierung

Gleiches Wort 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

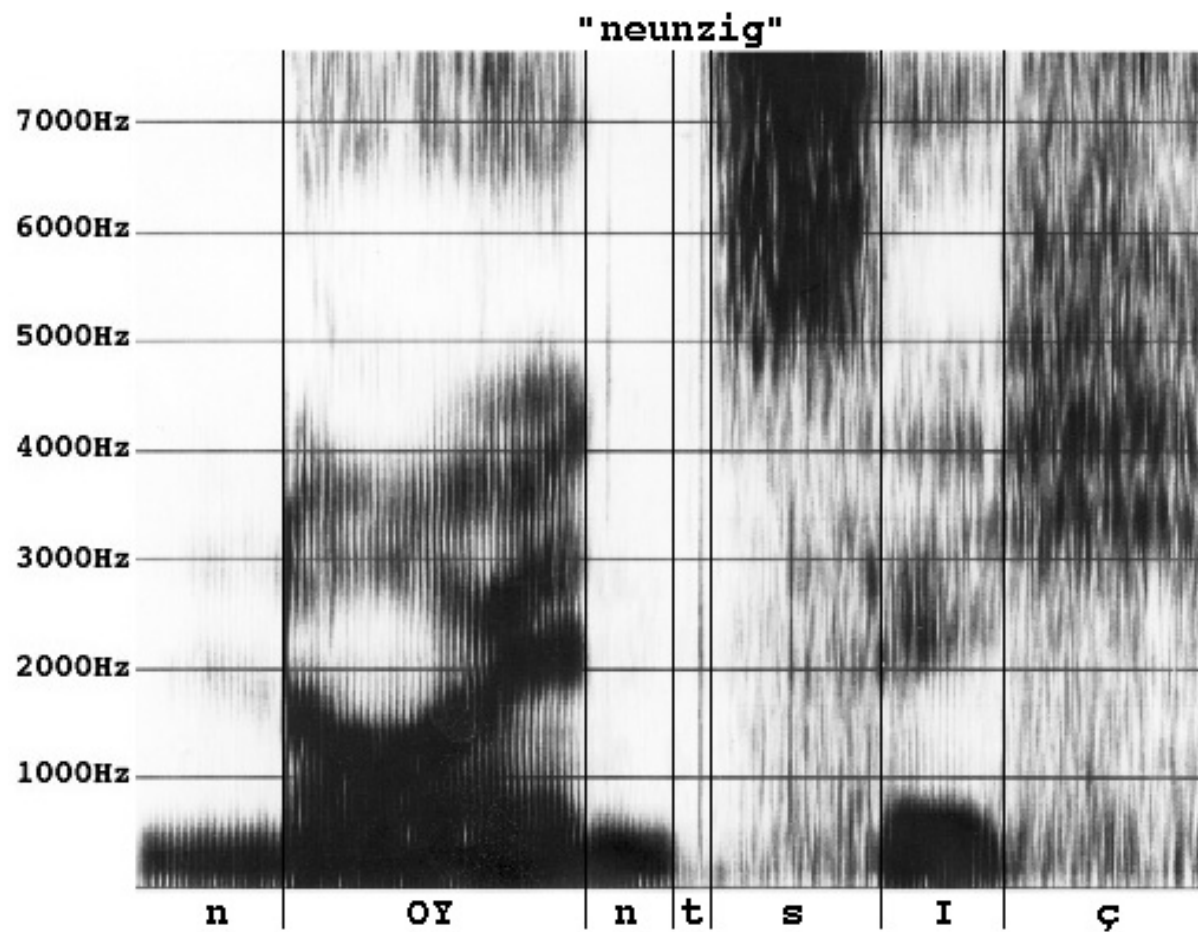
A majority of Labour voters welcome President George Bush's state visit to Britain which starts today, according to November's Guardian/ICM opinion poll. (The Guardian, 18.11.03)

- British
- Scottish
- American
- Australian



Quelle: Rhetorical Systems Demo Speech Synthesis

Spektrogramm für ein deutsches Wort



Lauterkennung

- Regelbasiert in der Praxis nicht möglich
- Statistisches Verfahren: **Lerne** Abbildung
 - **Merkmalsextraktion**: Bestimmung der Schallenergie in einzelnen Frequenzfenstern (z.B. Viertelton) und Zeitfenstern (z.B. 50 ms).
 - Manuelle phonetische **Transkription** von Sprachaufnahmen als Trainingsdaten
 - Trainiere **Hidden-Markov-Modell** auf Sequenzen von Merkmalsvektoren (Eingabe) und Transkriptionen (Ausgabe)
 - Modell (**Erkenner**) **klassifiziert** neue Eingaben

Finden der wahrscheinlichsten Sequenz (I)

- Gegeben eine Eingabe-Sequenz $I=(i_1,i_2,\dots)$ [Merkmale], finde die **wahrscheinlichste** Ausgabe-Sequenz $O=(o_1,o_2,\dots)$ [Laute]:

$$\max_O P(O|I) \quad \text{„das } O, \text{ für das } P \text{ maximal“}$$

- Bayessches Regel: Man kann die Wahrscheinlichkeiten umdrehen

$$\max_O P(O|I) = \max_O \frac{P(I|O)P(O)}{P(I)} = \max_O P(I|O)P(O)$$

- $P(O)$ ist die Wahrscheinlichkeit der Lautkette
 - linguistisches Wissen über **Grammatikalität** (**Sprachmodell**)
- $P(I|O)$ ist die Wahrscheinlichkeit, einen Merkmalsvektor für einen bestimmten Laut gesehen zu haben
 - Kann aus den **Trainingsdaten** bestimmt werden

Finden der wahrscheinlichsten Sequenz (II)

HMM hat Zustände für Laute

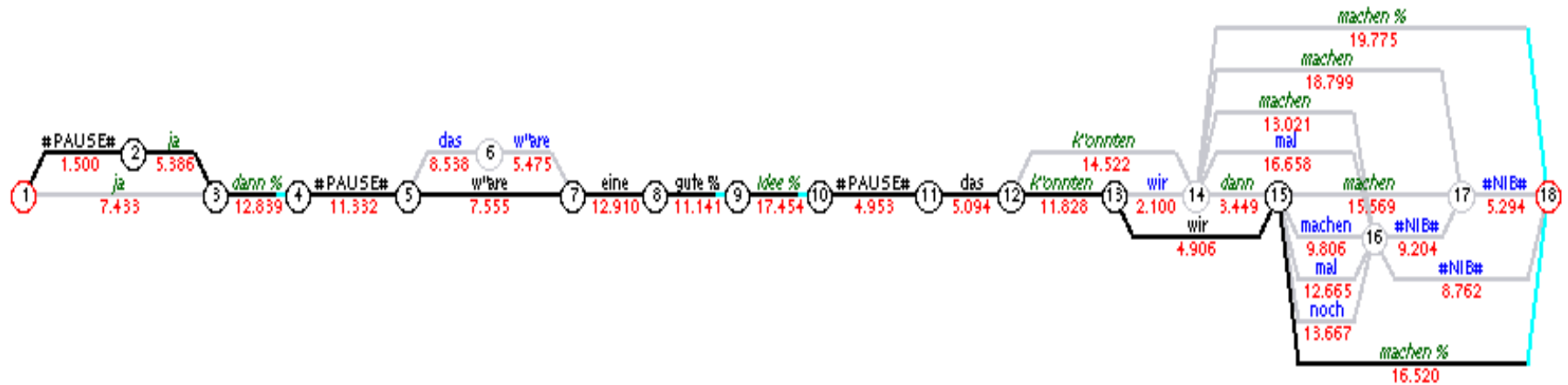
- Jeder Zustand S_n erzeugt einen Merkmalsvektor I_n mit Emissionswahrscheinlichkeit $P(I_n|S_n)$
- Jeder Zustand S_{n-1} ändert sich in Zustand S_n mit Übergangswahrscheinlichkeit $P(S_n|S_{n-1})$
- Training: lese alle $P(I_n|S_n)$ und $P(S_{n+1}|S_n)$ aus Trainingsdaten ab
- Anwendung: gegeben finde die wahrscheinlichste Lautsequenz = Zustandssequenz, die die gegebene Eingabesequenz erzeugt

$$\max_O P(I|O) = \max_O \prod_n P(S_n|S_{n-1})P(I_n|S_n)$$

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 Wahrscheinlichkeitswert abgetragen.

Ein Worthypothesengraph (WHG)



Quelle: Verbmobil, Terminvereinbarungsdialege:

„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 ($\text{Verarbeitungszeit} \leq \text{Sprechzeit}$) 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

Prosodie

- Das Sprachsignal enthält zusätzlich zur „segmentalen Struktur“, d.h. Information über die Abfolge von Lauten, die Wörter identifizieren, „suprasegmentale“ oder „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**.

Sprachsynthese

- Sprachsynthese erscheint einfacher als Spracherkennung. Trotzdem gibt es bisher keine Vollsynthese-Systeme mit einer Qualität, die an menschliche Sprecher heranreicht.
- Zwei Ansätze:
 - Text-to-speech (TTS): Lese Text vor
 - Concept-to-speech (CTS): Verbalisiere Sachverhalt (z.B. Datenbank)

Probleme der Sprachsynthese

- Fremdsprachige Wörter, Zahlen, Sonderzeichen
- Prosodie: Die richtige Rhythmik und Intonation setzt linguistisches und konzeptuelles Wissen voraus
- Orthographie (bei TTS)
- Generierung eines Satzes aus semantischer Repräsentation (bei CTS)

Sprachsynthese: 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 usw.

Sprachsynthese und –erkennung: Zum Ausprobieren

- Sprachsynthese:
 - Rhetorical Systems:
<http://www.rhetorical.com/cgi-bin/demo.cgi>
 - Logox:
<http://www.logox.de/cgi-bin/speechform.cgi>
- Dialogsysteme:
 - Deutsche Bahn (Philips)
[0241 – 60 40 20](tel:0241-604020)
 - SBB
[0041157 02 22](tel:00411570222)