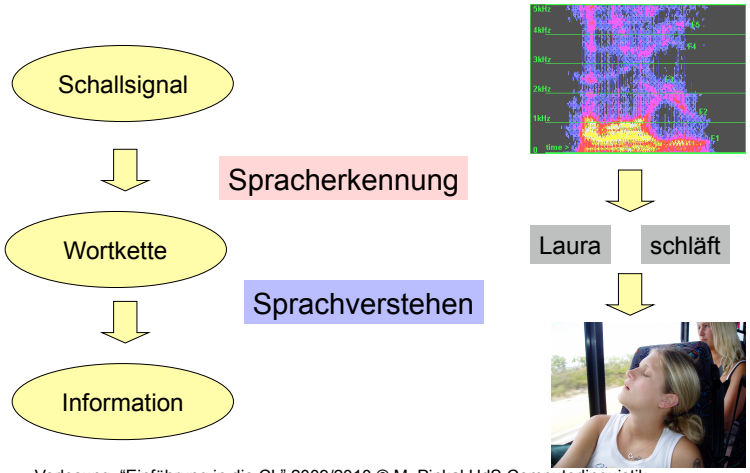
Einführung in die Computerlinguistik

Verarbeitung gesprochener Sprache

WS 2009/2010

Manfred Pinkal

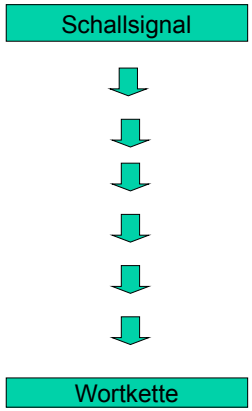
Was ist Sprachverarbeitung?



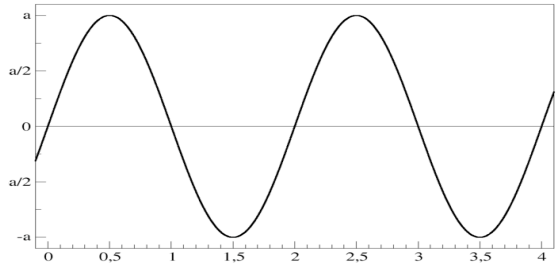
Grundaufgabe der Spracherkennung

- Die Grundaufgabe der Spracherkennung: Gegeben ist ein kontinuierliches Schallsignal. Welche Kette von Wörtern wurde vom Sprecher geäußert?

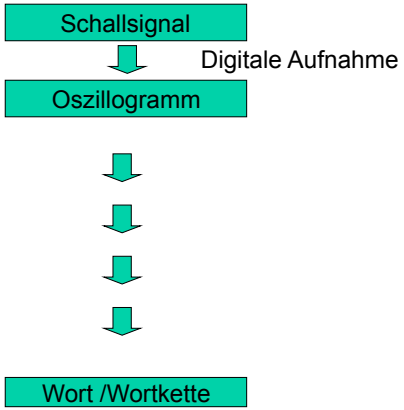
Spracherkennung: (Vereinfachtes) Schema



Reine Schwingung

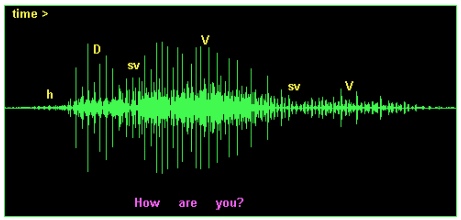


Spracherkennung: (Vereinfachtes) Schema

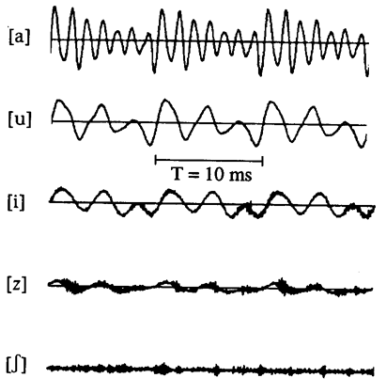


Ein Oszillogramm

- Das Oszillogramm für eine Äußerung des englischen Satzes „How are you"

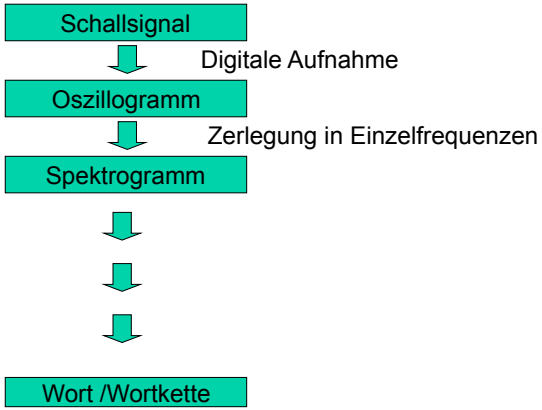


Einzelne Laute als Oszillogramme

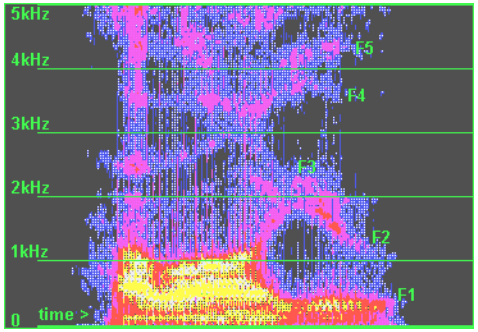


- Laute werden charakterisiert durch Kombination von Schwingungen verschiedener Frequenzen
- Im Oszillogramm **schwer erkennbar** (Überlagerung)
- Daher: Geschicktere Repräsentation durch Komponentenanalyse (Fourier-Transformation)
- Ergebnis: Zeit-Frequenz-Diagramm (**Spektrogramm**)

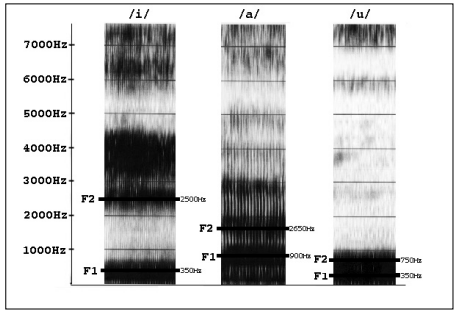
Spracherkennung: (Vereinfachtes) Schema



Ein Spektrogramm

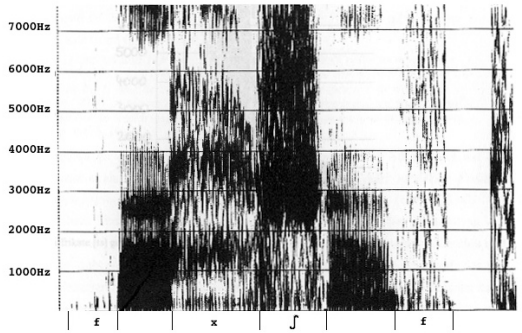


Spektrogramm für die Vokale i,a,u



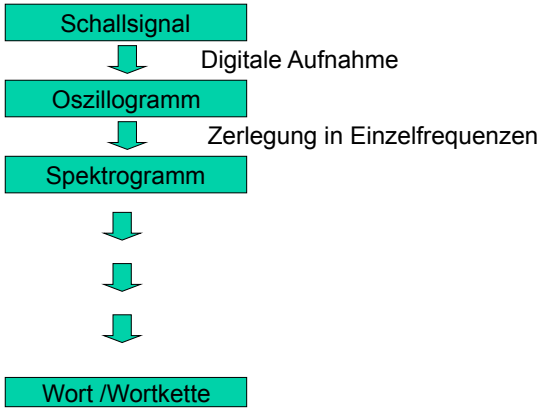
- Dunkle Färbung: große Schallenergie in einem bestimmten Frequenzbereich.
- Die Formanten (Obertöne) F1 und F2 sind für die charakteristische Vokalqualität verantwortlich.
- Der Verlauf des Basisformanten F0 (hier nicht sichtbar) gibt die Intonation der Äußerung wieder.

Spektrogramm für einige Konsonanten

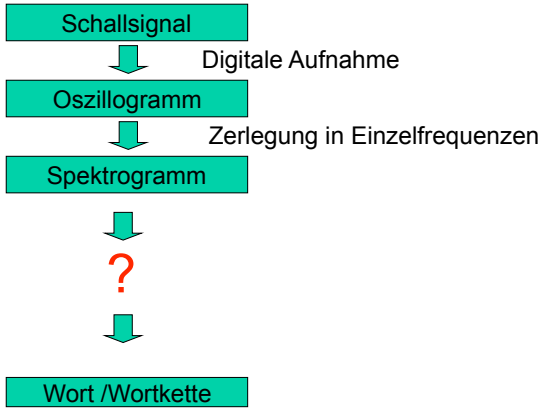


Frikative: f und ch-Laut („ach"-Laut); Sibillant: „sch"-Laut

Spracherkennung: (Vereinfachtes) Schema



Spracherkennung: (Vereinfachtes) Schema

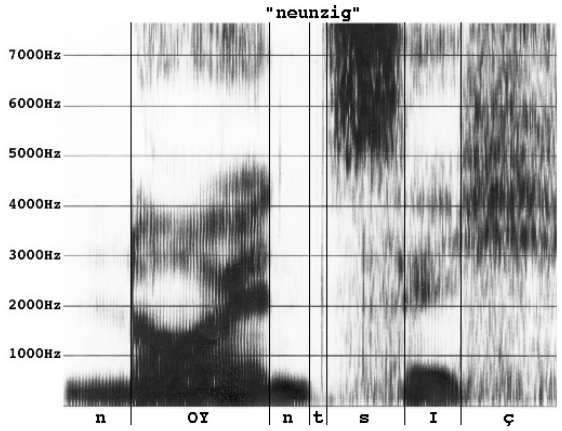


Worterkennung: Ein Versuch

- 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!

Spektrogramm für ein deutsches Wort



Problem 1: Kontinuität des Signals

- Die Laute eines Wortes lassen sich schwer abgrenzen
  - Wo hört Laut 1 auf, wo fängt Laut 2 an?
  - Dazu kommt das Phänomen der 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 2: Varianz der Realisierung

- Gleicher Laut/ gleiches Wort wird nicht immer gleich ausgesprochen
  - Verschiedene Dialekte
  - Verschiedene Sprecher
  - Unterschiedliche Sprechgeschwindigkeit
  - Physischer und emotionaler Zustand des Sprechers
  - Abhängig von Tonhöhe und Akzent

Problem3: Varianz des Signals

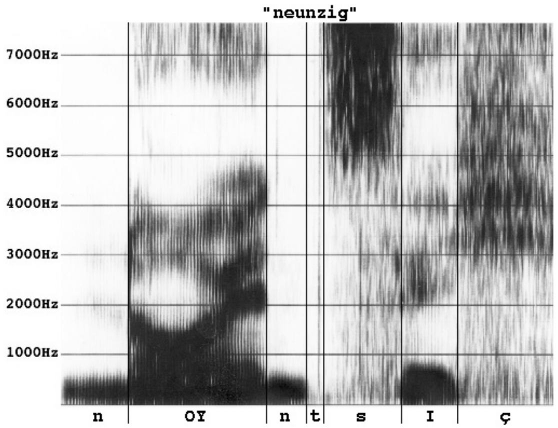
- Sprachexterne Einflüsse verändern das Signal
  - Raumakustik, Hall, Entfernung
  - Medium: Face-to-Face, Telefon, Handy
  - Mikrofonqualität und -charakteristik
  - Hintergrundgeräusche („Rauschen", „Noise")

Statistische Modellierung

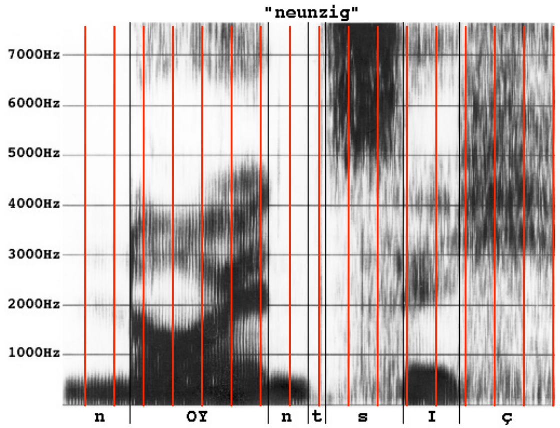
Spracherkennung ist eine Art Klassifikationsaufgabe:

- Ermittlung der wahrscheinlichsten Wortkette
  - $W = w_1 w_2 \ldots w_n$, die einem beobachteten akustischen Signal entspricht.
- Die akustische Information, die durch die Lautspektrographie gemessen wird, ist viel zu komplex für statistische Berechungen.
- Wir berechnen eine handhabbare Beschreibung der akustischen Information durch Merkmalsextraktion.

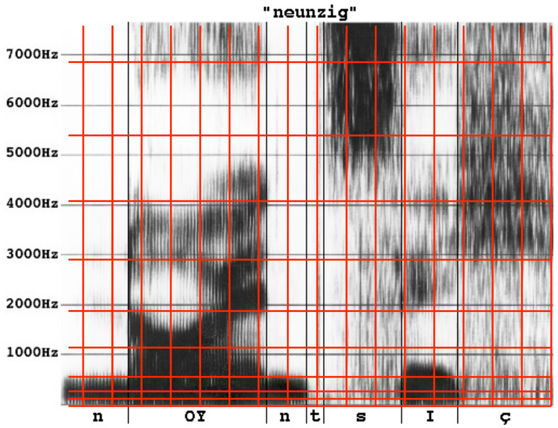
Spektrogramm für ein deutsches Wort



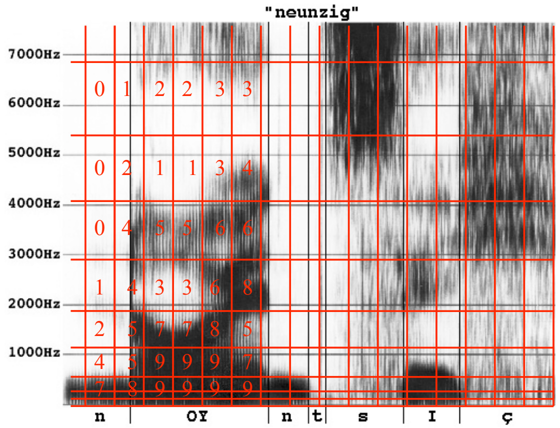
Spektrogramm für ein deutsches Wort



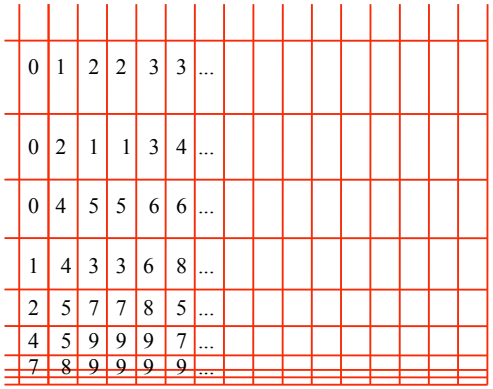
Spektrogramm für ein deutsches Wort



Spektrogramm für ein deutsches Wort



Merkmalsvektor, Ausschnitt



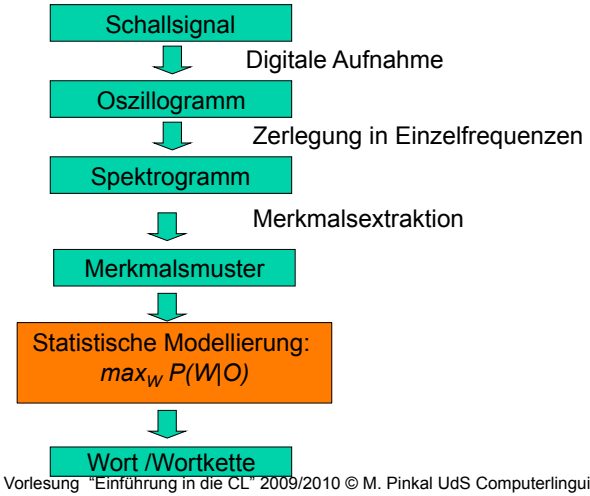
0	1	2	2	3	3	...
0	2	1	1	3	4	...
0	4	5	5	6	6	...
1	4	3	3	6	8	...
2	5	7	7	8	5	...
4	5	9	9	9	7	...
7	8	9	9	9	9	...

Merkmalsextraktion

- Bestimmung der Schallenergie in einzelnen Frequenzfenstern (z.B. Viertelton) und Zeitfenstern (z.B. 20 ms).
- Resultat ist eine Folge $O = o_1\ o_2\ ...\ o_m$ von (Einzel-) „Beobachtungen"
- Jedes o_i ist ein Merkmalsvektor, der die Schallenergie für die unterschiedlichen Frequenzfenster in einem bestimmten Zeitfenster angibt.
- Die erkannte Wortkette ist $max_W\ P(W|O)$: „das W, für das $P(W|O)$ maximal ist", oder ausbuchstabiert:

  $max_W\ P(W|O) = P(w_1 w_2\ ...\ w_n|\ o_1\ o_2\ ...\ o_m)$

Spracherkennung: (Vereinfachtes) Schema



Das Bayessche Theorem

- Wie bestimmen wir $P(W|O)$?
- Wie schließen wir vom akustischen Merkmalsmuster O auf die Wahrscheinlichkeit der tatsächlich geäußerten Wortkette W?
- Das Bayessche Theorem oder die Bayes-Regel:

$$P(E \mid F) = \frac{P(F \mid E) \cdot P(E)}{P(F)}$$

- Die Bayes-Regel ist ein elementares Gesetz der Wahrscheinlichkeitstheorie. Sie ist überall da nützlich, wo der Schluss von einer Größe F auf eine andere Größe E bestimmt werden soll (typischerweise von einer Wirkung/ einem Symptom auf die mögliche Ursache), die Abhängigkeit in der anderen Richtung (von der Ursache auf die Wirkung) aber besser zugänglich ist.

Das Bayessche Theorem

- Akustisches Merkmalsmuster *O* : Symptom
- Tatsächlich geäußerte Wortkette *W* : Ursache
- Mit Bayes-Regel :

$$P(W \mid O) = \frac{P(O \mid W) \cdot P(W)}{P(O)}$$

- Die wahrscheinlichste Wortkette:

$$\max_W P(W \mid O) = \max_W \frac{P(O \mid W) \cdot P(W)}{P(O)}$$

$$= \max_W P(O \mid W) \cdot P(W)$$

Akustisches Modell und Sprachmodell

$$\max_W P(W \mid O) = \max_W P(O \mid W) \cdot P(W)$$

- *P(O|W)* ist die Wahrscheinlichkeit, dass eine Wortfolge in einer bestimmten (durch den Merkmalsvektor bezeichneten) Weise ausgesprochen wird: Akustisches Modell
- *P(W)* ist die Wahrscheinlichkeit, dass eine bestimmte Wortfolge geäußert wird: „Sprachmodell"

Akustische Modelle

$$\max_W P(W \mid O) = \max_W P(O \mid W) \cdot P(W)$$

- Training von „Lautmodellen" auf Datensammlungen für gesprochene Sprache: Aufnahmen von Sprachlauten mit ihrer phonetischen Kategorie/ Umschrift
- Aussprachewörterbuch, das für jedes Wort die phonetische Umschrift enthält
  - Genauer: Die Umschrift für alternative Aussprachen, die in einem gewichteten endlichen Automaten kodiert sind.
- Für die statistische Zuordnung von Merkmalsmustern und Wörtern wird die HMM-Technik („Hidden Markov Models") verwendet.

Sprachmodelle

$$\max_W P(W \mid O) = \max_W P(O \mid W) \cdot P(W)$$

- Wie berechnen wir $P(W) = P(w_1 w_2 \ldots w_n)$?
- Grundlage ist die Frequenz von Wortfolgen in Korpora.
- Sparse-Data-Problem: Ganze Sätze kommen viel zu selten vor.
- Kettenregel erlaubt die Reduktion von $P(w_1 w_2 \ldots w_n)$ auf bedingte Wahrscheinlichkeiten:

  $P(w_1 w_2 \ldots w_n)$

  $= P(w_1) * P(w_2 \mid w_1) * P(w_3 \mid w_1 w_2) * \ldots * P(w_n \mid w_1 w_2 \ldots w_{n-1})$

  aber:

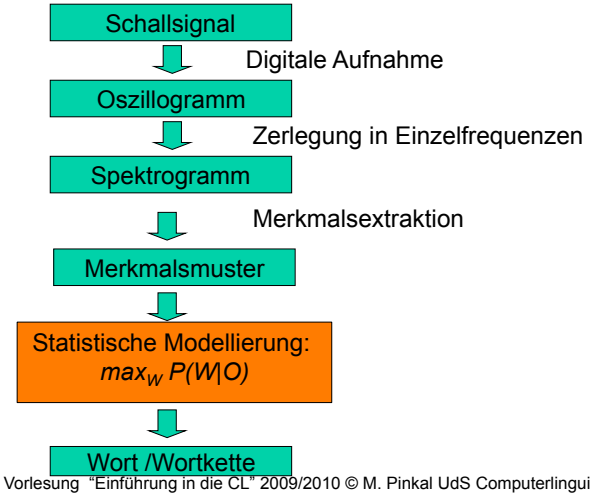
- $P(w_n \mid w_1 w_2 \ldots w_{n-1})$: Sparse-Data-Problem ist nicht beseitigt!

n-Gramme

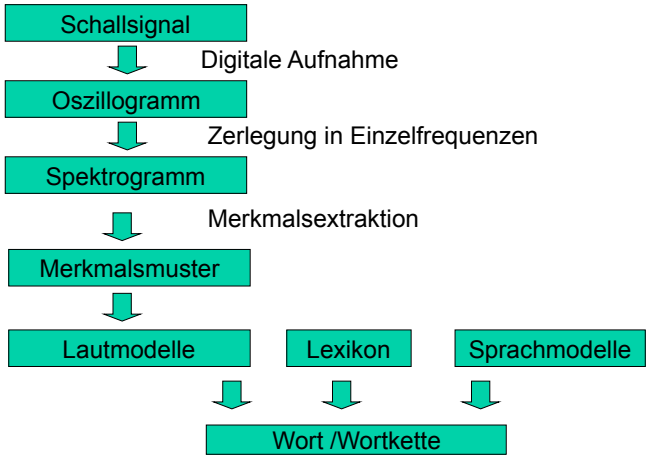
- n-Gramm-Technik:
  - Wir approximieren die Wahrscheinlichkeit, dass ein Wort w im Kontext einer beliebig langen Wortfolge auftritt, durch die relative Häufigkeit, mit der es in einem auf n Wörter begrenzten Kontext auftritt.
  - Dabei wird das Wort selbst mitgezählt. N-Gramm-Wahrscheinlichkeit berücksichtigt also einen Vorkontext von n-1 Wörtern.
- Meistens wird mit Bigrammen und Trigrammen gearbeitet.
- Beispiel Bigramm-Approximation:
  - $P(w_n|w_1w_2... w_{n-1}) \approx P(w_n|w_{n-1})$

    $P(w_1w_2 ... w_n) \approx P(w_1)*P(w_2|w_1)*P(w_3|w_2)* ... P(w_n|w_{n-1})$

Spracherkennung: (Vereinfachtes) Schema



Spracherkennung: Schema



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.
- Gängige Systeme analysieren in Echtzeit (Verarbeitungszeit ≤ Sprechzeit) und sind in der Wortfehlerrate in einem akzeptablen Bereich .

Erkennerperformanz ist abhängig von:

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

  Einfache Sprachsteuerungssysteme: 100-200 Wortformen

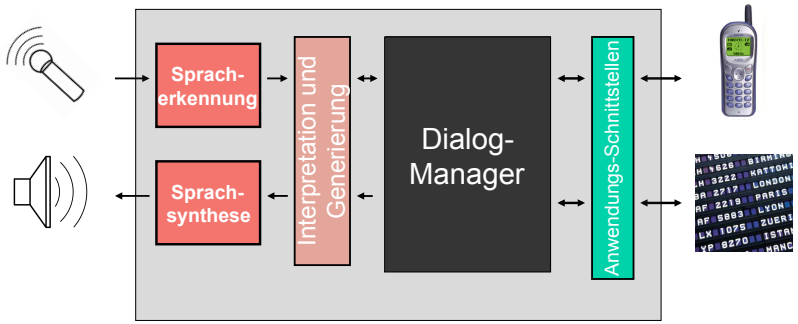
  Dialogsysteme: 500-1000 Wortformen (+ spezieller Wortschatz)

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

  beschränkte Domäne, gesteuerter Dialog: niedrige Perplexität

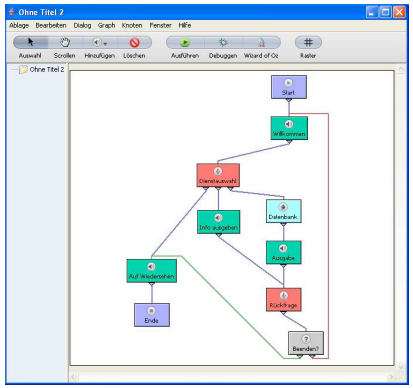
  keine Domänenbeschränkung, freie Rede: hohe Perplexität
- Eingabequalität
- Verarbeitungszeit

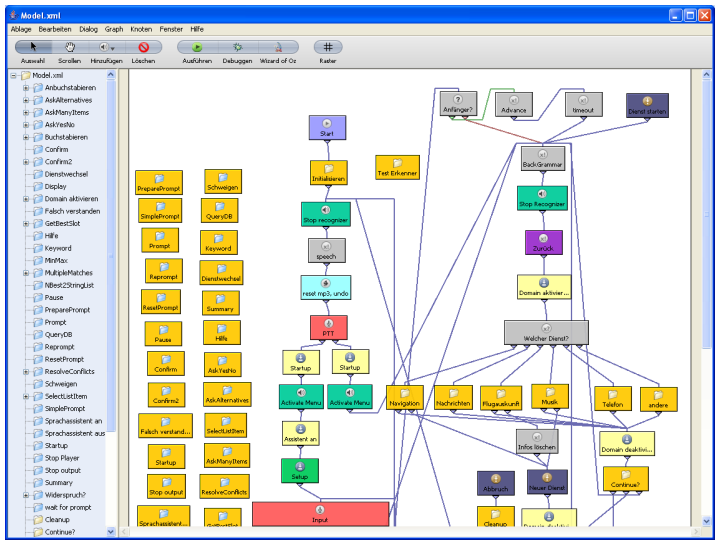
Architektur von Dialogsystemen



Dialogmodellierung

- Standard-Framework: endliche Automaten





Dialogsteuerung im Fahrzeug

Dialog im Fahrzeug