

Einführung in die Computerlinguistik

Verarbeitung gesprochener Sprache

WS 2010/2011

Manfred Pinkal

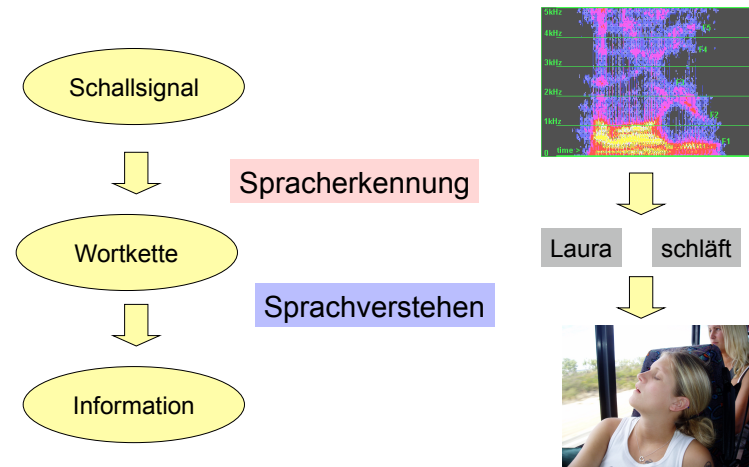
Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Grundaufgabe der Spracherkennung

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

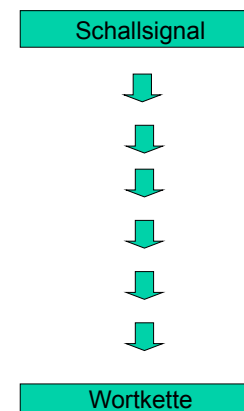
Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Was ist Sprachverarbeitung?



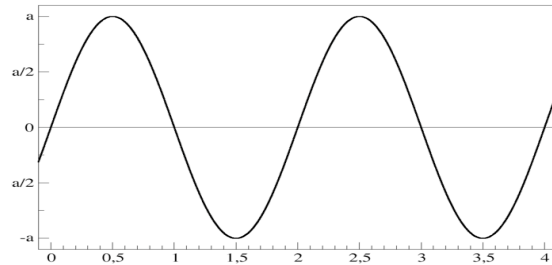
Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Spracherkennung: (Vereinfachtes) Schema

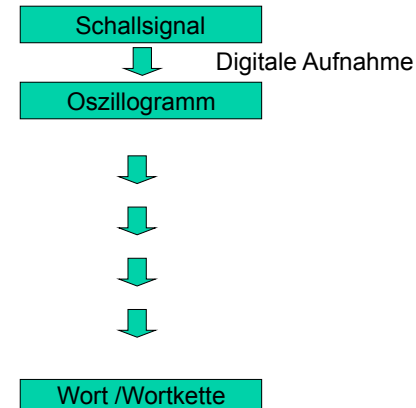


Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Reine Schwingung

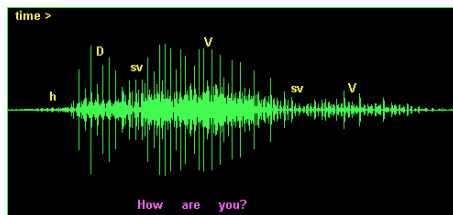


Spracherkennung: (Vereinfachtes) Schema

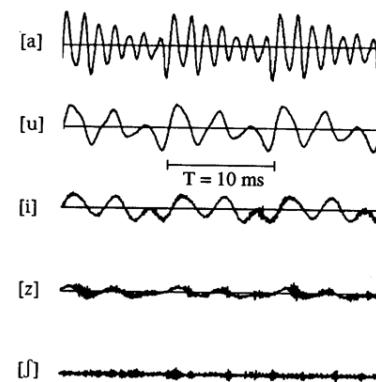


Ein Oszillogramm

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

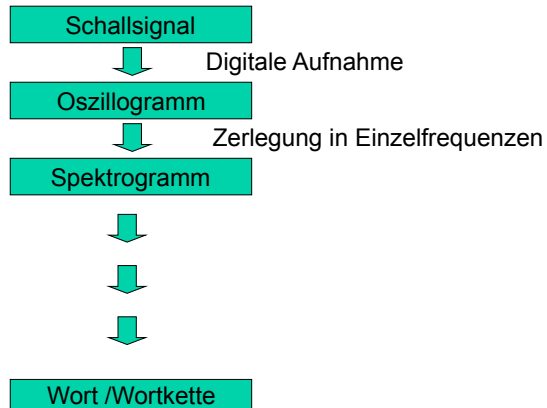


Einzelne Laute als Oszillogramme

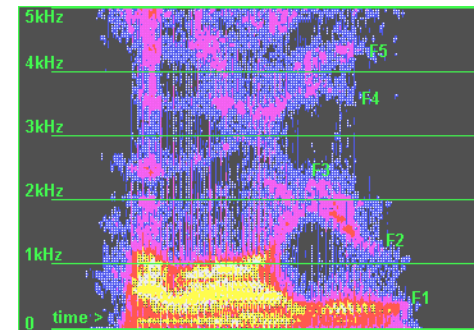


- Laute werden charakterisiert durch Kombination von Schwingungen verschiedener Frequenzen
- Im Oszillogramm **schwer erkennbar** (Überlagerung)
- Daher: Geschicktere Repräsentation durch Komponentenanalyse (Fourier-Transformation)
- Ergebnis: Zeit-Frequenz-Diagramm (**Spektrogramm**)

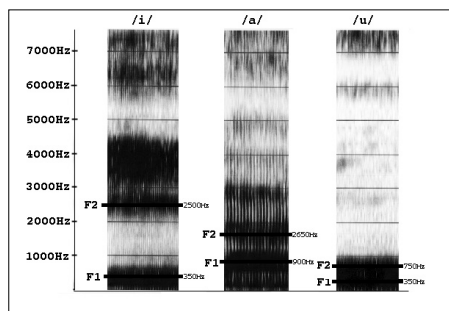
Spracherkennung: (Vereinfachtes) Schema



Ein Spektrogramm

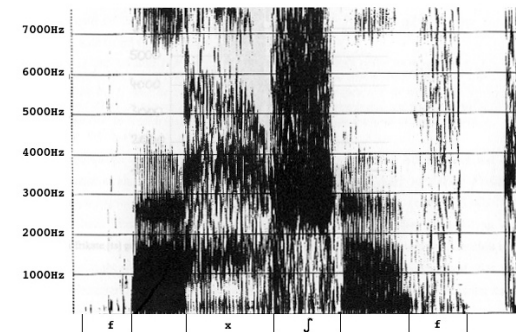


Spektrogramm für die Vokale i,a,u



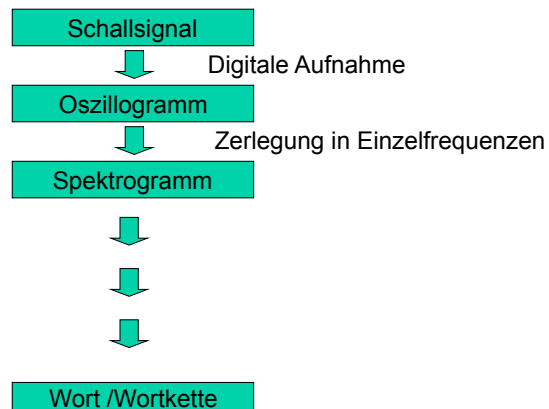
- Dunkle Färbung: große Schallenergie in einem bestimmten Frequenzbereich.
- Die **Formanten** (Obertöne) F1 und F2 sind für die charakteristische Vokalqualität verantwortlich.
- Der Verlauf des **Basisformanten** F0 (hier nicht sichtbar) gibt die Intonation der Äußerung wieder.

Spektrogramm für einige Konsonanten

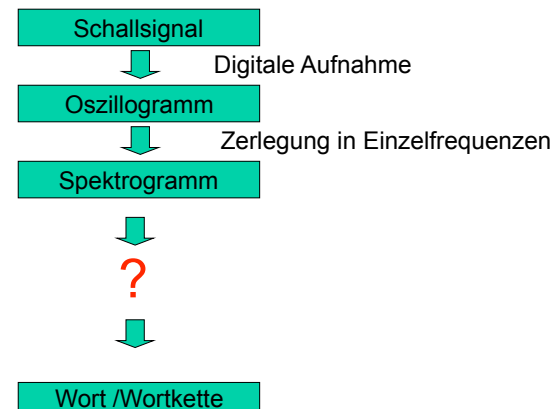


Frikative: f und ch-Laut („ach“-Laut); Sibillant: „sch“-Laut

Spracherkennung: (Vereinfachtes) Schema



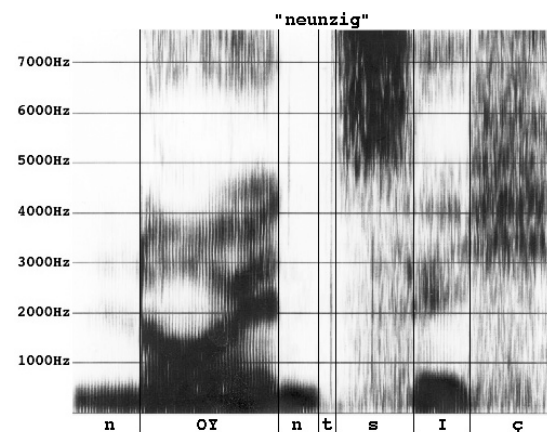
Spracherkennung: (Vereinfachtes) Schema



Worterkennung: Ein Versuch

- Schritt 1: Identifikation einzelner Spektrogramm-Schnipsel = Laute (Segmentierung)
 - Finde Grenzen im 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

Spektrogramm für ein deutsches Wort



Worterkennung: Ein Versuch

- Schritt 1: Identifikation einzelner Spektrogramm-Schnipsel = Laute (Segmentierung)
 - Finde Grenzen im 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!

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

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

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

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

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Problem3: Varianz des Signals

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

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

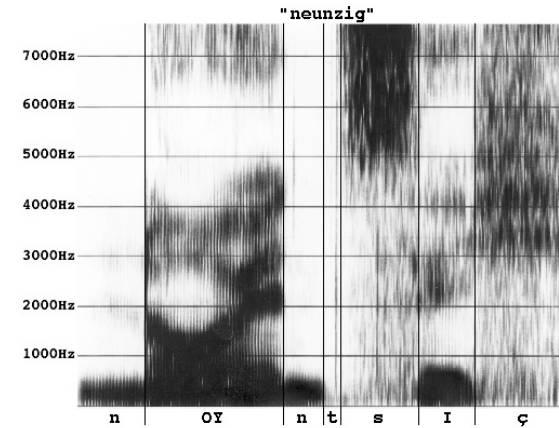
Statistische Modellierung

Spracherkennung ist eine Art Klassifikationsaufgabe:

- Ermittlung der **wahrscheinlichsten** Wortkette
 - $W = w_1 w_2 \dots w_n$, die einem beobachteten akustischen Signal entspricht.
- Die akustische Information, die durch die Lautspektrographie gemessen wird, ist zu differenzieren für statistische Berechnungen.
- Wir erzeugen eine handhabbare Charakterisierung der akustischen Information durch **Merkmals-extraktion**.

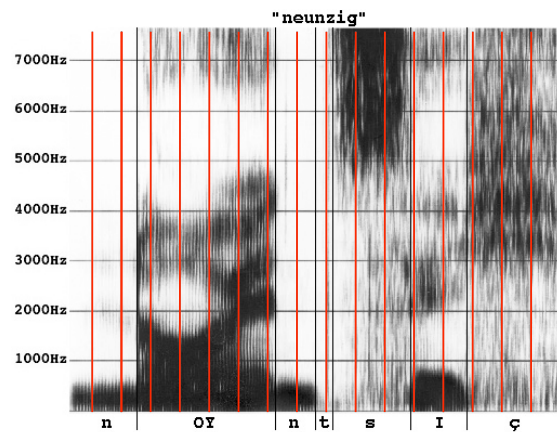
Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Spektrogramm für ein deutsches Wort



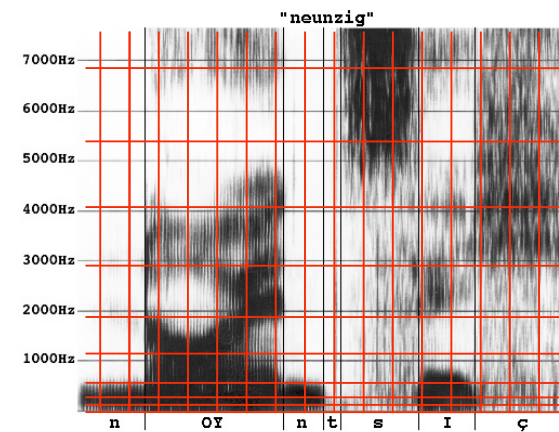
Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Spektrogramm für ein deutsches Wort



Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Spektrogramm für ein deutsches Wort



Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Wie bestimmen wir $P(W|O)$?

- Wie schließen wir vom akustischen Merkmalsmuster O auf die Wahrscheinlichkeit der tatsächlich geäußerten Wortkette W ?
- Akustisches Merkmalsmuster O : Symptom
- Tatsächlich geäußerte Wortkette W : Ursache
- Mit **Bayes-Regel** :

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

- Die wahrscheinlichste Wortkette:

$$\begin{aligned} \max_w P(W | O) &= \max_w \frac{P(O | W) \cdot P(W)}{P(O)} \\ &= \max_w P(O | W) \cdot P(W) \end{aligned}$$

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

Akustische Modelle

$$\max_w P(W | O) = \max_w P(O | W) \cdot P(W)$$

- 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.

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

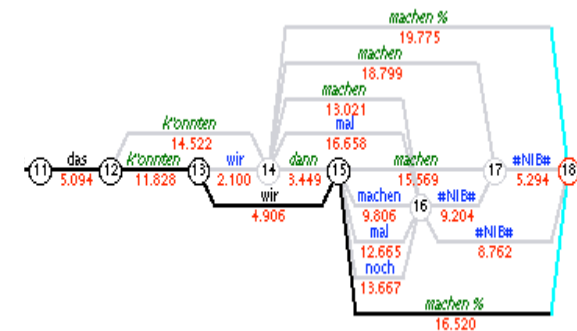
Akustisches Modell und Sprachmodell

$$\max_w P(W | O) = \max_w P(O | W) \cdot P(W)$$

- $P(O|W)$ 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**“

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

Worthypothesengraph



Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

Sprachmodelle

$$\max_w P(W|O) = \max_w P(O|W) \cdot P(W)$$

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

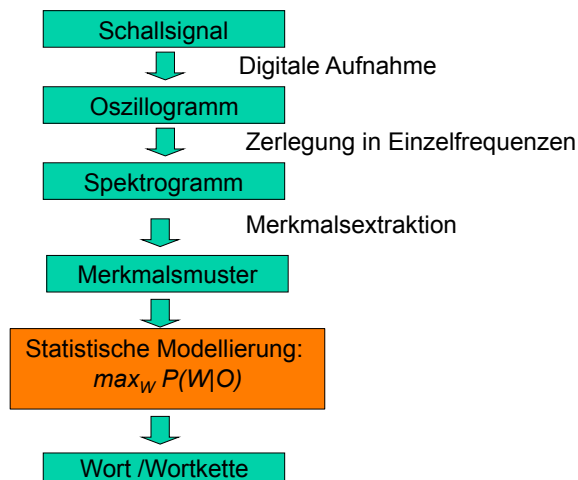
$$P(w_1 w_2 \dots w_n) = P(w_1) * P(w_2|w_1) * P(w_3|w_1 w_2) * \dots * P(w_n|w_1 w_2 \dots w_{n-1})$$

aber:

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

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

Spracherkennung: (Vereinfachtes) Schema



Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

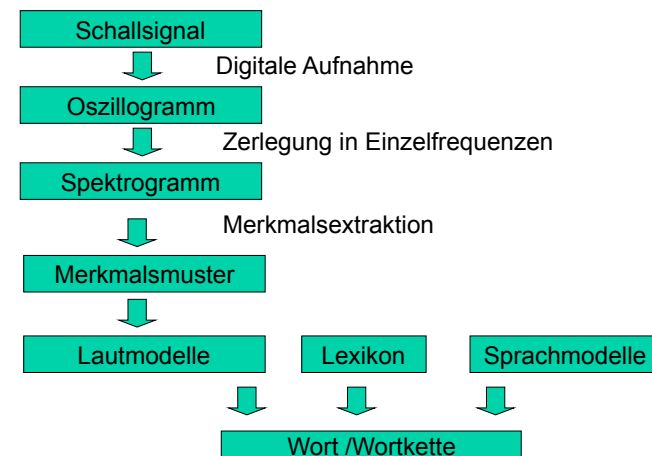
35

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

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

Spracherkennung: Schema



Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UoS Computerlinguistik

36

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 \leq Sprechzeit) und sind in der Wortfehlerrate in einem akzeptablen Bereich .

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik

Erkennungperformanz 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

Vorlesung "Einführung in die CL" 2010/2011 © M. Pinkal UdS Computerlinguistik