

Erkennung und Erzeugung gesprochener Sprache

Sebastian Pado

Einführung in die Computerlinguistik 2004/2005

1

Einführung

- Viele computerlinguistische Anwendungen gehen von **textbasierter** Ein/Ausgabe aus
 - (l.a.) einfacher
- Nicht überall möglich oder sinnvoll
 - Beispiele?
- Übersetzungen
 - Gesprochen » geschrieben: Spracherkennung
 - Geschrieben » gesprochen: Sprachsynthese

Einführung in die Computerlinguistik 2004/2005

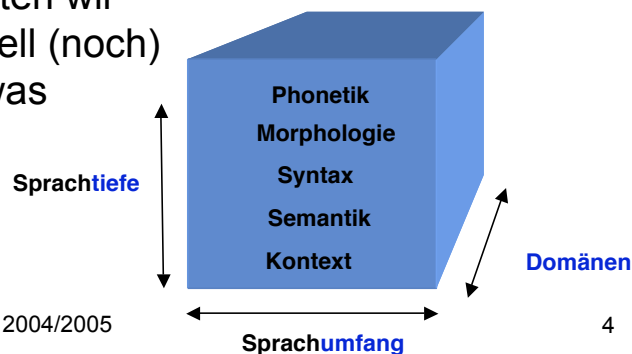
2

Spracherkennung und -synthese: Wozu?

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

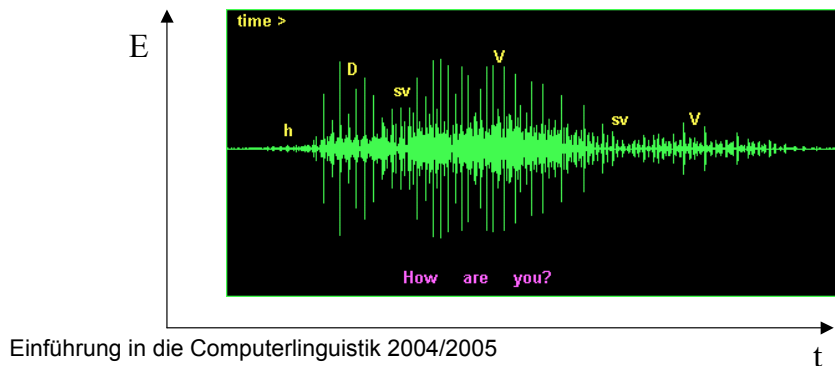
Spracherkennung

- Die Grundaufgabe der Spracherkennung: Gegeben ist ein kontinuierliches Schallsignal. Was wurde vom Sprecher geäußert?
- Ausgabe typischerweise: Kette von Worten
- NB. Eigentlich dürften wir laut Schichtenmodell (noch) gar nicht wissen, was Wörter sind



Oszillogramme

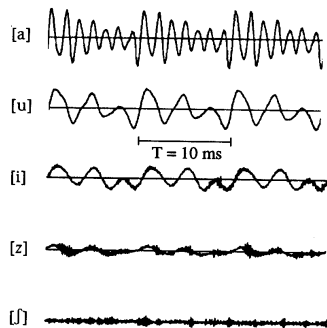
- Direkte Repräsentation des akustischen Signals
 - Zeit-Energie-Diagramm
- Beispiel: “How are you?”



Einführung in die Computerlinguistik 2004/2005

5

Einzelne Laute als Oszillogramme

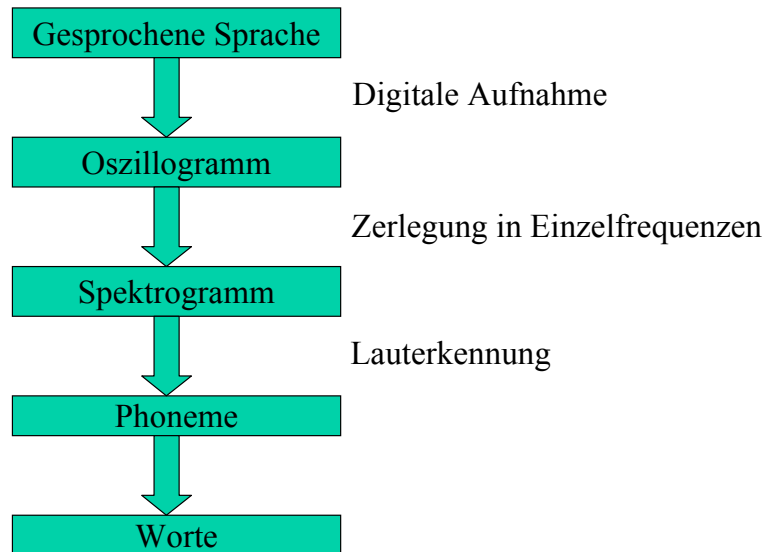


- Laute charakterisiert durch Kombination von Schwingungen verschiedener Frequenzen
- Im Oszillogramm schwer erkennbar (Überlagerung)
- Daher: Geschicktere Repräsentation durch Komponentenanalyse (Fourier-Transformation)
- Ergebnis: Zeit-Frequenz-Diagramm (Spektrogramm)
- (Notation: [Phon] /Phonem/)

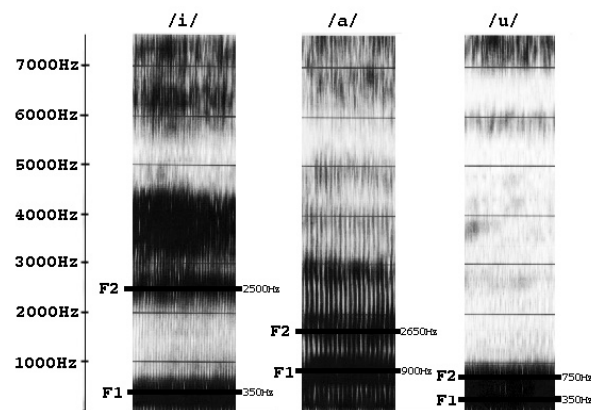
Einführung in die Computerlinguistik 2004/2005

6

Spracherkennung

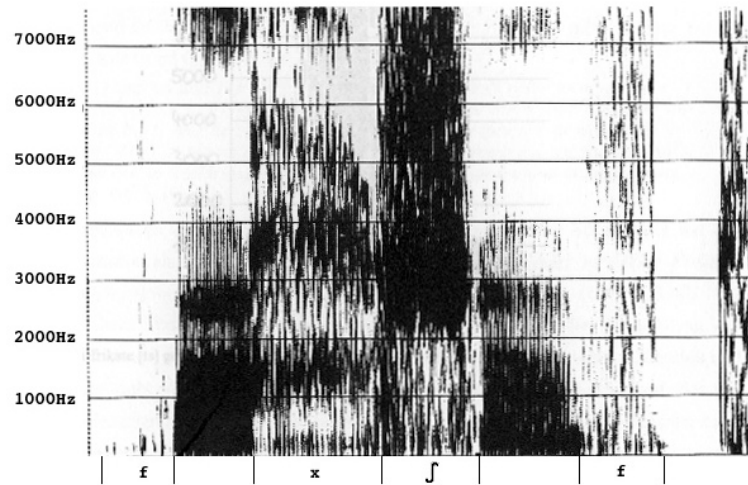


Spektrogramm für die Vokale i,a,u



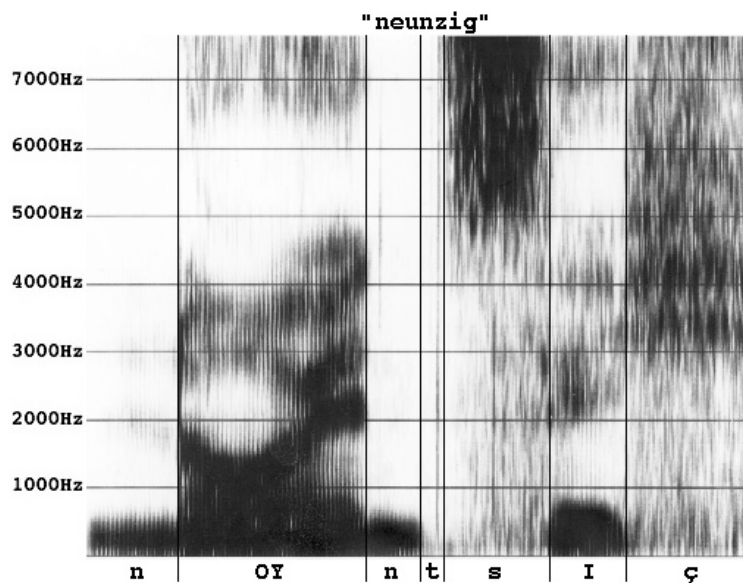
- Dunkle Färbung: große Schallenergie in einem bestimmten Frequenzbereich.
- Die **Formanten** (Obertöne) F1 und F2 sind für die charakteristische Vokalqualität verantwortlich sind.
- Der Verlauf des **Basisformanten** F0 (hier nicht sichtbar) gibt die Intonation der Äußerung wieder.

Spektrogramm für einige Konsonanten

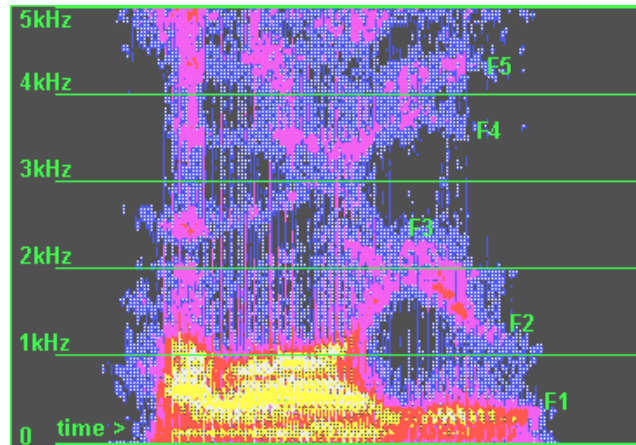


Frikative: f und ch-Laut („ach“-Laut);
Sibillant: „sch“-Laut

Spektrogramm für ein deutsches Wort



Ein buntes Spektrogramm



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

Naives Modell zur Spracherkennung

- 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]

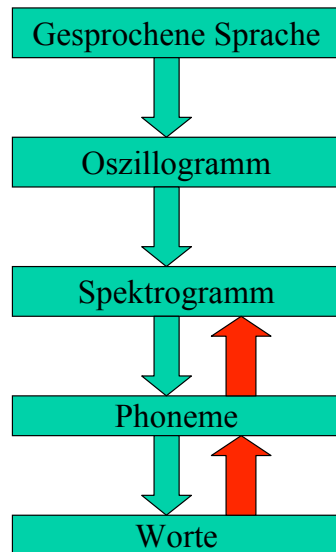
Beispiel: Dialekte

The White House today claimed a decisive victory in the 2004 US presidential election, even though the Democrats have yet to throw in the towel in the bitterly fought contest. (The Guardian, 03.11.03)

- English 
- Scottish 
- American 
- Hispanic 

Quelle: Rhetorical Systems Demo Speech Synthesis

Spracherkennung: 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. [j] in Straße

Wissen in der Spracherkennung Option 1: regelbasiert

- Idee: Bestimme Laute und wende Regeln an
 - Wenn Lautkette mit [çt] oder [jt] 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

Wissen in der Sprachverarbeitung

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 Wahrscheinlichkeitswert 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 \leq 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

Prosodie

- Das Sprachsignal enthält zusätzlich zur „segmentalen Struktur“ (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
- Automatische Erkennung / Verarbeitung von Prosodie schwierig

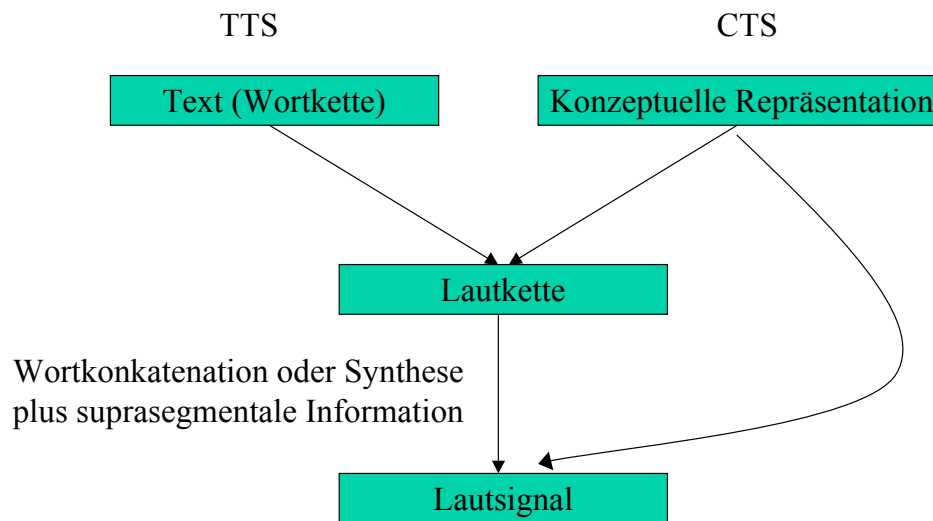
Informationsstruktur

- **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.
- Hauptunterscheidungen:
 - Grundlage der Generierung:
 - Text-to-speech (TTS): Lese Text vor
 - Concept-to-speech (CTS): Verbalisiere Sachverhalt (z.B. Datenbank)
 - Art der Generierung:
 - Wortkonkatenation
 - Synthese

Sprachsynthese



Probleme der Sprachsynthese

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

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)
 - Fränki Kino
[09131 - 6161116](tel:09131-6161116)
 - IBM Staumelder
[06221 593129](tel:06221-593129)