Einführung in die Computerlinguistik

3: Erkennung und Erzeugung gesprochener Sprache

WS 2006/2007

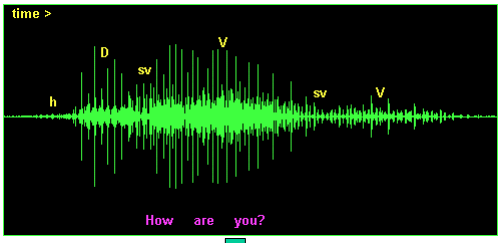
Manfred Pinkal

Motivation

- Spracherkennung:
  - Diktiersysteme, Steuerungssyssteme
- Sprachsynthese:
  - Vorlesesysteme, Ansagen mit variablem Inhalt
- Spracherkennung + Sprachsynthese:
  - Dialogsysteme

Grundaufgabe der Spracherkennung

- Gegeben: Kontinuierliches Schallsignal (Mikrophon)



"How are you"

- Gesucht: Wortkette (die tatsächlich geäußerte Wortfolge)

Spracherkennung: (Vereinfachtes) Schema

Schallsignal

Digitale Aufnahme

Oszillogramm

Laut/Phonem

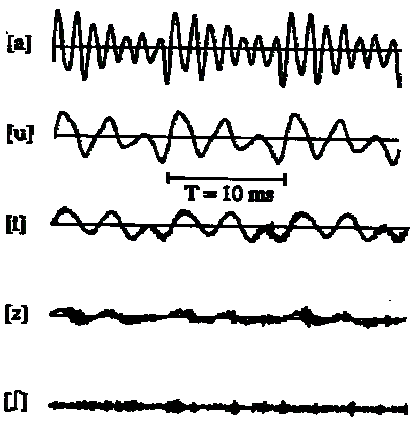
Wort-und Satzerkennung

Wort /Wortkette

Laut und Wort(-kette)

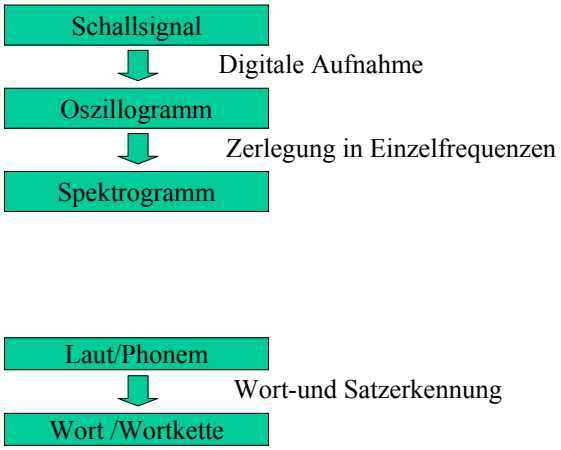
- Phonetik und Schreibung (Orthographie)
- Aussprachevarianten
- Homophone
  - Beispiele: *Rad – Rat , Lid – Lied*
- Wortgrenzen: Wörter sind nur in der Orthografie sauber getrennt.
  - In der gesprochenen Sprache gibt es zwischen Wörtern meistens keine Pause
  - Pausen kommen in spontaner Sprache auch innerhalb von Wörtern vor

Einzelne Laute als Oszillogramme

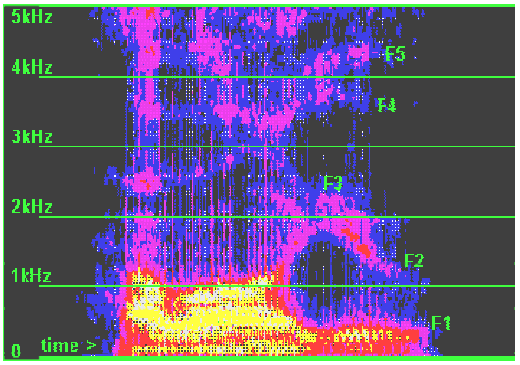


- Laute charakterisiert durch Kombination von Schwingungen verschiedener Frequenzen
- Im Oszillogramm **schwer erkennbar** (Überlagerung)
- Daher: Geschicktere Repräsentation durch Komponentenanalyse (Fourier-Transformation)
- Ergebnis: Zeit-Frequenz-Diagramm (**Spektrogramm**)

Spracherkennung: (Vereinfachtes) Schema

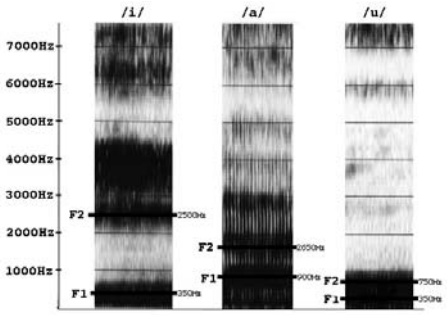


Ein buntes Spektrogramm



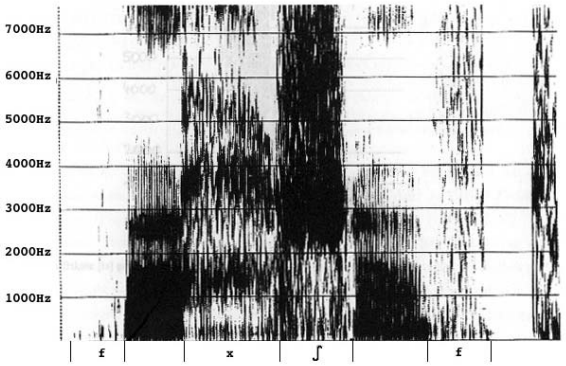
... für den englischen Satz „How are you?“

Spektrogramm für die Vokale i,a,u



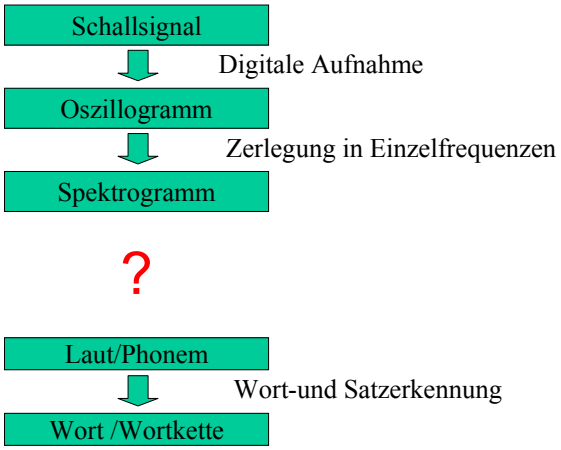
• Dunkle Färbung: große Schallenergie in einem bestimmten Frequenzbereich.
• Die Formanten (Obertöne) F1 und F2 sind für die charakteristische Vokalqualität verantwortlich sind.
• Der Verlauf des Basisformanten F0 (hier nicht sichtbar) gibt die Intonation der Äußerung wieder.

Spektrogramm für einige Konsonanten



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

Spracherkennung: (Vereinfachtes) Schema

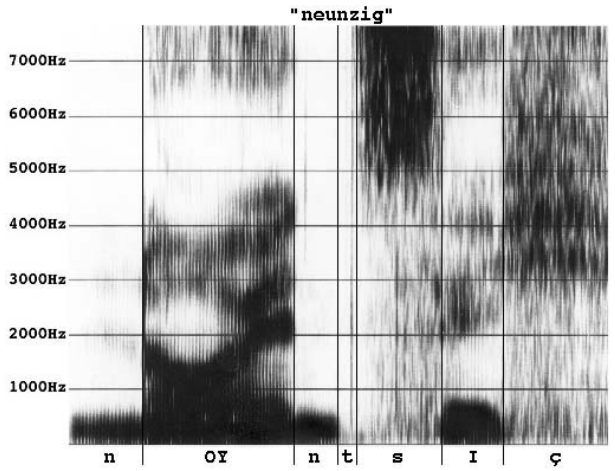


Naives Modell zur Lauterkennung

- 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!

Spektrogramm für ein deutsches Wort



Problem1: Kontinuität des Signals

- 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

Problem: Varianz der Realisierung

Gleicher Laut wird nicht immer gleich ausgesprochen

- Verschiedene Dialekte
- Verschiedene Sprecher
- Unterschiedliche Sprechgeschwindigkeit
- Physischer und emotionaler Zustand des Sprechers

Problem: Varianz des Signals

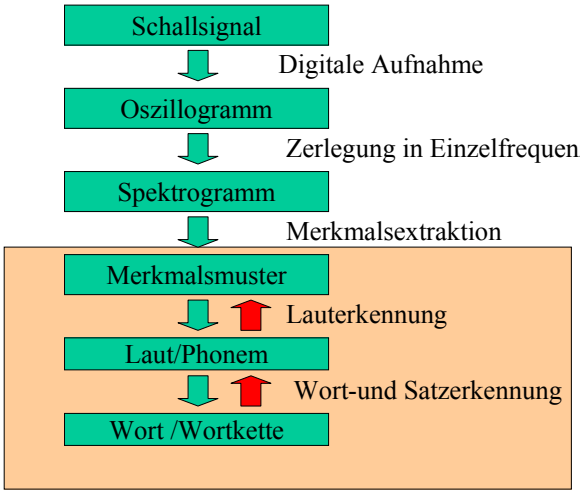
Sprachexterne Einflüsse verändern das Signal

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

Statistische Lauterkennung

- Idee: Regelmäßigkeiten werden nicht explizit repräsentiert, sondern automatisch aus Datenquelle gelernt
- Merkmalsextraktion:
  - Daten in sehr kurze Zeitscheiben (zB 50 ms) einteilen
  - Intensität von ca. 25 Frequenzbändern bestimmen (Formanten!): Merkmalsvektor
- Annotierte Spektrogramme
  - Phonetiker ergänzen eine grosse Anzahl Spektrogramme (wortweise) mit phonetischer Information
- Maschinelles Lernen:
  - 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: Folge von Merkmalsvektoren
  - Ausgabe: wahrscheinlichste Phonemfolge

Spracherkennung: (Vereinfachtes) Schema

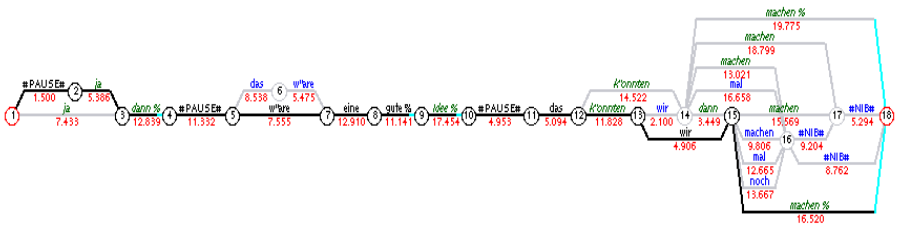


Kombinierte statistische Modelle für die Spracherkennung

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.
- 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 Steuerungssysteme: 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

Sprachsynthese

- Sprachsynthese erscheint einfacher als Spracherkennung. Trotzdem gibt es bisher keine Vollsynthese-Systeme mit einer Qualität, die an menschliche Sprecher heranreicht.

- Ansätze:
  - Einzellautsynthese (unmöglich)
  - Voraufgenommene Sprache (unpraktisch)
  - Diphon-Synthese
  - Wortkonkatenation
  - Unit Selection

Sprachsynthese, Beispiele

Prosodie

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