

Das LIFT-System

- Die Analysekomponente des LIFT-System arbeitet im wesentlichen mit Schlüsselwort-Erkennung (keyword spotting) (und ist darin dem Pattern Matching von ELIZA ähnlich):

Ich möchte in den vierten Stock.

Fahr mich in die vierte Etage.

In den vierten Stock bitte.

Vierter Stock!

Vier!

Wärest Du so nett, mich in den vierten Stock zu bringen?

Das LIFT-System [2]

- Die Schlüsselwörter werden allerdings – im Rahmen der Aufzugsminiwelt – interpretiert:
Ich möchte in den vierten Stock/ zu Professor Barry/ zum Seminarraum/ in die Phonetik.
- und die relevante Information wird in sprachliche oder nicht-sprachliche Aktionen umgesetzt (der Dialog ist wissensbasiert, wie in SHRDLU, allerdings in stark eingeschränktem Sinn):
 - Fahrkommandos
 - Erläuterungen: „*Sie sind bereits im vierten Stock*“
„*Das Büro von Professor Crocker ist im Gebäude 17.1. Soll ich Sie ins Erdgeschoss fahren?*“
 - Rückfragen: „*Zur Phonetik bitte.*“ – „*Möchten Sie in den vierten oder in den fünften Stock?*“

Das LIFT-System [3]

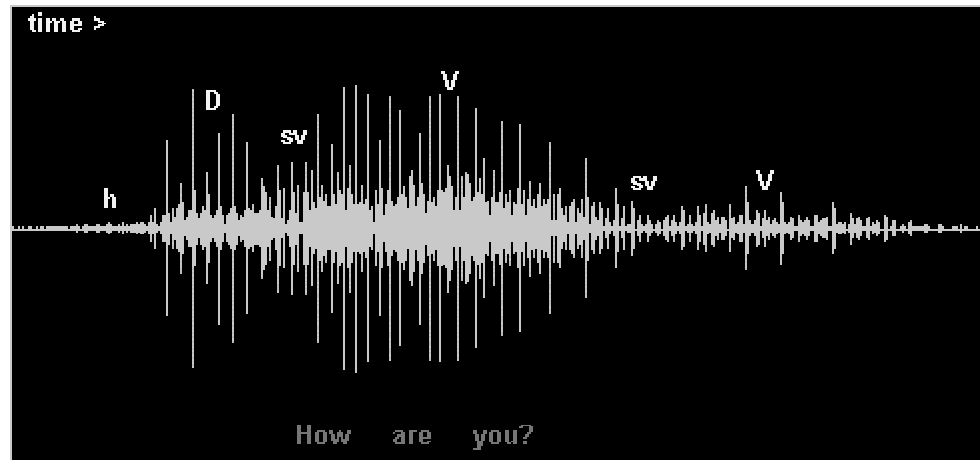
- LIFT-Dialoge sind strukturiert – (wie SHRDLU, anders als ELIZA). Die Dialogstruktur ist in Form von sog. „endlichen Automaten“ kodiert.
- LIFT verwendet gesprochene Eingabe und Spracherkennung, die unzuverlässig ist. Deshalb:
 - Nachfragen: „Ich habe Sie nicht verstanden. Bitte wiederholen Sie Ihre Eingabe.“
 - Verständigungssicherung und Möglichkeit der Benutzerkorrektur: „Ich fahre Sie in den vierten Stock.“
 - Fallback: „Bitte drücken Sie die Knöpfe“
- Aktuelle kommerzielle Dialoganwendungen kombinieren Pattern Matching und selektive wissensbasierte Verarbeitung.

Spracherkennung

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

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: Varianz des Signals

- Varianz der Realisierung von Lauten und Wörtern:
 - Verschiedene Dialekte
 - Verschiedene Sprecher
 - Unterschiedliche Sprechgeschwindigkeit
 - Physischer und emotionaler Zustand des Sprechers
 - Kontext, in dem ein Laut/Wort auftritt
- Varianz der Signalqualität
 - Raumakustik, Entfernung
 - Medium: Face-to-Face, Telefon, Handy
 - Mikrophon-Qualität und -Charakteristik
 - Störgeräusche („Rauschen“, „Noise“)

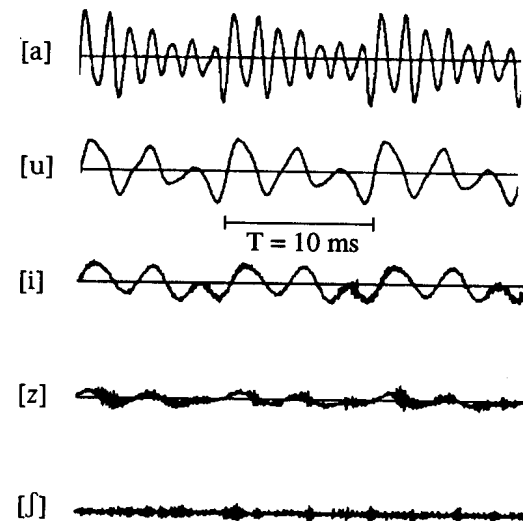
Spracherkennung: Kontinuität des Signals

- Die Laute eines Wortes lassen sich nicht gegeneinander abgrenzen, sondern gehen ineinander über: Man kann durch einen Schnitt 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. Umgekehrt kommen Pausen in spontaner Sprache auch innerhalb von Wörtern vor.

Spracherkennungstechnik

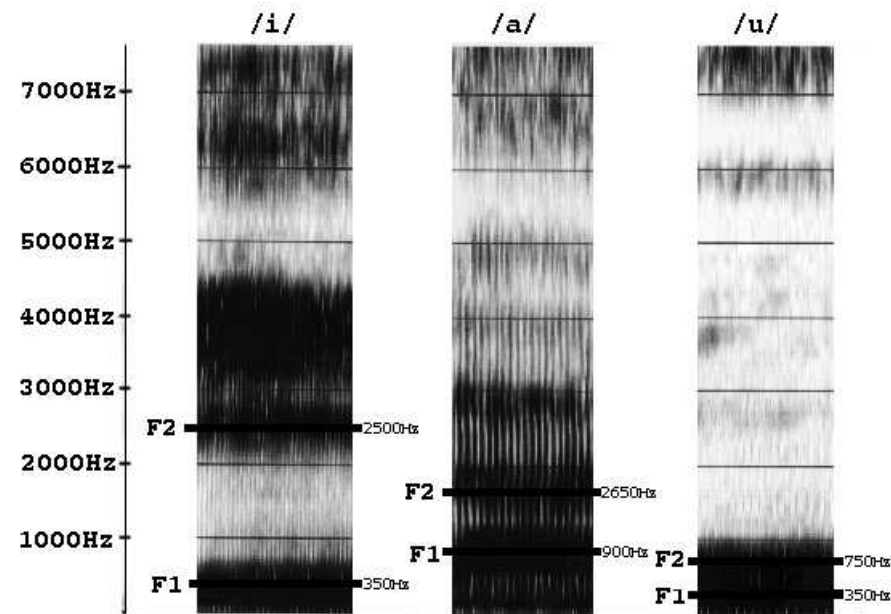
- Digitale Aufnahme des Sprachsignals (visualisiert durch Oszillogramm)
- Zerlegung der komplexen Schwingung in Einzelfrequenzen (Erzeugung des Lautspektrogramms)

Ein paar Oszillogramme



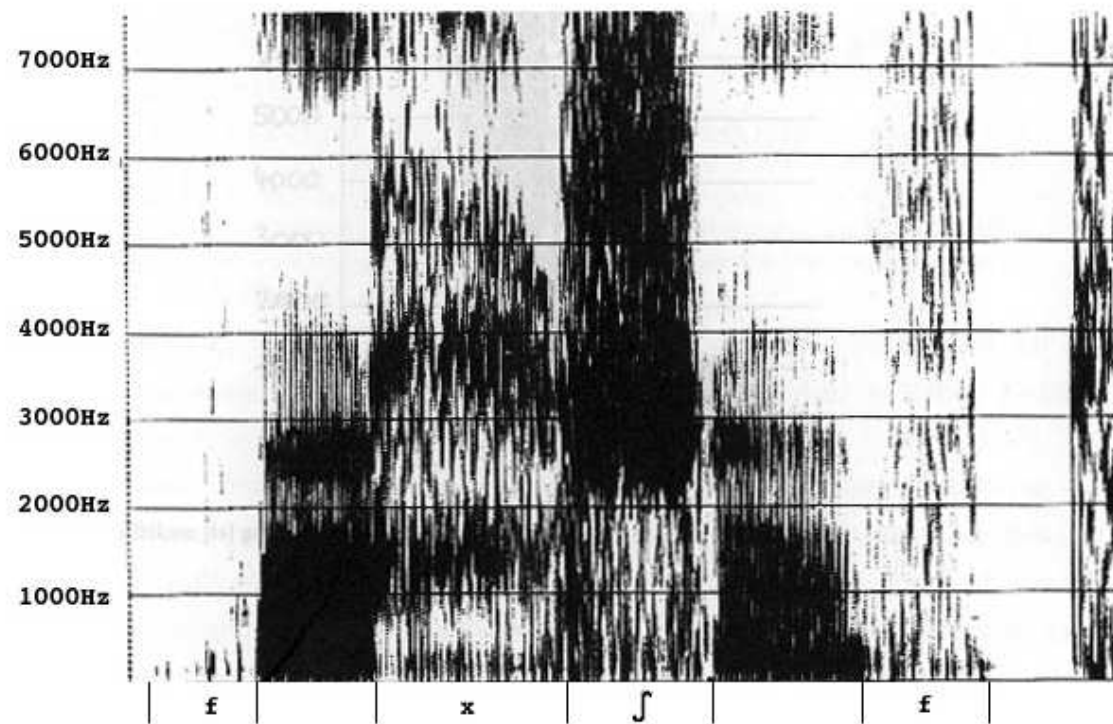
Anmerkung zur Notation: Eckige Klammern werden zur Bezeichnung von Lauten/phonetischen Einheiten verwendet, Schrägstriche zur Bezeichnung von Phonemen (funktionalen Einheiten der Phonologie). Die Notation geht, wie die Beispiele zeigen, manchmal durcheinander. Zum grundsätzlichen Unterschied s. Einführung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

Spektrogramm für die Vokale i,a,u



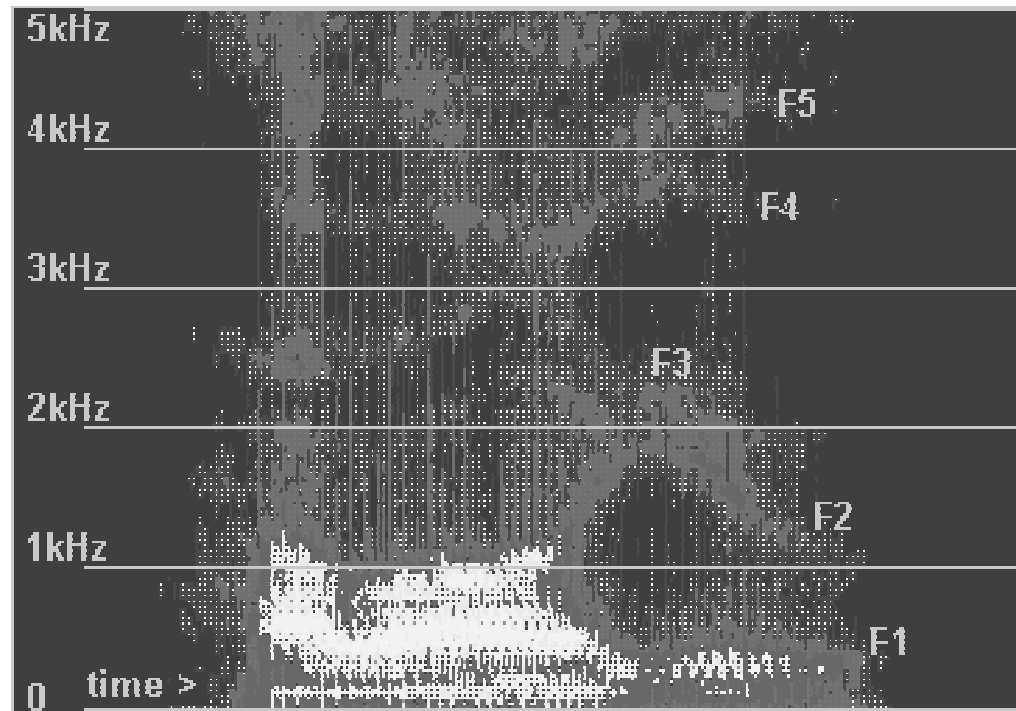
Die x-Achse ist die Zeitachse, auf der y-Achse sind Frequenzen abgetragen. Dunkle Färbung bedeutet große Schallenergie in einem bestimmten Frequenzbereich. Die „Formanten“ F1 und F2 sind die Obertöne, die für die charakteristische Vokalqualität verantwortlich sind. Der Verlauf des Basisformanten F0 (hier 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?“

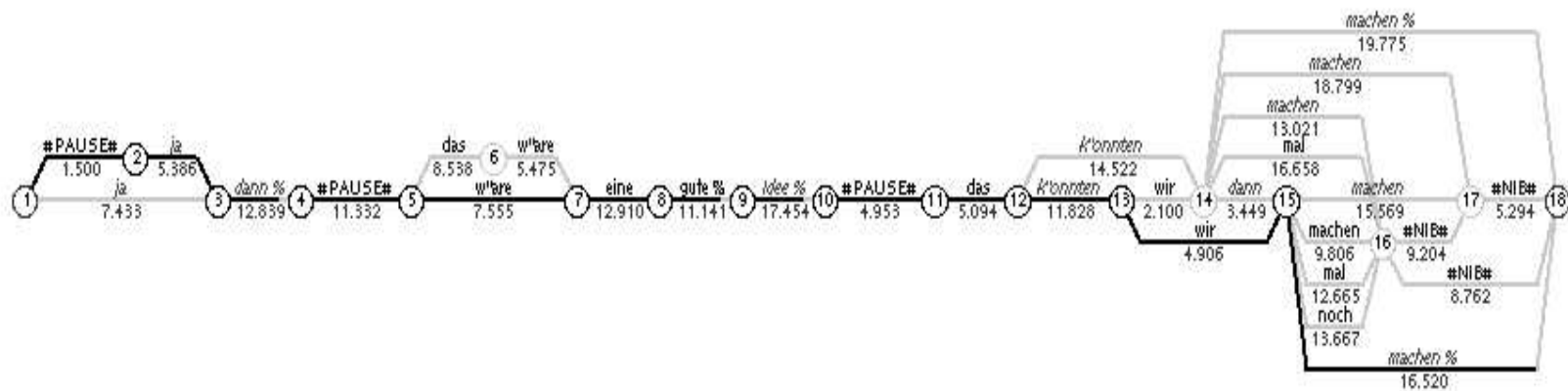
Spracherkennungstechnik

- Digitale Aufnahme des Sprachsignals
- Zerlegung der komplexen Schwingung in Einzelfrequenzen (Erzeugung des Lautspektrogramms)
- Merkmalsextraktion: Bestimmung der Schallenergie in einzelnen Frequenzfenstern (z.B. Viertelton) und Zeitfenstern (z.B. 50 ms).
- Phonetische Transkription/Annotation der Sprachaufnahmen
- Spracherkenner mit statistische Lernverfahren (basierend auf „Hidden-Markov-Modellen“) werden mit Trainingsdaten gefüttert:
 - Aufbereitetes Sprachsignal (Sequenzen von Merkmalsvektoren)
 - Phonetische Transskription
- Output des trainierten Spracherkenners: „Signal A entspricht im Kontext B der Laut/ das Wort C mit Wahrscheinlichkeit w.“

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