Grundlagen der Sprachsignalerkennung

- Spracherkennung als technisches und akademisches Problem
- Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung
- Merkmalsextraktion
- Worterkennung
- Phon-basierte Worterkennung
- Erkennung fließender Sprache
- Systemarchitekturen

Literatur: Grundlagen, Überblick


Klaus Fellbaum.
Sprachverarbeitung und Sprachübertragung, volume 12 of *Nachrichtentechnik*.
Springer, Berlin, 1984.


Alex Waibel and Kai-Fu Lee, editors.
Readings in Speech Recognition.
Morgan Kaufmann, San Mateo CA, 1990.


B. Eppinger and E. Herter.
Sprachverarbeitung.
Hanser, München, 1993.


Daniel Jurafsky and James H. Martin.
Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.
Prentice Hall, Upper Saddle River, NJ, 2000.

Literatur: Grundlagen, Überblick


Manfred R. Schroeder.
Computer speech : recognition, compression, synthesis.
Springer, Berlin, 2004.


Andreas Wendemuth.
Grundlagen der stochastischen Sprachverarbeitung.
Oldenbourg, München, 2004.


Stephen Euler.
Grundkurs Spracherkennung : vom Sprachsignal zum Dialog ; Grundlagen und Anwendung verstehen.
Vieweg, Wiesbaden, 2006.

Literatur: Merkmalsextraktion


Ronal W. Schafer and Lawrence R. Rabiner.
Digital representations of speech signals.
Proceedings of the IEEE, 63(4):662–667, 1975.


Peter Vary, Ulrich Heute, and Wolfgang Hess.
Digitale Sprachsignalverarbeitung.
B. G. Teubner, Stuttgart, 1998.

Literatur: Worterkennung


Sadaoki Furui.
Digital Speech Processing, Synthesis, and Recognition.
Marcel Dekker, New York, 1989.


L. R. Rabiner and B. H. Juang.
An introduction to hidden markov models.
IEEE Acoustics, Speech, and Signal Processing Magazine, 3(1):4–16, 1986.


Harvey F. Silverman and David P. Morgan.
The application of dynamic programming to connected speech recognition.
IEEE Acoustics, Speech, and Signal Processing Magazine, 7(3):6–25, 1990.


Lawrence R. Rabiner.
A tutorial on hidden markov models and selected applications in speech recognition.
Proceedings of the IEEE, 1989.


Joseph Picone.
Continuous speech recognition using hidden markov models.
IEEE Acoustics, Speech, and Signal Processing Magazine, 7(3):26–41, 1990.

Literatur: Worterkennung


X. D. Huang, Y. Ariki, and M. A. Jack.
Hidden Markov Models for Speech Recognition.
Edinburgh University Press, Edinburgh, 1990.


K. F. Lee.
Automatic Speech Recognition: The Development of the Sphinx System. 4th edition.
Kluwer, Boston, 1999.

Literatur: Sprachmodellierung

F. Jelinek.
Self-organized language modeling for speech recognition.
In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann, San Mateo, 1990.

Frederick Jelinek.
Statistical methods for speech recognition.
MIT Press, Cambridge, Mass., 1998.

Eugene Charniak.
Statistical Language Learning.
MIT Press, Cambridge, MA, 1993.

Literatur: Dialogmodellierung

Michael H. Cohen, James P. Giangola, and Jennifer Balogh.
Voice user interface design.
Addison-Wesley, Boston, Mass., 2004.

Deborah Dahl, editor.
Practical spoken dialog systems.
Kluwer Academic Publ., Dordrecht, 2004.

Randy Allen Harris.
Voice interaction design : crafting the new conversational speech systems.
Elsevier/Morgan Kaufmann, Amsterdam, 2005.

Michael F. McTear.
Spoken dialogue technology : toward the conversational user interface.
Springer, London, 2004.

Grundlagen der Sprachsignalerkennung

- Spracherkennung als technisches und akademisches Problem
- Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung
- Merkmalsextraktion
- Worterkennung
- Phon-basierte Worterkennung
- Erkennung fließender Sprache
- Systemarchitekturen

Spracherkennung als technisches und akademisches Problem

- Gesprochene Sprache
- Spracherkennung als technisches Problem
- Parameter und Einschränkungen
- Anwendungsbereiche
- Spracherkennung als akademisches Problem
- Herausforderungen
- Chronik
- Entwicklungsstand und Ausblick

Spracherkennung als technisches und akademisches Problem

- Gesprochene Sprache
- Spracherkennung als technisches Problem
- Parameter und Einschränkungen
- Anwendungsbereiche
- Spracherkennung als akademisches Problem
- Herausforderungen
- Chronik
- Entwicklungsstand und Ausblick

Gesprochene Sprache

- Stellung in der Sprachtheorie
  - gesprochene Sprache ist primär
  - geschriebene Sprache ist sekundär
  - *de facto* jedoch einseitige Dominanz einer idealisierten Sprache (Schriftsprache)
  - starke Auswirkungen auf die (frühen) Arbeiten zur Spracherkennung

- ein simples Modell
  - Buchstabenfolge $\leftrightarrow$ "Phonemfolge"
    $\int$ *pra*:*xɛrkɛnuɳ*
  - Analogie zur Schriftsprache
  - entspricht dem subjektiven Eindruck:
    gesprochene Sprache = Folge von Wörtern = Folge von Lauten
    - "Perlenkette diskreter Elemente"
    - klingen in jedem Kontext gleich

Gesprochene Sprache

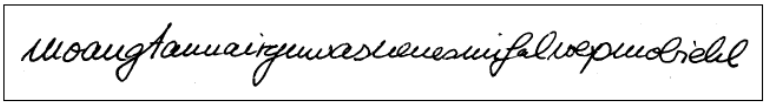
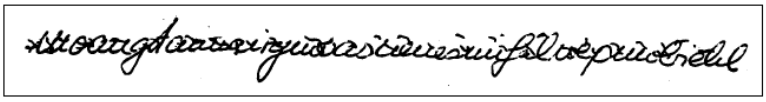
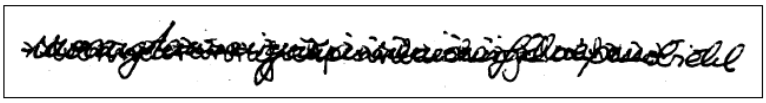
- zu grobe Beschreibung für gesprochene Sprache
- objektive Tatbestände:
  - Sprachsignal ist kontinuierlich, keine Grenzmarkierungen für Laute und Wörter
  - extrem große, kontextabhängige Varianz: z.B. Assimilationen
  - unvollständige Lautrealisierungen (Elisionen, Verschleifungen)
- subjektiver Eindruck ist das Ergebnis einer hochkomplexen internen Verarbeitungsprozedur

Gesprochene Sprache

- Sprachverstehen ist nicht nur ein (passiver) Analyseprozeß
- aktive Rekonstruktion aus einem hochredundanten Code
  - Phonotaktik
  - Morphologie
  - Syntax
  - Semantik
  - Weltwissen
  - Domänenwissen

Gesprochene Sprache

- Warum ist Spracherkennung schwierig?

Guten Morgen, Herr Hauptkomissar Thanner. Gibt es irgendetwas Neues im Fall "Verbmobil"?	der Text in "Schönschrift"
Morgen, Thanner. Irgendwas Neues im Fall "Verbmobil"?	spontan gesprochene Sprache
morgen thanner irgendwas neues im fall verbmobil	Großschreibung? Satzzeichen?
morgenthannerirgendwasneuesimfallverbmobil	kontinuierliche Sprache
moangtannairgnwasneuesimfalwerpmobiehl	Aussprachevarianten
	artikulatorische Verschleifung
	Störungen und Verzerrungen
	Fremdstimmen „Cocktailparty“

Gesprochene Sprache

- Modellerweiterungen sind nötig
  - Artikulation ist ein asynchroner Prozess
    → artikulatorische Varianz
  - Artikulation ist kontextuell beeinflusst
    → Koartikulation, Assimilation
  - suprasegmentale Eigenschaften (Prosodie)
    → Rhythmus, Betonung, Melodie
  - anatomische und psychische Individualitäat
    → Sprechervarianz
- Ökonomieprinzip: Sprecher versucht immer, sein Kommunikationsziel mit minimalem Aufwand zu erreichen
  - hohe Redundanz → geringe Qualität der Artikulation

Spracherkennung als technisches und akademisches Problem

- Gesprochene Sprache
- Spracherkennung als technisches Problem
- Parameter und Einschränkungen
- Anwendungsbereiche
- Spracherkennung als akademisches Problem
- Herausforderungen
- Chronik
- Entwicklungsstand und Ausblick

Spracherkennung als technisches Problem

- Sprachsignal enthält sehr viel Varianz
  - Sprecher
  - Kontext
  - Umgebung
  - . . .
- Spracherkennung ist Suche nach Invarianten
  - Gemeinsamkeiten von Lauten/Silben/Worten
  - Wie kann man die Varianz "herausfiltern"?
  - Spracherkennung = Dekodierung?
- das Sprachsignal ist ein unterbestimmter Kode
  - Verschleifungen, Weglassungen, ...
  - Hörer muss Information ergänzen
  - Spracherkennung ist auch ein aktiver Prozess der Informationsrekonstruktion

Spracherkennung als technisches Problem

- Repräsentationsformen für Sprache
  - analoges Signal
  - digitalisiertes Signal
  - (digitalisierte) parametrische Signalbeschreibung
  - lineare Symbolfolgen
    - phonetisch: Laute, Diakritika
    - graphematisch: Buchstaben, Interpunktion
  - textuell: Morphe, Wortformen, Sätze
  - strukturelle Beschreibungen (für Silben, Wörter, Sätze, Texte)
  - "logische" (semantische) Beschreibung
  - (Re-)aktionsklassen
    - Steuerimpulse
    - Datenbankanfragen
    - Dialogantworten
    - Übersetzungen

Spracherkennung als technisches Problem

Zielstellungen bei der Verarbeitung gesprochener Sprache

- AD-/DA-Wandlung (z.B. Telefon)
  - analog $\longrightarrow$ digital
  - digital $\longrightarrow$ analog
- Sprachkodierung, Sprachkompression, Sprachverschlüsselung
  - digital $\longleftrightarrow$ digital
- Sprach(voll)synthese
  - digital $\longrightarrow$ analog
  - parametrisch $\longrightarrow$ analog
  - phonetisch $\longrightarrow$ analog
  - graphemisch $\longrightarrow$ analog
  - "strukturell" $\longrightarrow$ analog
- Sprecheridentifikation, -verifikation, -evaluation
- Sprachidentifikation

Spracherkennung als technisches Problem

- Spracherkennung (ASE, ASR)
  - analog $\rightarrow$ textuell
- Sprachverstehen (ASV, ASU)
  - analog $\rightarrow$ "logisch"
- Sprachdialog-, Sprachübersetzungs-, Sprachkommandosysteme
- Verarbeitung natürlicher Sprache (NLP)
  - textuell $\rightarrow$ "logisch"
  - textuell $\rightarrow$ strukturell
  - textuell $\rightarrow$ textuell
  - textuell $\rightarrow$???

- ASV = ASE + NLP ???

Spracherkennung als technisches Problem

- historischer Ausgangspunkt: Spracherkennung
  - Nachrichtentechnik
  - Kanalbandbreiten (Transatlantikkabel)
    - digitalisierte Sprache $\approx$ 100 Kbit/s
    - Kanalvocoder $\approx$ 10Kbit/s
    - Formantvocoder $\approx$ 1Kbit/s
    - Lautzeichenübertragung $\approx$ 0,1Kbit/s
  - Mensch-Mensch-Kommunikation
  - Vision: akustische Schreibmaschine

- erste Erfolge: Erkenner für isolierte Laute (Vokale), später auch für isoliert gesprochene Einzelworte
  - Kommandointerpreter
  - "akustische Tastatur"
  - Anwendungen in zahlreichen Gebieten

Spracherkennung als technisches Problem

- erste Wende: Sprachverstehen
  - Reaktion auf Schwierigkeiten
  - menschliches Vorbild
  - Voraussetzung: technische Systeme, die Sprache "verstehen" können
  - akustische Programmierung, z.B. Datenbankabfrage
- zweite Wende: zurück zur Spracherkennung
  - verbesserte Lösungen bei der Signalverarbeitung und Klassifikation
  - unüberschaubare Systeme
  - extrem hoher Erarbeitungsaufwand für hochgradig spezialisierte Lösungen

Spracherkennung als technisches Problem

- Symbiose 1: Sprachdialogsysteme
  - zielorientierter Dialog erfordert Sprachverstehen
  - Beschränkung auf spezielle Einsatzgebiete
- Symbiose 2: automatisches Dolmetschen
  - Übersetzen erfordert wenigstens partielles Verstehen
  - Übersetzen ist aber auch ohne detailliertes Spezialwissen möglich
- Langzeitperspektive: Integration von ASE und NLP

Spracherkennung als technisches und akademisches Problem

- Gesprochene Sprache
- Spracherkennung als technisches Problem
- Parameter und Einschränkungen
- Anwendungsbereiche
- Spracherkennung als akademisches Problem
- Herausforderungen
- Chronik
- Entwicklungsstand und Ausblick

Parameter und Einschränkungen

- akustische Aufnahmebedingungen:
  - Raum:
    reflexionsfrei / Studio / Telefonzelle

    → lineare und nichtlineare Transformationen
  - Mikrofon:
    Kehlkopfmikrofon / Kopfgeschirr / Handmikrofon / Telefon

    → Frequenzgang, Bandbreite
  - Störgeräusche:
    Sprecherkabine / Büro / Fahrerkabine / Maschinenraum / Bahnhofshalle / Cockpit

    → Störabstand
- Nutzungsbedingungen: stationär / mobil

Parameter und Einschränkungen

- Beschränkungen des Nutzerkreises
  - Sprecherzahl:
    ein / mehrere / viele Sprecher, bekannte / unbekannte Sprecher
  - Sprechertyp
    weiblich/männlich, Kind / Erwachsener
    individuelle Besonderheiten: Anatomie, Dialekt, Sprachfehler, ...
  - Sprecherdisziplin (teilweise sehr hohe Anforderungen)
    Wortpausen: Einzelworte / fließende Sprache
    Artikulation: Lesesprache / Diktiersprache / Spontansprache
    Sprachumfang: Lexik / Syntax
  - Sprecherdisposition:
    Stress / Ermüdung / Konzentration / Gesundheitszustand
  - Kooperativität
  - Vertrautheitsgrad

Parameter und Einschränkungen

- Anwendungsgebiete
  - Beispiele

    Versandbestellung, home banking
    Bahnauskunft, Flugreservierung
    Terminabsprache
    Patentrecherche, ärztliche Diagnose
    Telefonseelsorge
  - Unterscheidungskriterien

    Wortschatz / Grammatik
    Semantik (atomar / kompositionell)

Parameter und Einschränkungen

- Beispiele für Wortschatzgrößen (Vollformen)

  grunt detector: 1
  ja/nein-Erkenner: 2
  Ziffern/Zahlen: 10 + x
  Kommandos: 20 - 200
  Auskünfte: 500 - 2000
  Alltagssprache: 8000 - 20000
  Diktiermaschine: 20000 - 50000
  Deutsch (ohne Fremdwörter) $\approx$ 300000

Parameter und Einschränkungen

- Dialoggestaltung
  - geführter Dialog / verteilte Initiative
- Systemmodifikation, -adaptation
  - Lernen / Nachführen / Neulernen / Umprogrammieren

  für
  - neue Sprecher / neue Wörter / neue Anwendungen
- Extrem viele Freiheitsgrade beim Systementwurf
  - Universelle Lösung des "Spracherkennungsproblems" ist nicht zu erwarten!
    $\rightarrow$ Spektrum speziell angepasster Systemvarianten

Spracherkennung als technisches und akademisches Problem

- Gesprochene Sprache
- Spracherkennung als technisches Problem
- Parameter und Einschränkungen
- Anwendungsbereiche
- Spracherkennung als akademisches Problem
- Herausforderungen
- Chronik
- Entwicklungsstand und Ausblick

Anwendungsbereiche

Verglichen mit anderen Eingabekanälen ist Spracherkennung *immer* mit hoher Unsicherheit verbunden.

Wozu dann eigentlich?

Anwendungsbereiche

- Anwendungsdruck ist augenscheinlich sehr hoch:
  - natürlicher Kanal der Mensch-Mensch-Kommunikation
  - unterstützt kollektives Problemlösen
  - erfordert im Idealfall wenig Übung,
    von Natur aus mnemonisch (Klartext, keine Kürzel)
  - verhältnismäßig schneller Eingabekanal
    - Tastatureingabe: 10-20 ... 100-150 W/min
      (untrainiert/trainiert)
    - Handschrifteingabe: 25 W/min mit/ohne Übung
    - Sprechen 120-250 W/min mit/ohne Übung
      Diktiermaschine 40 W/min
  - wichtigster Kanal in multimedialen Kommunikationssituationen

Anwendungsbereiche

- hoher Anwendungsdruck (Forts.)
  - alternative Interaktionskanäle möglicherweise blockiert
    Diktieren am Operationsmikroskop
    unkritische Bedienfunktionen im Auto
    Cockpitassistenz
    Behinderungen
  - alternative Kanäle möglicherweise unzumutbar
    - naive Nutzer in öffentlichen Informationsdiensten
      aber problematisch: unkannte Sprecher, Sprecherdisziplin
    - ubiquitous / wearable computing
  - Bewegungsfreiheit
  - geringer Raumbedarf (nur Mikrofon)
  - lichtunabhängig
  - öffentlich, separat protokollierbar

Anwendungsbereiche

- aber: Vorteile oft nur unter idealisierten Bedingungen
  - meist preiswerte technische Alternativen verfügbar
    - Tastatur
    - grafische Bedienoberflächen
    - direkte Datenerfassung
  - (idealisierter) Geschwindigkeitsvorteil oft nicht so entscheidend (z.B. Diktiergerät), wohl aber die schnelle Verfügbarkeit der eingegebenen Texte
  - Medienkonvergenz
    - Telefon wird zum Multimediaterminal
    - Telefonie über den Arbeitsplatzrechner
  - direkte Manipulation oft einfacher als die Formulierung sprachlicher Befehle (z.B. Operationsmikroskop)
    - Joystick, Maus, Touchscreen, Pedale, ...
  - leicht abhörbar

Anwendungsbereiche

- Unsicherheit kann durch Redundanz kompensiert werden
  - Eingabewiederholung / Eingabebestätigung
  - erfordern aber Zeit und Geduld

- bisher realisierte Anwendungen
  - Maschinensteuerung (auch Software)
  - Auskunftssysteme
  - Diktiersysteme
- Prototyplösungen
  - Dolmetschen für spezielle Anwendungen
  - Multimediaretrieval
  - Gesprächsprotokollierung

Spracherkennung als technisches und akademisches Problem

- Gesprochene Sprache
- Spracherkennung als technisches Problem
- Parameter und Einschränkungen
- Anwendungsbereiche
- Spracherkennung als akademisches Problem
- Herausforderungen
- Chronik
- Entwicklungsstand und Ausblick

Spracherkennung als akademisches Problem

- Ziel: tieferes Verständnis für humane Sprachperzeption
  - Validierung perzeptionspsychologischer und psycholinguistischer Theorien
  - Entwicklung von Problembewusstsein
    → Aufgabenstellungen für weitere psychologische Untersuchungen

- humane Spracherkennung als Vorbild
  - nicht unbedingt:
    technische Spracherkennung = humane Sprachperzeption
    "Flugzeuge schlagen auch nicht mit den Flügeln"
  - aber: Anregungen
    z.B. Nutzung gehörbezogener Parameter
  - Motor für die weitere Entwicklung der Forschung
    Aufzeigen der Lücke zwischen technischer Realität und humanem Vorbild

Spracherkennung als akademisches Problem

- teilweise heftige Kontroverse
  - "Each time I fired a phonetician - recognition rate increased by 3%".
  - "With friends like statistics - who needs linguistics."
  - "Die KI in der Spracherkennung hat uns zehn Jahre gekostet."

Spracherkennung als technisches und akademisches Problem

- Gesprochene Sprache
- Spracherkennung als technisches Problem
- Parameter und Einschränkungen
- Anwendungsbereiche
- Spracherkennung als akademisches Problem
- Herausforderungen
- Chronik
- Entwicklungsstand und Ausblick

Herausforderungen

- Kontinuität
  - keine Diskontinuitäten im Signal (Laut, Silben, Wort, Phrasengrenzen)
- Variabilität
  - Anatomie / Dialekt / sprachlicher Kontext / Sprechtempo / Emotion / Aufnahmebedingungen
- Komplexität

12	distinktive Merkmale
40 - 50	Phoneme
100 - 200	Allophone
1000 - 2000	Diphone
1000 - 2000	Halbsilben
$\approx$ 10000	Silben
100000	Grundformen (Komposita?)
potentiell unendlich	Sätze

Herausforderungen

- Ambiguitäten auf verschiedenen Ebenen
  - Homophonie:
    to/too, Rad/Rat, Beeren/Bären, ...
  - partielle Ähnlichkeiten (Überschneidungen):
    Fahrt/fort
  - Wortgrenzen:
    grey tape/great ape, Wach-traum/Wacht-raum, Drucker-zeugnis/Druck-erzeugnis
  - Unterordnung in der Satzstruktur:
    Der Mann sah das Mädchen mit dem Fernglas.
  - Referenz in der logischen Struktur:
    das Tonband, das Nixon vernichtete
  - phrasale Gliederung:
    Der gute Mann denkt an sich [,] selbst zuletzt.
    Der Professor [,] sagt [,] der Student sei unterqualifiziert.

Herausforderungen

- lexikalische Bedeutung wird erst durch den Kontext endgültig determiniert
  - *Bienenhonig/Imkerhonig*
  - *Schweineschnitzel/Jägerschnitzel*
- Bedeutung ist nicht statisch und kann im Dialog ausgehandelt werden

Spracherkennung als technisches und akademisches Problem

- Gesprochene Sprache
- Spracherkennung als technisches Problem
- Parameter und Einschränkungen
- Anwendungsbereiche
- Spracherkennung als akademisches Problem
- Herausforderungen
- Chronik
- Entwicklungsstand und Ausblick

Chronik

1844	Telegraph
1858	transatlantischer Telegraph
1876	Telefon
1927	transatlantische Funkverbindung
1939	Kanalvocoder
1946	Klangspektrograph: visible speech
1948	akustische Theorie der Sprachproduktion
1951	Formantsynthese für Einzellaute
1956	transatlantisches Telefon
1958	kommerzielle Digitalrechner
1962	PCM, Satellitenkommunikation
1963	großer Optimismus Morris Halle: ein Problem - zwei Jungs - drei Jahre $\rightarrow$ Detektoren für zwölf Merkmale
1965	Fast Fourier Transformation (FFT)

Chronik

1966	digitale Filterung
1968	linear predictive coding (LPC)
1970	erste Arbeitsplatzrechner
1970	erstes DARPA-Projekt: Speech Understanding Research
1971	Worterkennung mit Schablonenvergleich (DTW)
1972	Sprecherverifikation
1973	Dennis Klatt: Hören oder Lesen
1975	Hidden-Markov-Modelle (HMM)
1976	Bruce Lowerre: HARPY (HMM mit 15000 Zuständen)
1976	Text-to-Speech-Synthese
1980	Mel-Cepstrum, Vektorquantisierung
1981	Signalprozessoren (DSP)
1985	kontextabhängige Phonmodelle
1985	Backpropagation für KNN

Chronik

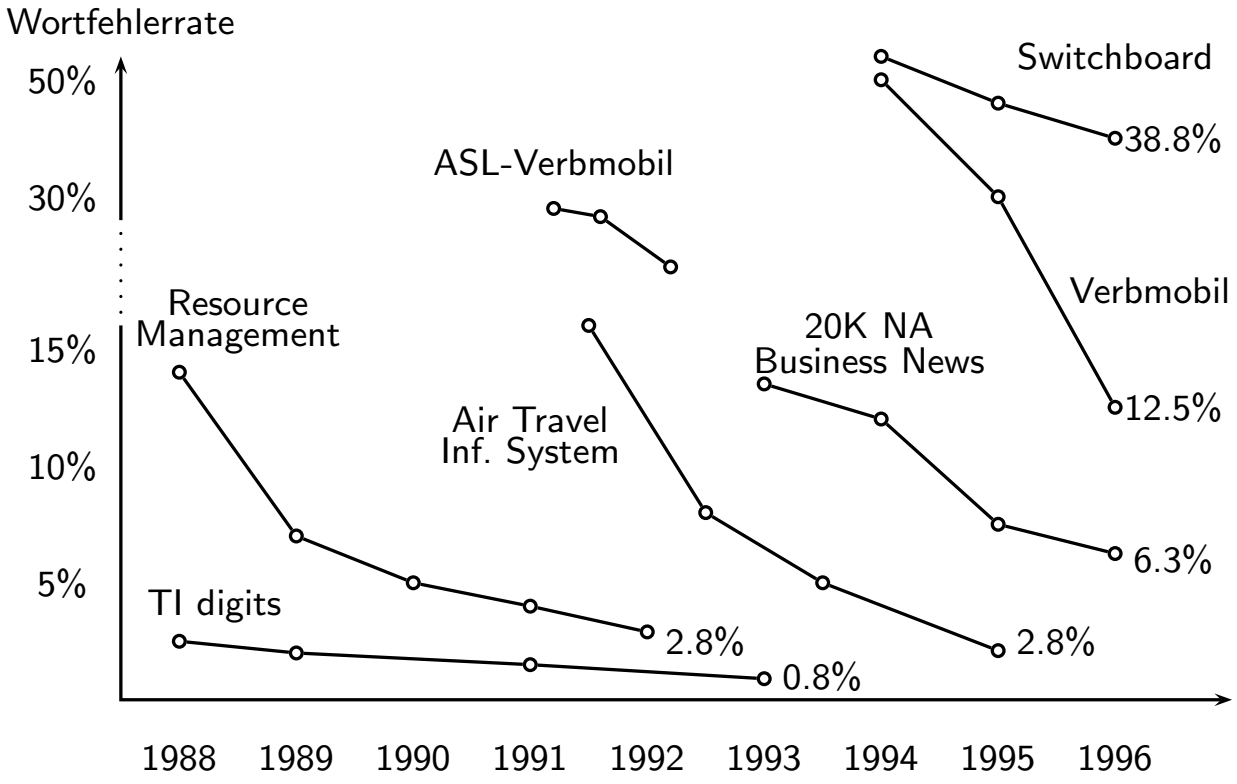
1989	Sprecherunabhängige kontinuierliche Spracherkennung mit 1000 Wörtern
1993	Diktiergerät für kontinuierliche Sprache mit 25000 Wörtern
1993	Kommandoerkennung in Apples System 7
1995	Sprecheradaption mit linearer Regression
1996	Erkennung von Spontansprache
1996	Sprecheradaption durch Vokaltraktnormalisierung
2000	Audio-visuelle Erkennung für großen Wortschatz
2000	Prototypen für das Maschinelle Dolmetschen
2000	XML-Standard für Dialogsysteme (Draft)
2003	Telefoniesysteme im breiten Einsatz
2004	XML-Standard für Dialogsysteme (Recommendation)
2007	Spracherkennung in Windows Vista (einschließlich Diktat)

Spracherkennung als technisches und akademisches Problem

- Gesprochene Sprache
- Spracherkennung als technisches Problem
- Parameter und Einschränkungen
- Anwendungsbereiche
- Spracherkennung als akademisches Problem
- Herausforderungen
- Chronik
- Entwicklungsstand und Ausblick

Entwicklungsstand und Ausblick

- Aufgaben mit steigendem Schwierigkeitsgrad



Entwicklungsstand und Ausblick

- Steigender Bedarf an Trainingsdaten

Gelesene Sprache (CSR)

Zeit-raum	Korpus	Trainings-daten		Voka-bular	Per-plexi-tät	Wort-fehler-rate %
		h	Wort-formen	Wort-formen		
87-92	Ressource Management (RM)	4	-	1k	60	4
92-94	Wall Street Journal (WSJ)	12	38M	5k	50	5
93-94	Wall Street Journal (WSJ)	66	120M	20k	150	10
94-95	North Am. Business News (NAB)	66	400M	65k	150	7
95-	Broadcast News (BN)	100	130M	65k	200	20

Young & Chase 1998

Entwicklungsstand und Ausblick

- Steigender Bedarf an Trainingsdaten

Spontansprache (LVCSR)

Zeit-raum	Korpus	Trainings-daten		Voka-bular	Per-plexi-tät	Wort-fehler-rate %
		h	Wort-formen	Wort-formen		
93-95	Switchboard (SWB)	60	2.2M	10k	120	45
96-97	Switchboard (SWB)	120	3.1M	20k	80	30
97-	Switchboard-2 (SWB-2)	140	3.1M	20k	75	35
96-	Call Home English (CH Eng)	20	200k	25k	120	40
95-	Call Home Spanish (CH Span)	18	200k	23k	120	55
96-	Call Home Arabish (CH Arab)	18	300k	16k	–	60
95-	Call Home Mandarin (CH Mand)	16	160k	20k	175	55
96-	Call Home German (CH Germ)	16	170k	20k	120	60

Young & Chase 1998

Entwicklungsstand und Ausblick

- Wachstum des Wortschatzes

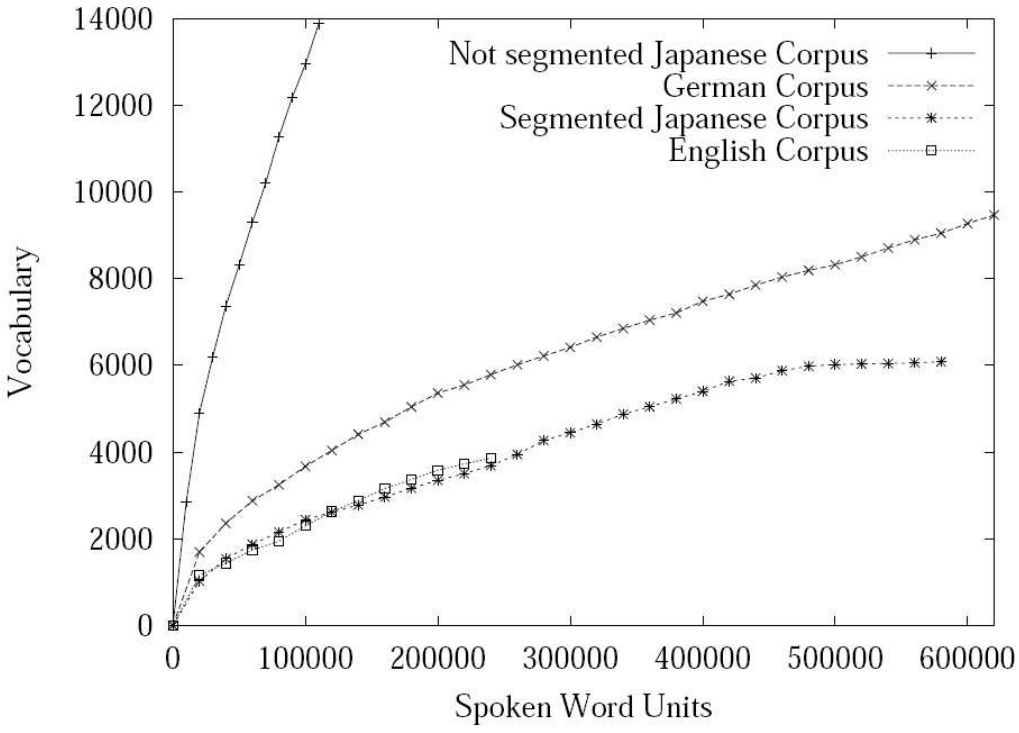
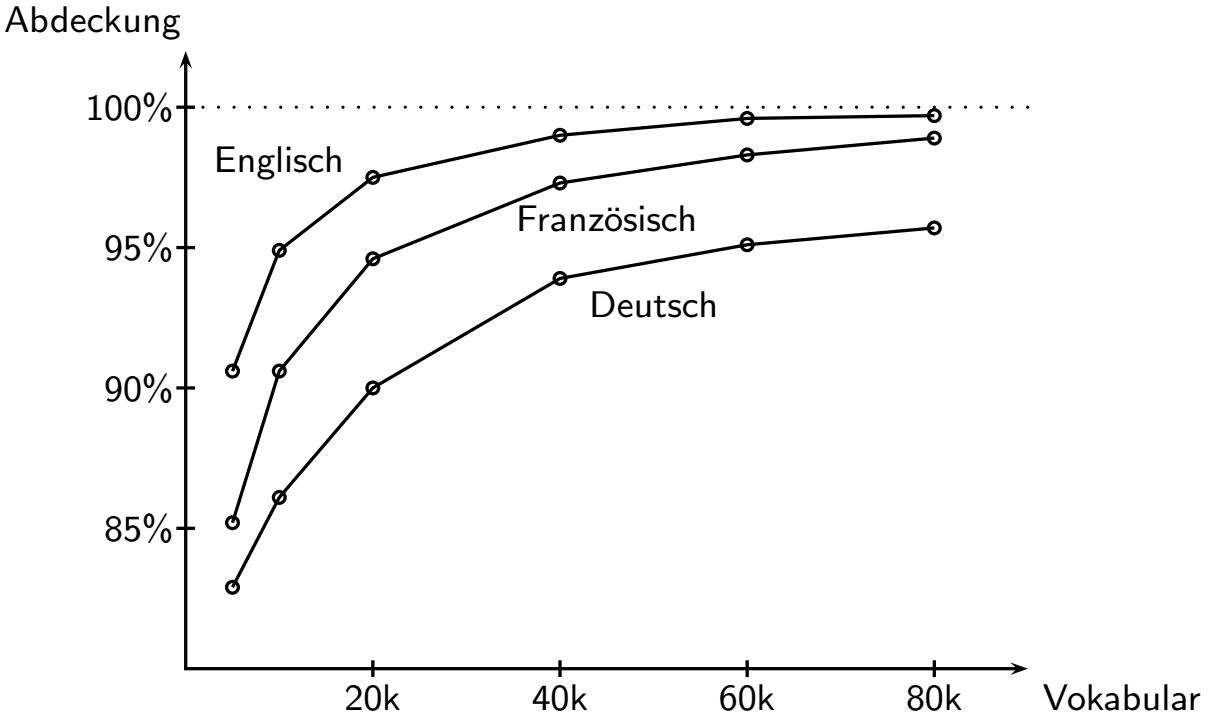


Figure 1. Vocabulary growth comparing German, English and Japanese

Entwicklungsstand und Ausblick

- Wie groß müssen die Wörterbücher sein?



Young et al. 1996

Entwicklungsstand und Ausblick

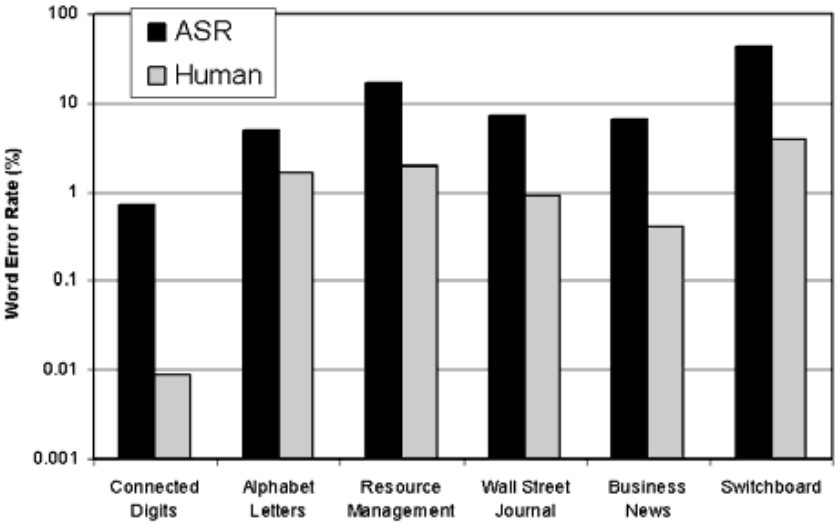
- Wie groß müssen die Wörterbücher sein?

	Englisch	Französisch	Deutsch
Datenquelle	Wall Street Journal	Le Monde	Frankfurter Rundschau
Korpusgröße (Wortformen)	37.2M	37.7M	31.5M
Anzahl unterschiedlicher Wortformen	165k	280k*)	500k*)
Bigram-Perplexität	198	178	430
Trigram-Perplexität	135	119	336
Mittl. Anz. Phone/Wortform	4.16	3.53	5.09
Einzelphon-Wortformen	3%	19%	0.5%

*) Unter Berücksichtigung der Groß-/Kleinschreibung!

Entwicklungsstand und Ausblick

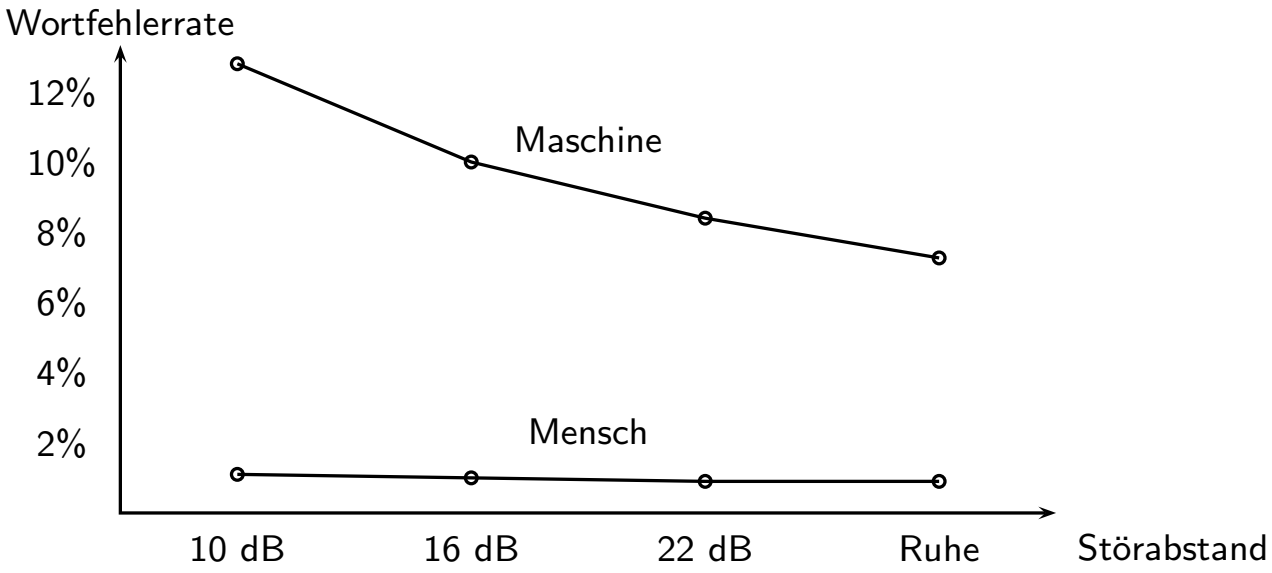
- Wie gut ist die Erkennungssicherheit?



Moore 2003

Entwicklungsstand und Ausblick

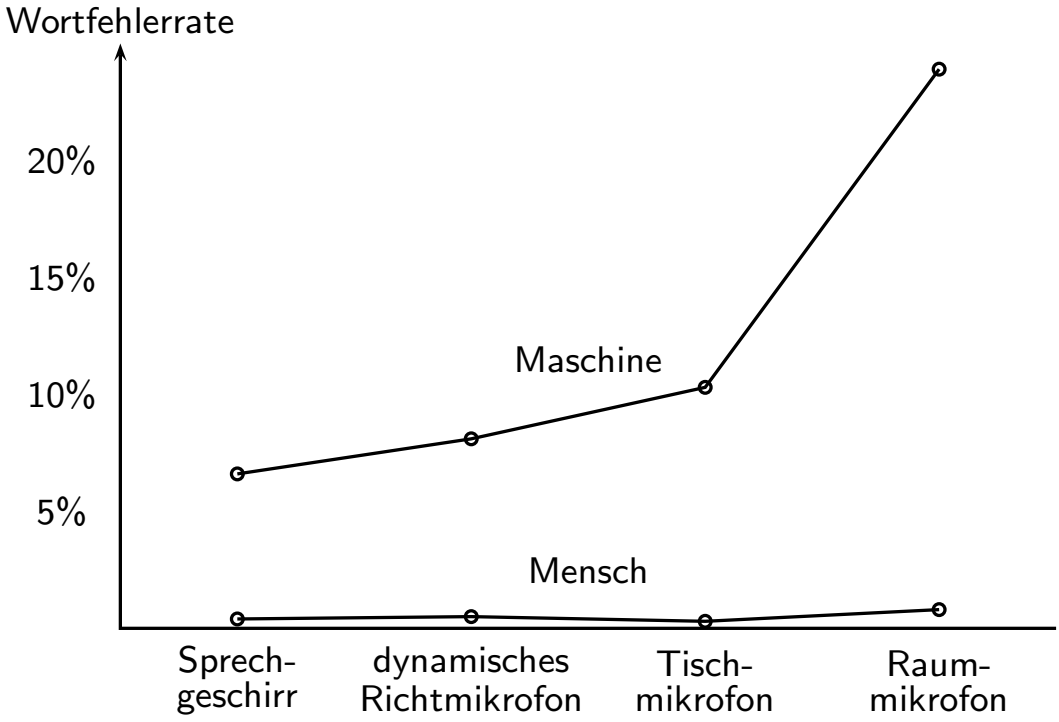
- Wie gut ist die Erkennungssicherheit?



Lippmann 1997

Entwicklungsstand und Ausblick

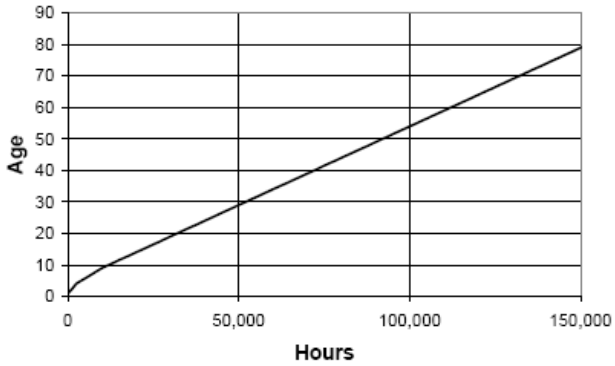
- Wie gut ist die Erkennungssicherheit?



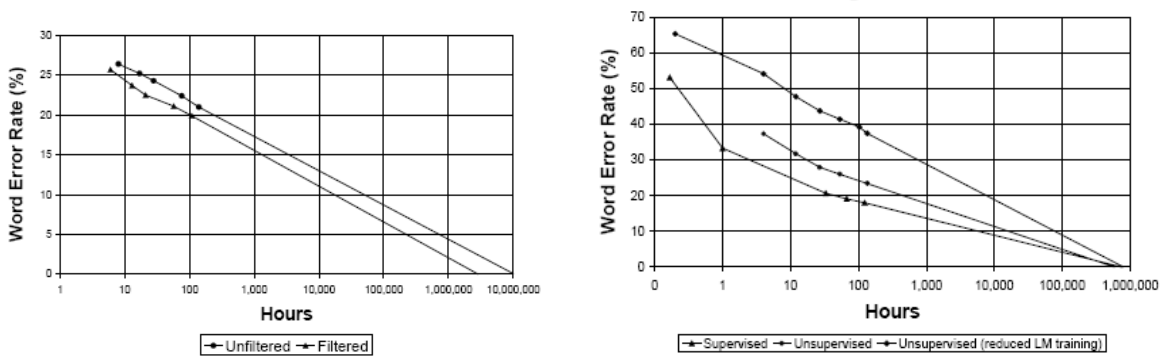
Lippmann 1997

Entwicklungsstand und Ausblick

- Wieviele Daten brauchen wir eigentlich? (Moore 2003)



Geschätzte Gesprächsdauer, der ein Mensch ausgesetzt ist



Extrapolierte Wortfehlerraten für wachsende Größe der Trainingskorpora

Grundlagen der Sprachsignalerkennung

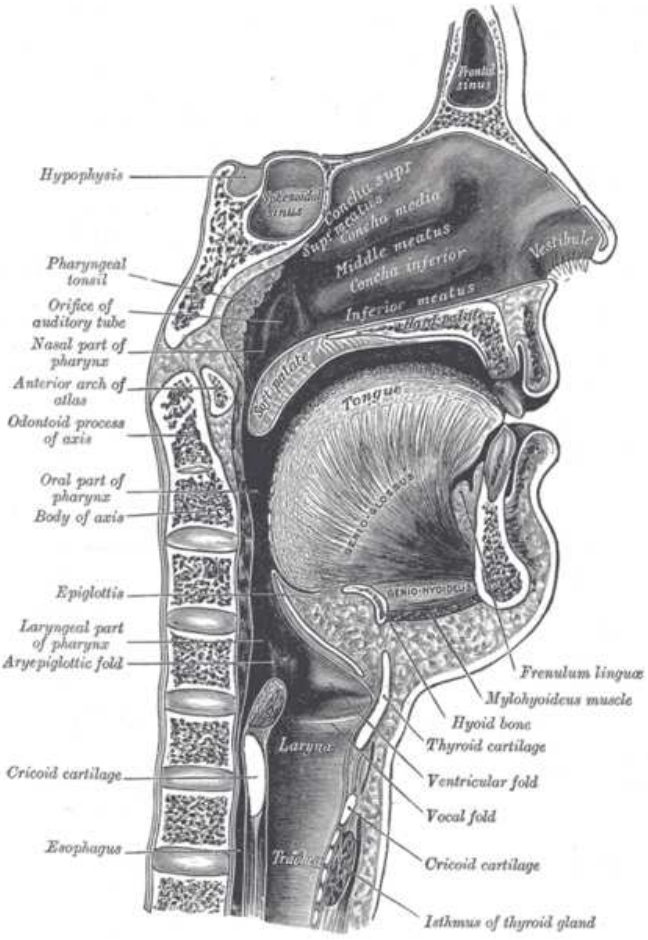
- Spracherkennung als technisches und akademisches Problem
- Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung
- Merkmalsextraktion
- Worterkennung
- Phon-basierte Worterkennung
- Erkennung fließender Sprache
- Systemarchitekturen

Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung

- Sprachsignalerzeugung
- Grundbegriffe aus Phonetik und Phonologie
- Physik des Sprachsignals
- Modellierung der Sprachsignalerzeugung
- Sprachperzeption

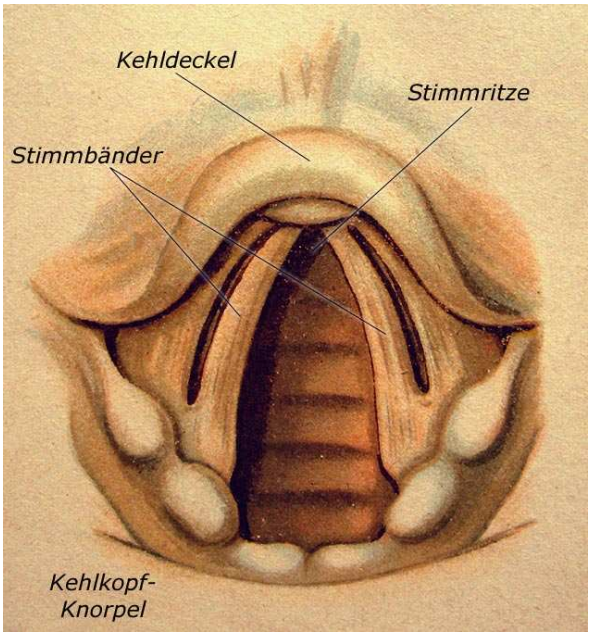
Sprachsignalerzeugung

- Artikulationsorgane
  - Lunge: Druckaufbau Energiezufuhr
  - Stimmlippen: Anregung
  - Vokaltrakt: Filterung



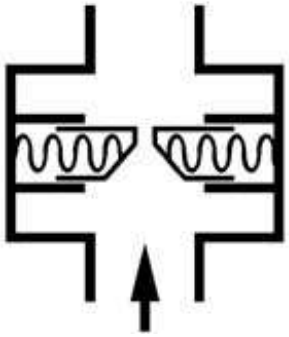
Sprachsignalerzeugung

- Anregung
  - Stimmlippen



Sprachsignalerzeugung

- stimmhafte Anregung
  - Polsterpfeife



Fellbaum (1984)

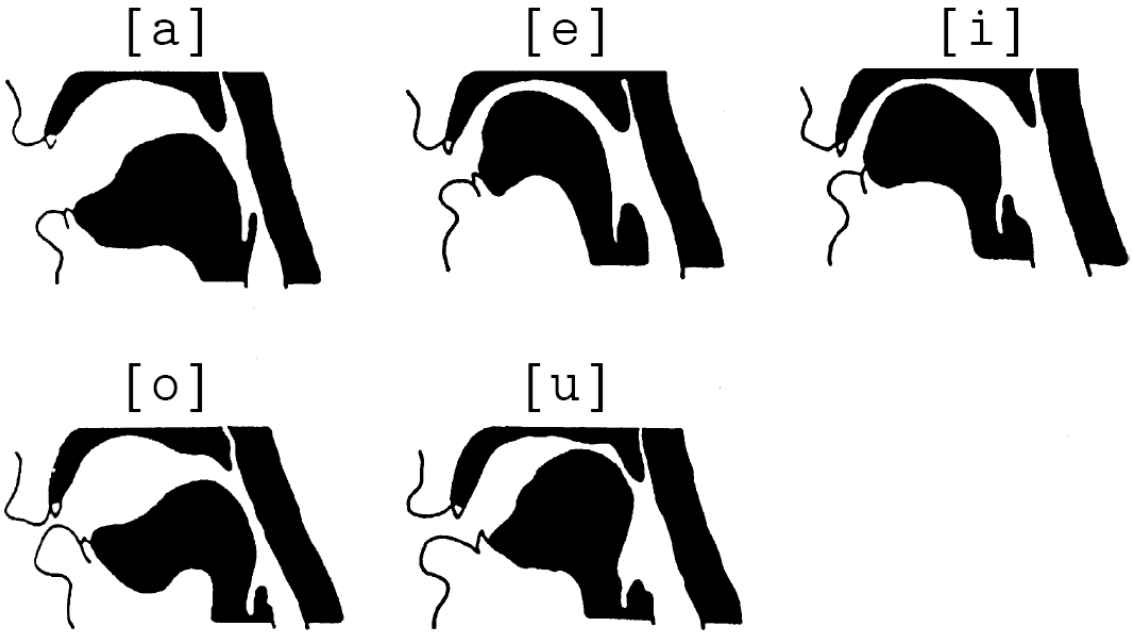
  - Dreiecksfunktion

Hess (1983)

- stimmlose Anregung: rauschförmiges Anregungssignal (Engestelle)
- "gemischte" Anregung

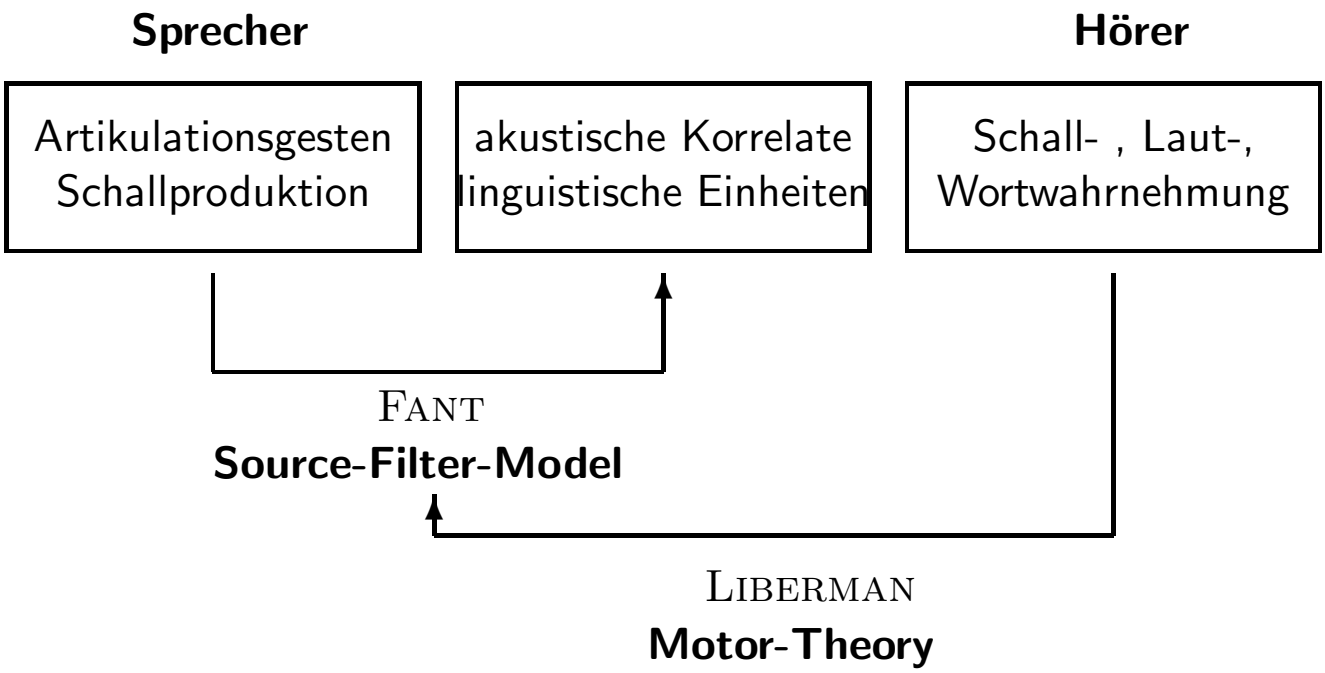
Sprachsignalerzeugung

- Vokaltrakt (Resonanzkörper)



Fellbaum (1984)

Sprachsignalerzeugung



Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung

- Sprachsignalerzeugung
- Grundbegriffe aus Phonetik und Phonologie
- Physik des Sprachsignals
- Modellierung der Sprachsignalerzeugung
- Sprachperzeption

Grundbegriffe aus Phonetik und Phonologie

- akustische Phonetik
  - Beschreibung der Laute nach dem subjektiven Höreindruck
- physikalische Phonetik
  - Untersuchung des Sprachsignals mit physikalischen Methoden
- artikulatorische Phonetik
  - Beobachtung der Artikulationsbewegungen
  - Röntgenaufnahmen des Mundraumes
- Phonologie
  - Klassifikation der Laute nach ihrer Funktion im Sprachsystem (Bedeutungsdifferenzierung)

Grundbegriffe aus Phonetik und Phonologie

- artikulatorische Phonetik: Klassifizierungsprinzipien
  - Artikulationsart
  - Artikulationsort
  - Verschlußlaute/stationäre Laute
  - Konsonanten/Vokale
  - Anregung (stimmhaft/stimmlos)

Laute und Lautklassen

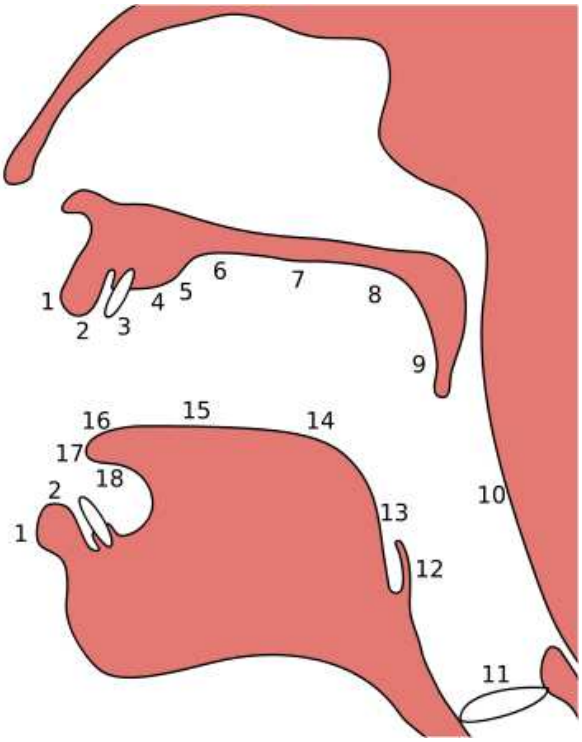
- Klassifizierung der Konsonanten nach der Artikulations*art*
  - Verschlußlaute (Plosive)
    - Druckaufbau (Verschlußphase)
    - Verschlußlösung (Explosivphase, "Burst")
    - stimmhaft ([b], [d], [g]), stimmlos ([p], [t], [k])
  - Reibelaute (Frikative, Spiranten)
    - Engebildung im Mund- oder Rachenraum
    - rauschartiger Laut
    - stimmhaft ([v], [z], [j]), stimmlos ([f], [s], [$\int$])
  - Nasale
    - Mundhöhle weitgehend verschlossen, Luft entweicht durch die Nase
    - immer stimmhaft ([m], [n], [η])

Laute und Lautklassen

- Klassifizierung der Konsonanten nach der Artikulations*art*
  - Seitenlaute (Laterale)
    - Luftstrom führt rechts und links an der Zunge vorbei
    - immer stimmhaft ([l])
  - Intermittierende (Vibranten)
    - Zäpfchen oder Zunge schwingen vom Luftstrom angeregt
    - immer stimmhaft ([r], [R])

Laute und Lautklassen

- Klassifizierung der Konsonanten nach dem Artikulations*ort*



2+2 Lippen: bilabial ([p], [b])

2+3 Lippen + Zähne: labiodental ([f])

16+3 Zunge + Zähne: dental ([s])

16+4 Zunge + Zahnfächer: alveolar ([d], [t])

15+7 Zunge + harter Gaumen: palatal ([ç])

14+8 Zunge + weicher Gaumen: velar ([k])

14+9 Zunge + Zäpfchen: uvular ([R])

11 Stimmritze: glottal ([h])

Laute und Lautklassen

- Überblick: Konsonanten

Art/Ort	lab.	alv.	p-alv.	pal.	vel.	uvul.	glot.
Plos:	p b	t d			k g		ʔ
Affrik:	pf	ts	tʃ				
Frik:	f v	s z	ʃ ʒ	ç	x	ʁ	h ɦ
Son:	m	n			ŋ		
Approx:		l		j			

Barry 2003

Laute und Lautklassen

- Klassifizierung der Vokale nach der Artikulations*art*
  - offen: [a:]
  - geschlossen: [u:]
  - gerundet: [o:]
  - nicht gerundet: [e:]

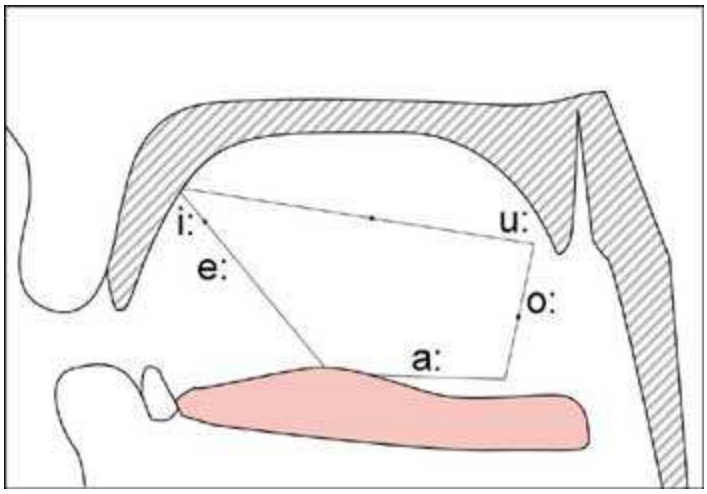
Laute und Lautklassen

- Klassifizierung der Vokale nach dem Artikulations*ort* (Zungenwölbung)
  - Horizontalposition für den höchsten Punkt des Zungenrückens
    - Vorderzungenvokale, palatale Vokale
      in der Nähe der Zähne
      helle Vokale ([i:]))
    - Hinterzungenvokale, alveolare Vokale
      in der Nähe des Zäpfchens
      dunkle Vokale ([u:])
  - Vertikalposition des Zungenrückens
    - hoch ([i:])
    - tief ([a:]))

Laute und Lautklassen

- Sonderformen
  - Diphthonge (Gleitlaute) ([a͡o], [a͡e], [o͡ø])
  - nasalierte Vokale ([ɛ̃]))

- Vokalviereck



Grundbegriffe aus Phonetik und Phonologie

- Laut / Phon
  - kleinste akustisch/artikulatorisch wahrnehmbare Einheit
  - Schreibweise: eckige Klammern
  - elementare Lautverbindungen: Diphone (Triphone, ...)

- Phonem
  - kleinste bedeutungs*differenzierende* Einheit
    - am Sprachsystem orientiert, nicht am Signal!
  - Schreibweise: Schrägstriche
  - Minimalpaare (/ta:k/, /la:k/): morphemdifferenzierend

Grundbegriffe aus Phonetik und Phonologie

- Allophone
  - unterschiedliche lautliche Realisierungen für ein Phonem
    - freie Allophone (kontextunabhängig) [r] ↔ [R]
    - stellungsbedingte Allophone (kontextabhängig) [iç] ↔ [ax] (auch: [h])
    - teilweise sehr viele Allophone
      z.B. ca. 100 [k] -Allophone
      Kiste, Kasten, Keller, Kohle, kühl, klein, wacker, ...

Grundbegriffe aus Phonetik und Phonologie

- Morph
  - kleinste bedeutungs*tragende* lautliche Einheit
  - am Sprachsystem orientiert!

- Morphem
  - ein oder mehrere bedeutungsgleiche Morphe, die sich in der gleichen lautlichen Umgebung gegenseitig ausschließen (Allomorphe)

- Allomorph
  - stellungsbedingte: arbeit-*et* / schläf-*t*
  - freie: *Wandr*-er / *Wander*-er

Grundbegriffe aus Phonetik und Phonologie

- (Sprech-)Silbe
  - artikulatorische Einheit
  - alternative Gliederung zur Morphemzerlegung
  - Silbengliederung

    Silbe $\rightarrow$ Onset Rhyme

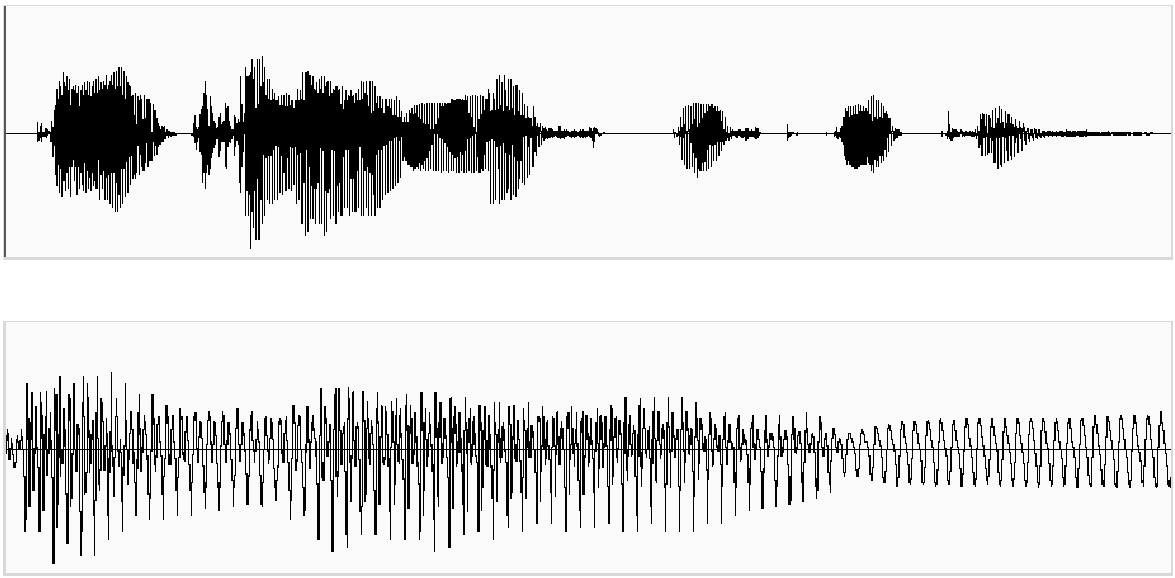
    Rhyme $\rightarrow$ Peak Coda
  - Silbenmodelle
    - CV ([da:])
    - CVC ([ta:k])
    - VCC ([ast])
  - Halbsilben (Onset + Peak, Peak + Coda)

Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung

- Sprachsignalerzeugung
- Grundbegriffe aus Phonetik und Phonologie
- Physik des Sprachsignals
- Modellierung der Sprachsignalerzeugung
- Sprachperzeption

Physik des Sprachsignals

- Schalldruckzeitfunktion $p(t)$



Physik des Sprachsignals

- Schwankungen von 20 μPa bis 60 Pa
- Schalldruckpegel

$$L = 20 \lg \frac{\tilde{p}}{p_0} \qquad \text{(in dB)}$$

$$\tilde{p} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} |p(t)| \, dt$$

$$p_0 = 20 \ \ \mu\text{Pa} \qquad \text{(Hörschwelle bei 1 kHz)}$$

Physik des Sprachsignals

- Schalleistungspegel (energiebezogen)
  - Energie

$$E = \int_{t_1}^{t_2} p^2(t)\, dt$$

  - Pegel

$$L = 10 \lg \frac{P}{P_0} \qquad \text{(in dB)}$$

$$P = \frac{E}{t} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} p^2(t)\, dt$$

$$P_0 = 10^{-12}\ \text{W} \qquad \text{(Hörschwelle bei 1 kHz)}$$

- Schallintensitätspegel (Leistung pro Flächeneinheit)

Physik des Sprachsignals

- Sprachsignal ist nichtstationär
  - (quasi-) periodische und nichtperiodische Abschnitte (stimmhaft/stimmlos)
  - periodisch:

$$\forall t : x(t) = x(t + T)$$

T	Periode
$f_0 = \frac{1}{T}$	Grundfrequenz

Physik des Sprachsignals

- Beschreibung der Frequenzcharakteristik (Spektrum)
- Sprachsignal kann dargestellt werden als Überlagerung sehr vieler reiner Sinusschwingungen

$$x(t) = A \sin(\omega t + \varphi)$$

A	maximale Amplitude
$\omega = 2\pi f$	Kreisfrequenz
φ	Phasenverschiebung

Physik des Sprachsignals

- Spektrum: Summierung über viele Sinusschwingungen

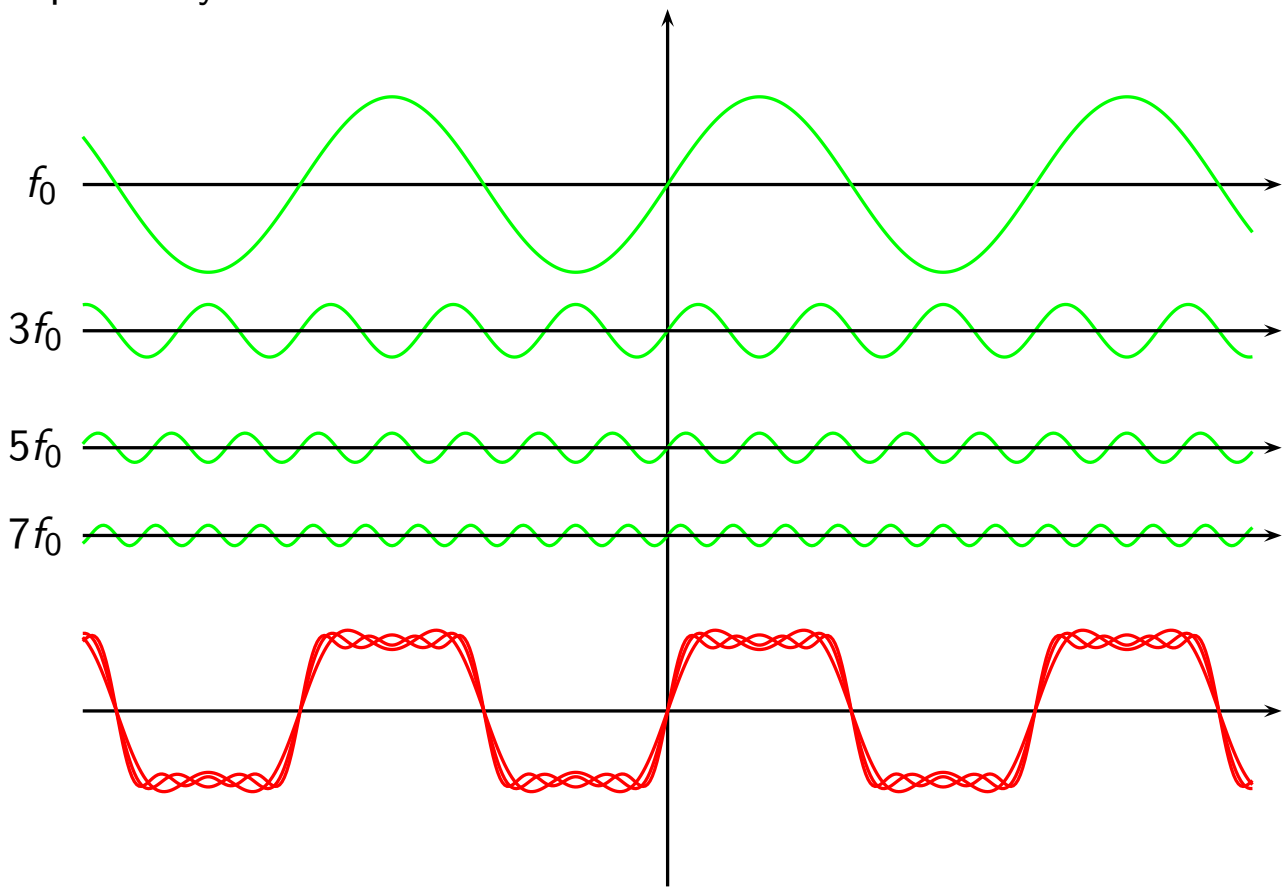
$$x(t) = \int_0^\infty A(\omega) \sin(\omega t + \phi(\omega)) \, d\omega$$

$A(\omega)$	Amplitudenspektrum
$\phi(\omega)$	Phasenspektrum

- Phasenspektrum wird meist vernachlässigt
- Frequenzumfang der Sprache: 60 Hz bis 10 kHz

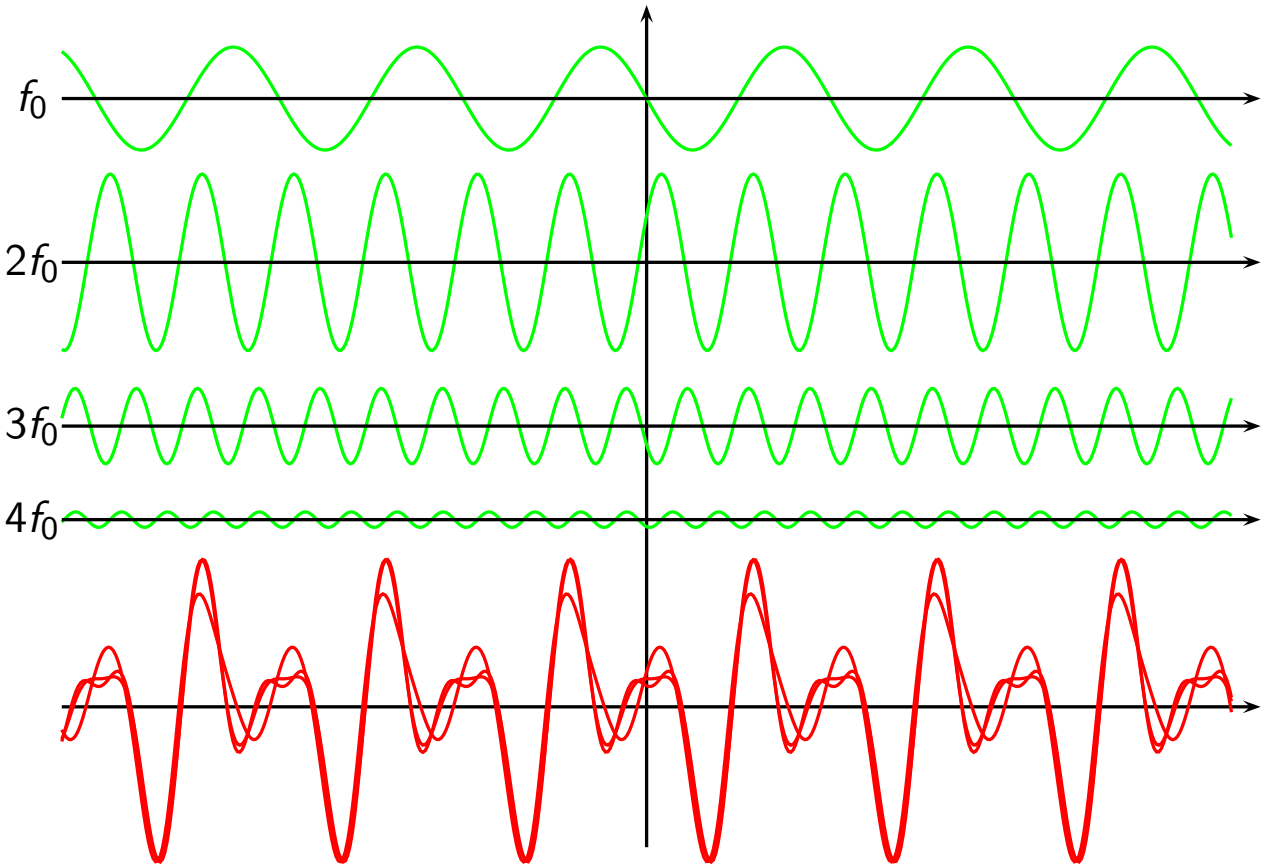
Physik des Sprachsignals

- Spektralsynthese: Rechteckfunktion



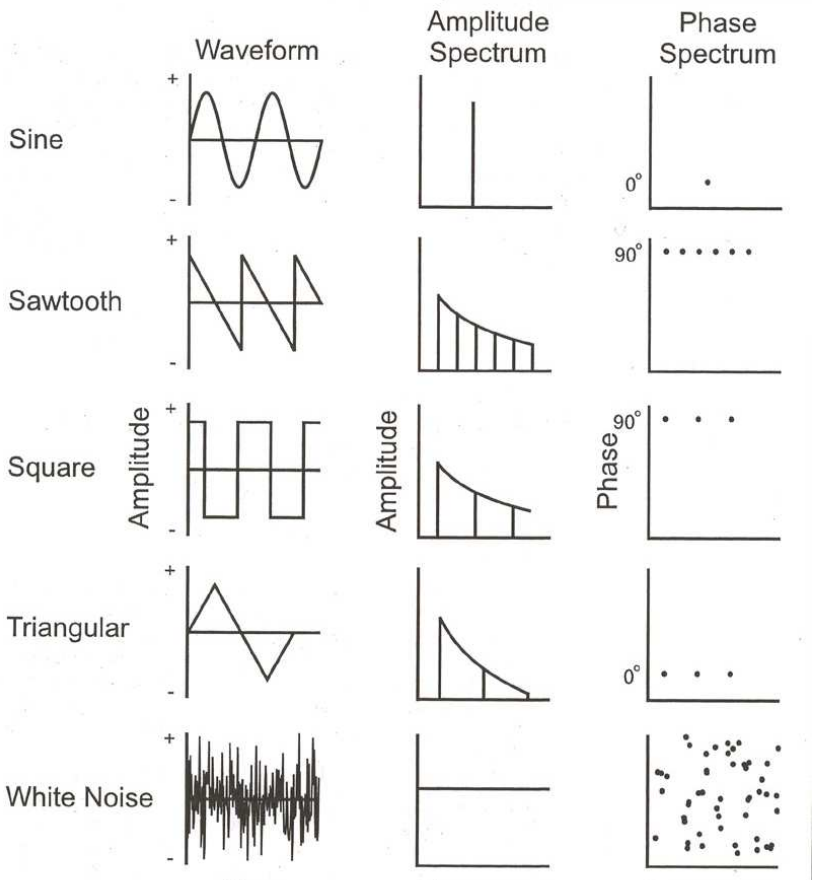
Physik des Sprachsignals

- Spektralsynthese: vokalischer Laut



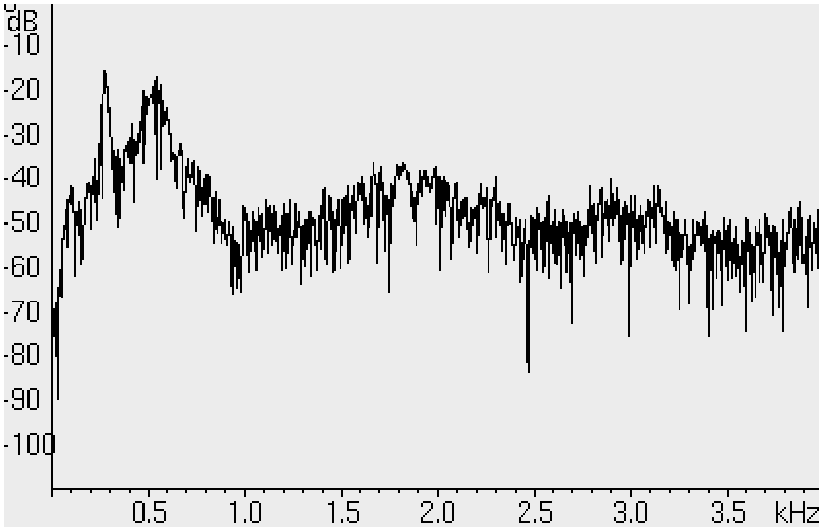
Physik des Sprachsignals

- Zeitfunktionen und ihre Spektraldarstellung



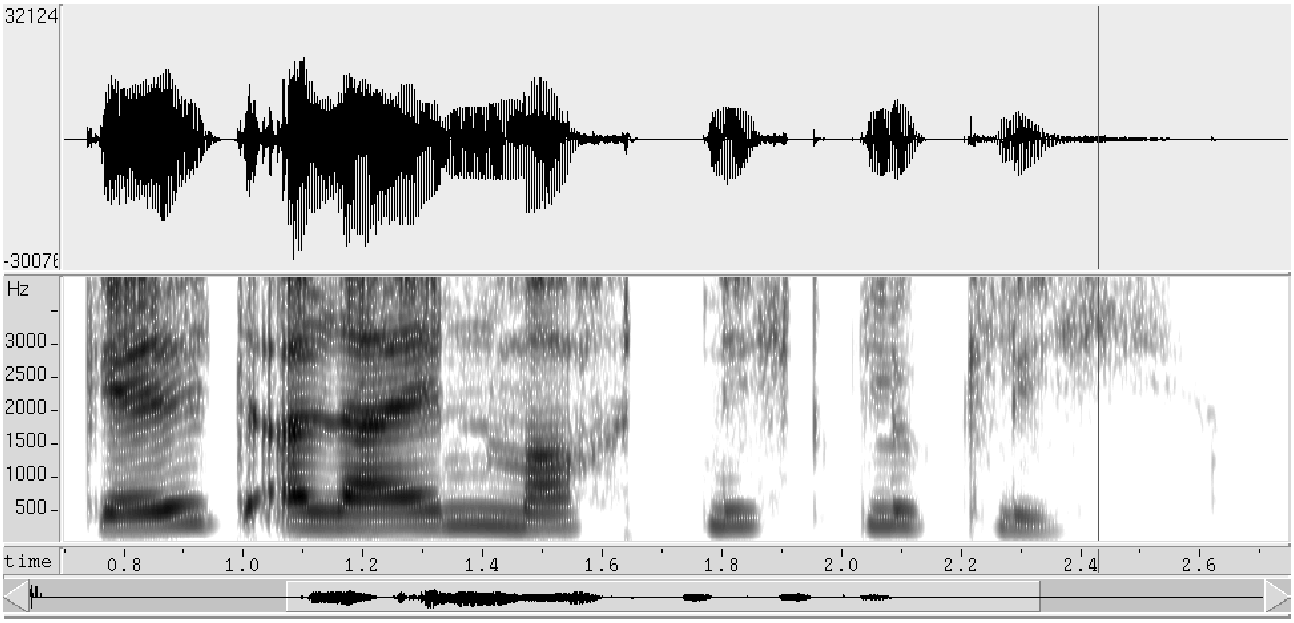
Physik des Sprachsignals

- lokale Maxima im Spektrum: Formanten $F_1, F_2, F_3, \ldots$



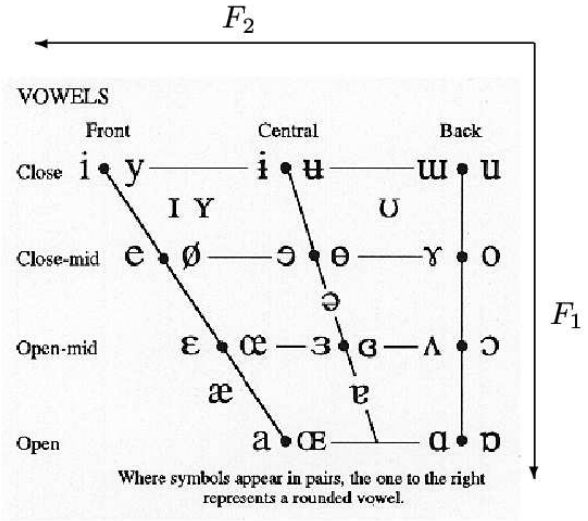
Physik des Sprachsignals

- Formanten im Spektrogramm



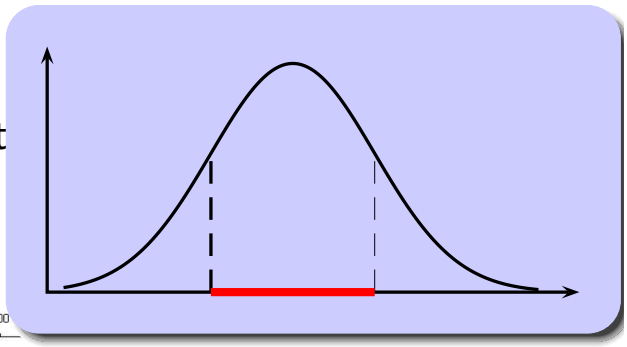
Physik des Sprachsignals

- Vokaldetektion: F_1/F_2

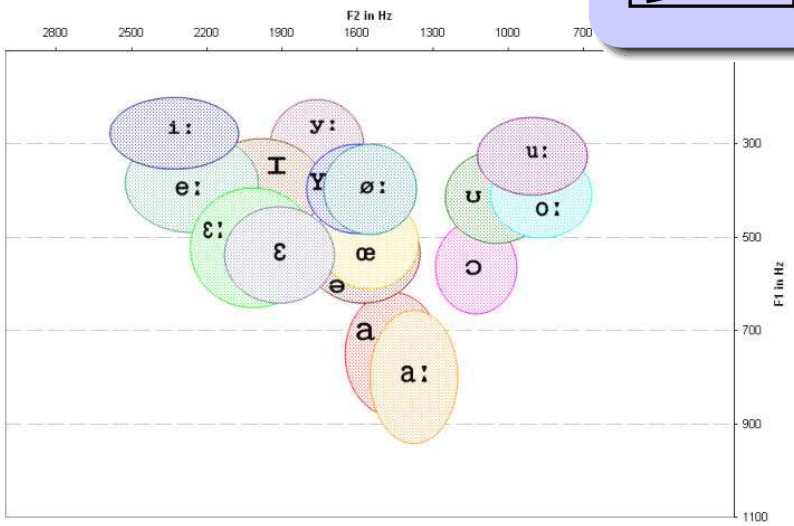


- Voraussetzungen
  - gut artikuliert, stationär, ein Sprecher, gleiche Aufnahmebedingungen
  - bei kontinuierlicher Sprache: Verschiebung der Vokalschwerpunkte zum schwa-Laut
- Plosive: Charakteristische Vokalbewegungen
- frühe Idee: Verwendung von Formanten zur Lauterkennung (akustische Schreibmaschine, Bandbreitenkompression)

Physik des Sprachsignals



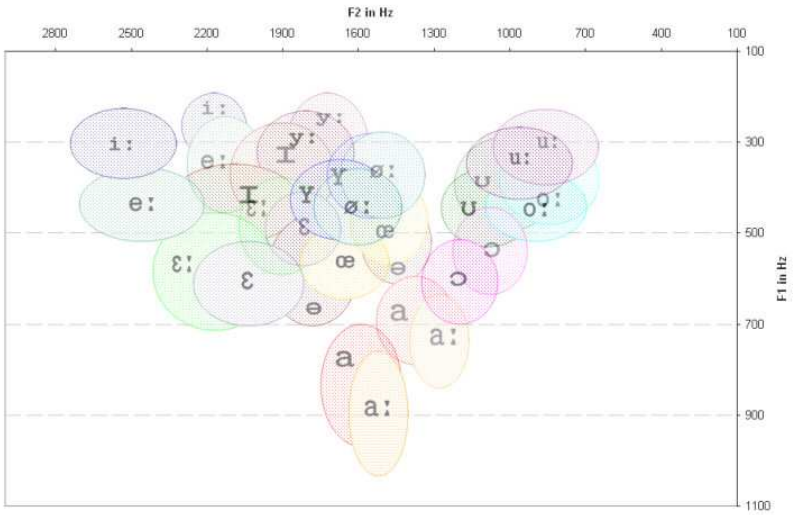
- Formanten sind nicht verlässlich detekt
  lokal Maxima im Spektrum)
- große Überlappungsbereiche



Sendlmeier and Seebode 2006

Physik des Sprachsignals

- große interpersonelle Unterschiede (hier: Geschlecht)



Sendlmeier and Seebode 2006

Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung

- Sprachsignalerzeugung
- Grundbegriffe aus Phonetik und Phonologie
- Physik des Sprachsignals
- Modellierung der Sprachsignalerzeugung
- Sprachperzeption

Modellierung der Sprachsignalerzeugung

- Lineares System: Superpositionsprinzip

$$Tr\left(a \cdot u(t) + b \cdot v(t)\right) = a \cdot Tr\left(u(t)\right) + b \cdot Tr\left(v(t)\right)$$

- Es können keine zusätzlichen Frequenzen erzeugt werden!
- Spezialfall eines linearen Systems: Filter
  - im Zeitbereich: Faltung des Signals $s(t)$ mit der Impulsantwort $h(t)$ des Filters

$$x'(t) = x(t) \otimes h(t) = \int_{-\infty}^{+\infty} x(t-\tau) \cdot h(\tau)\, d\tau$$

  - im Frequenzbereich: Multiplikation der Spektren

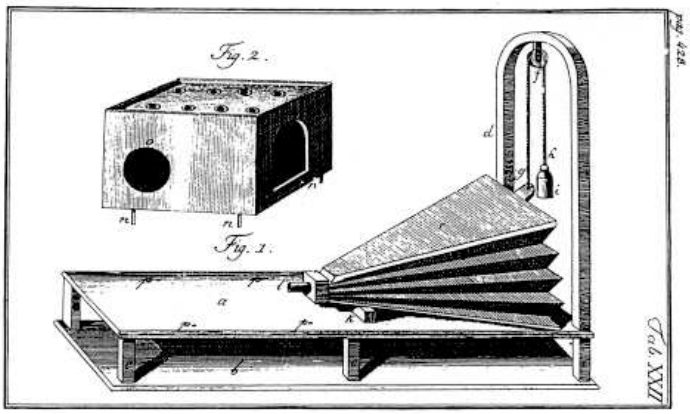
$$X'(\omega) = X(\omega) \cdot H(\omega)$$

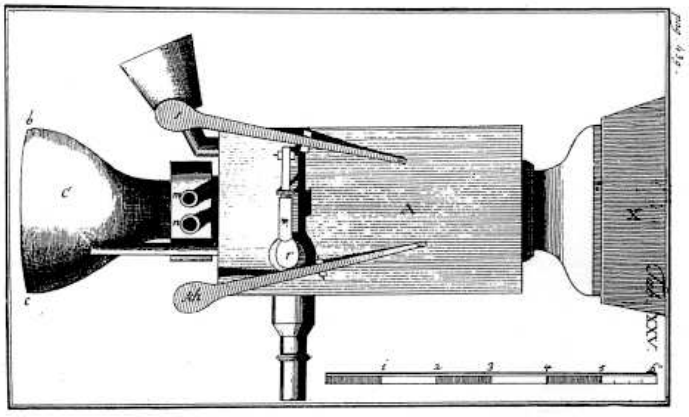
Modellierung der Sprachsignalerzeugung

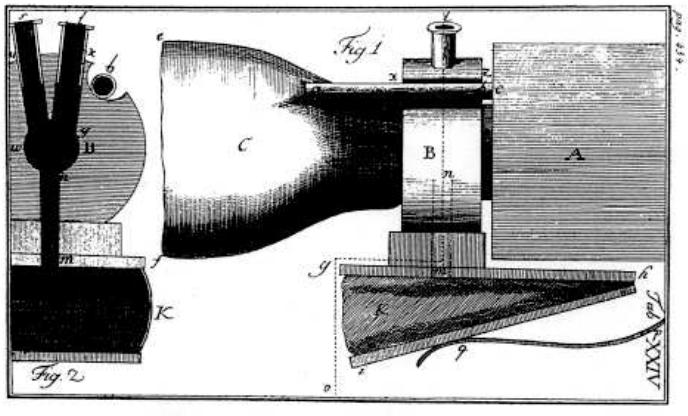
- Filtertypen
  - Tiefpass
  - Hochpass
  - Bandpass
- Verwendung zur Modellierung von Vokaltrakt und Abstrahlungscharakteristik

Modellierung der Sprachsignalerzeugung

- Wolfgang von Kempelen (1791)







Modellierung der Sprachsignalerzeugung

- zu modellierende Phänomene
  - Anregung (Laryngalsignal)
  - Form des Vokaltraktes als Funktion der Zeit
  - Verluste an den Vokaltraktwänden (Wärmeleitung, Flüssigkeitsverformung, Reibung)
  - Elastizität der Wände
  - Abstrahlung von den Lippen
  - Berücksichtigung des Nasenraums
- Vereinfachungen erforderlich

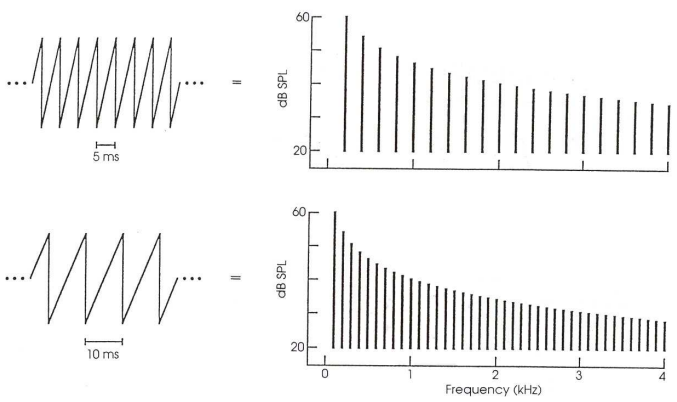
Modellierung der Sprachsignalerzeugung

- Source-Filter-Modell (Gunnar Fant 1960)

$$x(t) = r(t) \otimes v(t) \otimes a(t)$$

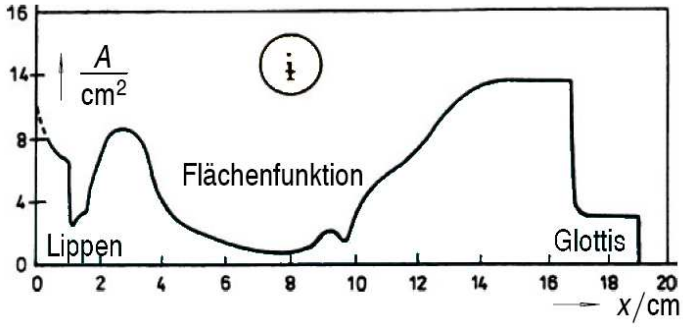
$r(t)$	Abstrahlung von den Lippen (Hochpass)
$v(t)$	Vokaltraktform (Resonator, Bandpass)
$a(t)$	Anregungsfunktion (Folge von Dreiecksimpulsen bzw. weißes Rauschen)

  Abstrahlung wird meist mit in die Vokaltraktcharakteristik eingerechnet
- Anregung für Vokale ist in erster Näherung Dreiecksfunktion



Modellierung der Sprachsignalerzeugung

- Filterwirkung des Vokaltrakts bestimmt durch seine Flächenfunktion (ortsabhängige Querschnittsfläche)

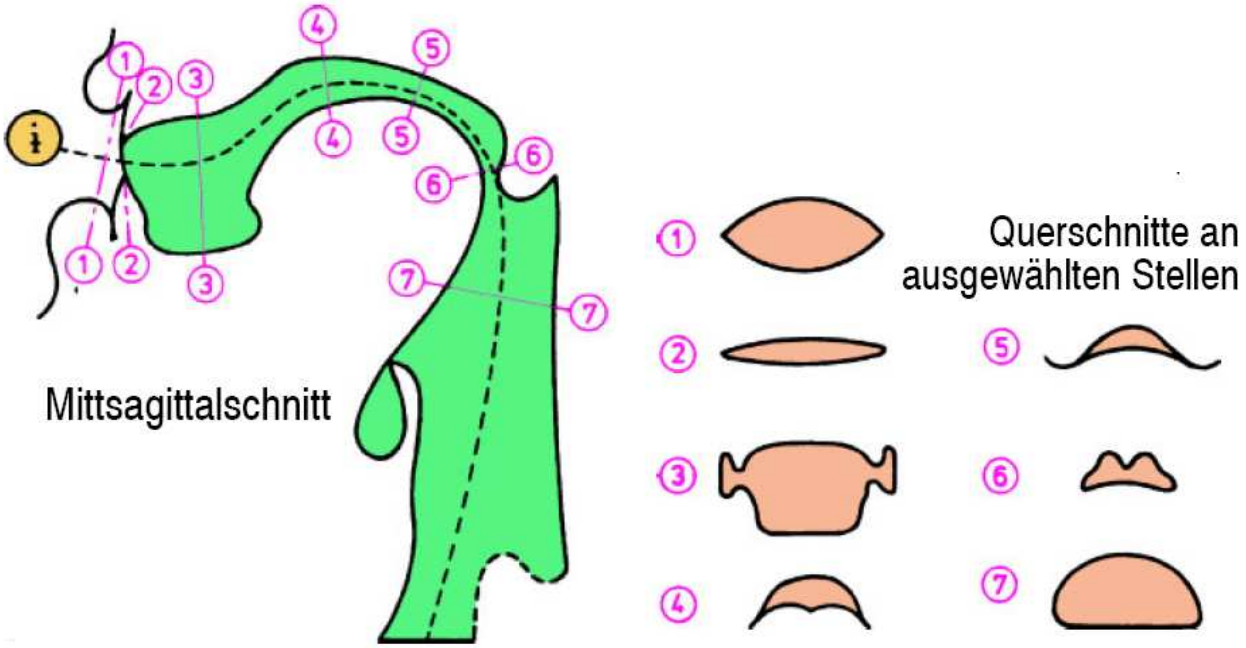


nach Fant (1960)

- Resonanzen des Vokaltrakts entsprechen den Formanten im Signalspektrum

Modellierung der Sprachsignalerzeugung

- Vokaltraktquerschnitte



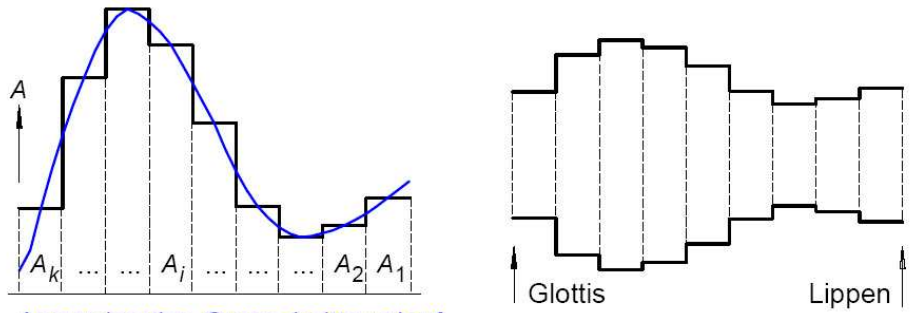
(nach Fant (1960))

Modellierung der Sprachsignalerzeugung

- Ungeheuer (1962): Querschnittform ist akustisch irrelevant
- Approximation durch Röhrenmodell
  - kreisrunder Querschnitt
  - diskrete Längsschnittsfunktion (Treppenfunktion)
  - (meist) gleiche Länge der Teilabschnitte
  - Durchmesser ist klein im Vergleich zur Schallwellenlänge
  - Wand ist schallhart (keine Verluste)

Modellierung der Sprachsignalerzeugung

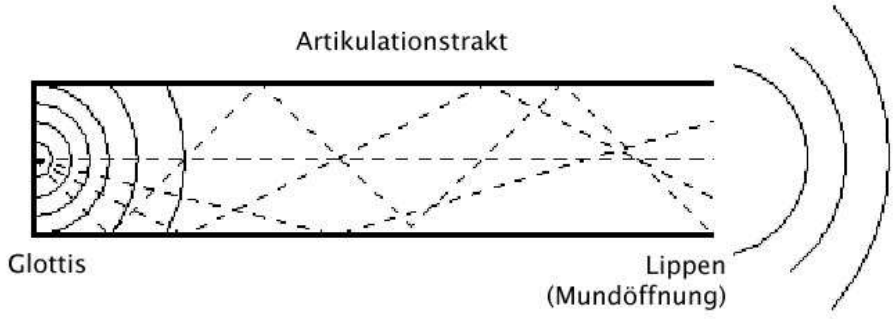
- Röhrenmodell des Vokaltrakts



Approximation Querschnittsverlauf durch stückweise homogene Leitung

Röhrenmodell

Hess (1983)



Hess (1983)

Modellierung der Sprachsignalerzeugung

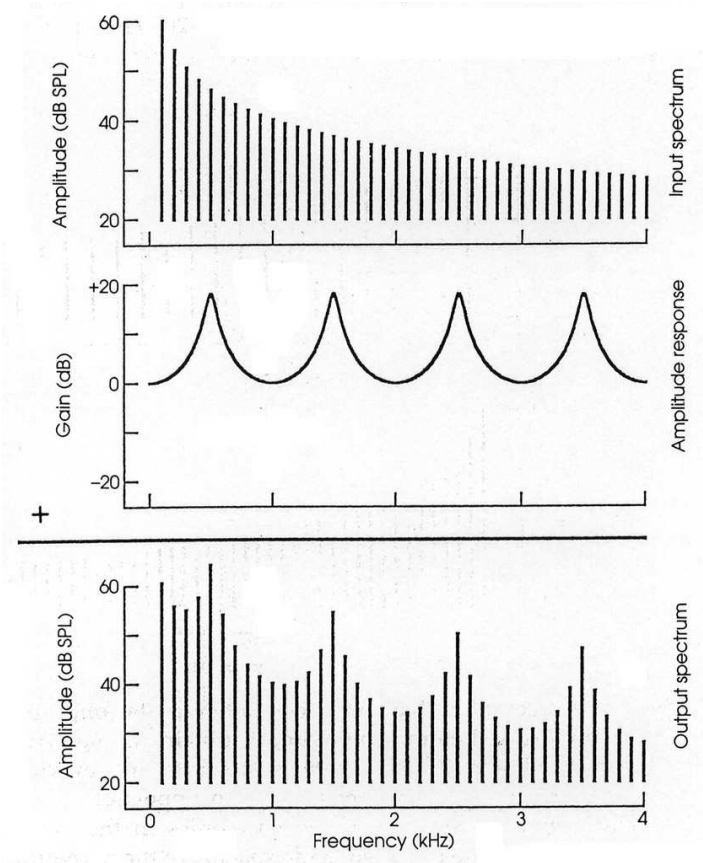
- Zielstellungen:
  1. Nachbildung realer Längsschnittsfunktionen des Vokaltrakts
     → vocal tract analog
  2. Variation der Vokaltraktparameter, bis optimale Übereinstimmung mit einem vom Menschen produzierten Signal gegeben ist.
     → terminal tract analog

- Röhrenmodell ist schlecht geeignet für Nasale und nasalierte Vokale
  - Ankopplung eines zusätzlichen Resonanzraumes (Antiformanten)

- Anwendung in der Spracherkennung: Vokaltraktadaption durch Längennormierung des Spektrums

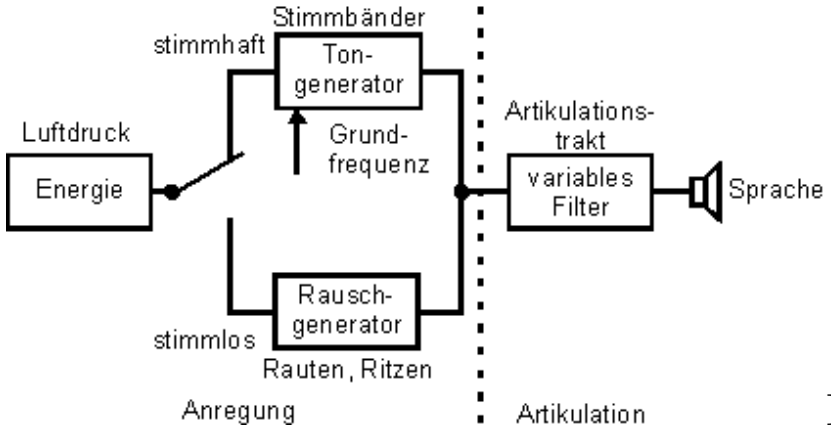
Modellierung der Sprachsignalerzeugung

- Superposition: Faltung ist Multiplikation im Frequenzbereich



Modellierung der Sprachsignalerzeugung

- "Schaltbild" der menschlichen Spracherzeugung



Fellbaum 1984

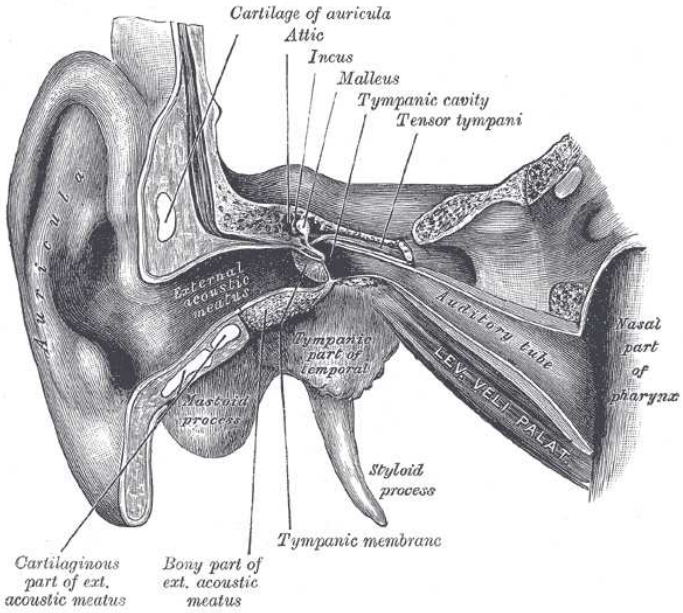
- Source-Filter-Model hat insgesamt Tiefpasscharakter
  - Dreiecksfunktion als Anregung: 12 dB/Oktave abfallender Frequenzgang
  - Abstrahlung (Hochpaß): 6 dB/Oktave ansteigender Frequenzgang
  - zusammen: 6 dB/Oktave abfallender Frequenzgang

Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung

- Sprachsignalerzeugung
- Grundbegriffe aus Phonetik und Phonologie
- Physik des Sprachsignals
- Modellierung der Sprachsignalerzeugung
- Sprachperzeption

Sprachperzeption

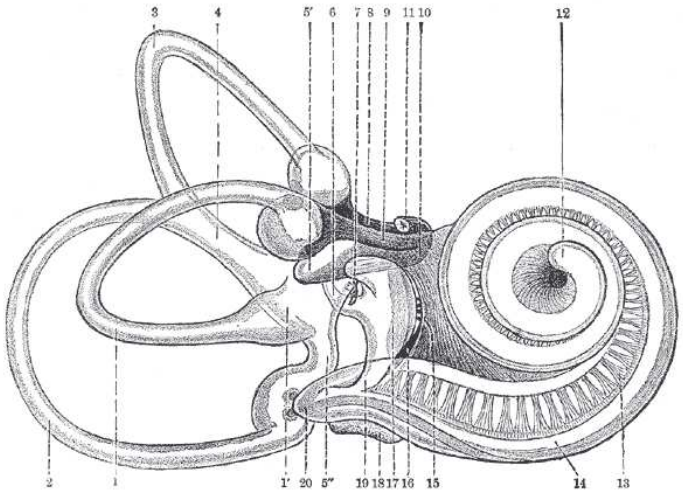
- Das Ohr als Nachrichtenempfänger (Zwicker 1967)
- bildet akustische Stimuli auf neuronale Aktivität ab

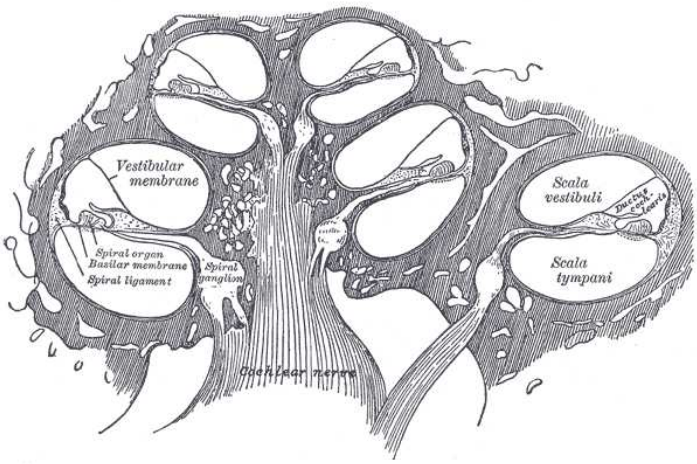


- äußeres Ohr, Mittelohr, Innenohr
- Trommelfell → ovales Fenster
- Verbindung über Gehörknöchelchen: malleus/Hammer, incus/Amboss, stapes/Steigbügel

Sprachperzeption

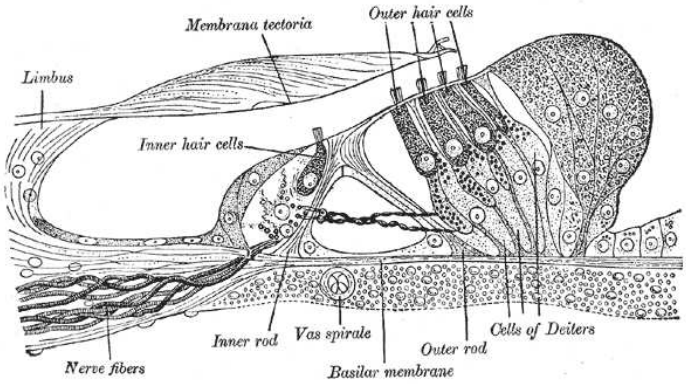
- Innenohr: Vestibül, Bogengänge, cochlea (Schnecke)





Sprachperzeption

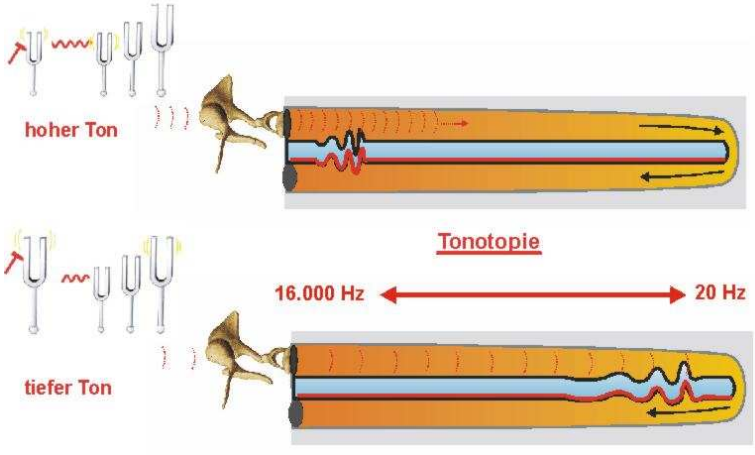
- akustischer Reiz → psychoakustische Empfindungsgröße
  - $f < 1$ kHz: direkte Umsetzung in entsprechende Nervenimpulse
  - $f > 1$ kHz: Abbildung auf einen geometrischen Ort auf der Basilarmembran



- Phasenwahrnehmung ist irrelevant für Sprache und Musik

Sprachperzeption

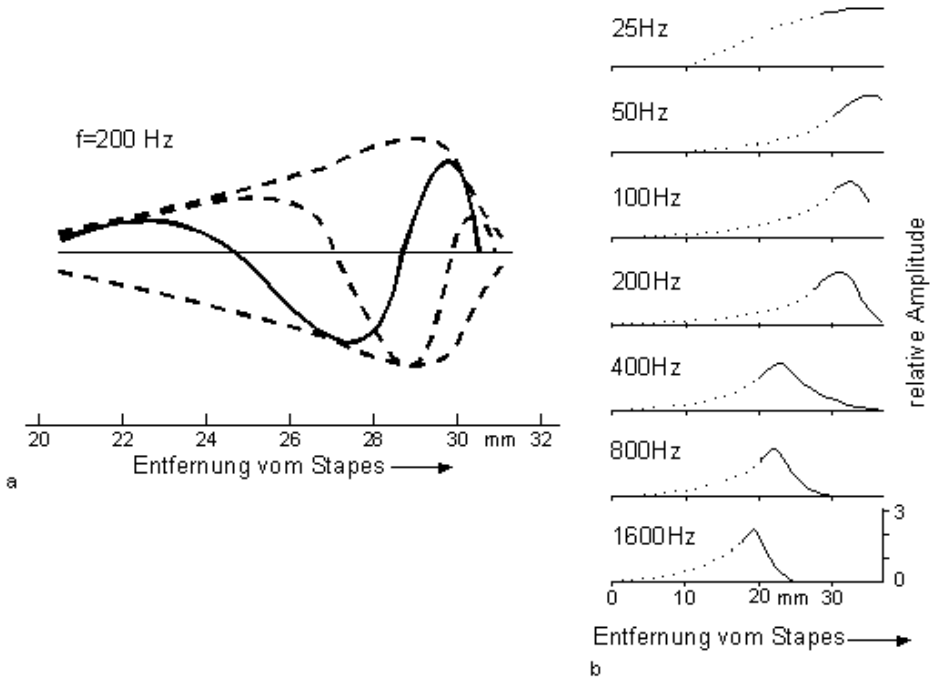
- Frequenz-Orts-Kodierung über Wanderwellen



- äußere Haarzellen: Verstärkung (60 dB)
- innere Haarzellen: Frequenzkodierung (3000-4000 Töne)

Sprachperzeption

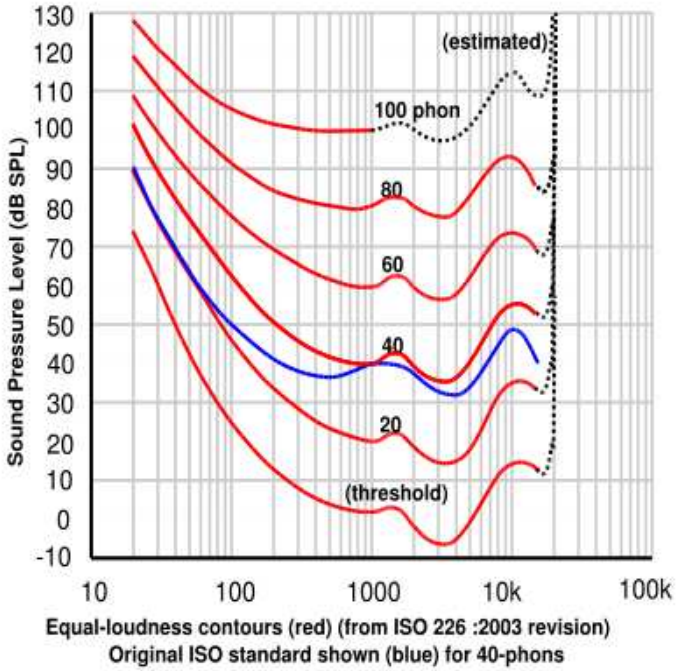
- Frequenz-Orts-Kodierung über Wanderwellen



Fellbaum (1984) nach Bekesy (1960)

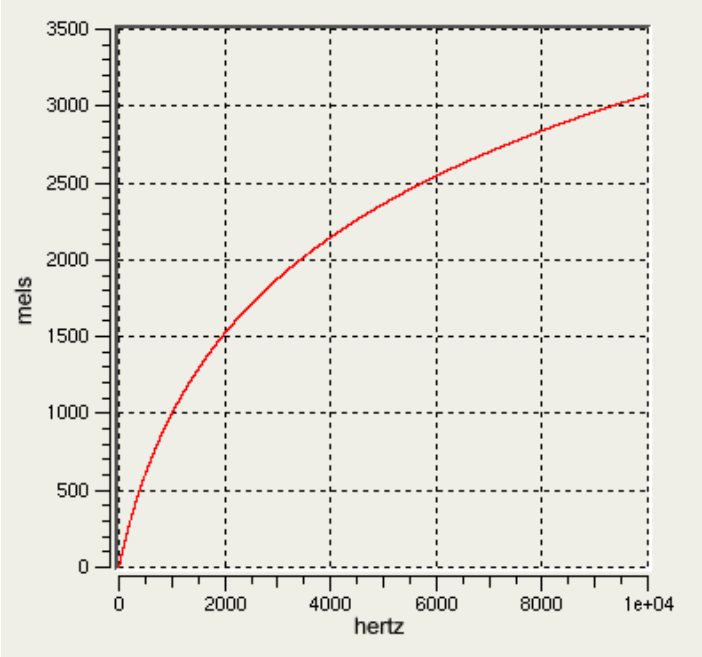
Sprachperzeption

- das Ohr ist kein idealer Sensor
- Lautstärkeempfindung (in phon) ist nicht proportional zum Schallpegel (in dB)
- Lautstärkeempfindung hängt von der Frequenz ab



Sprachperzeption

- nichtlineare Transformation von Information:
  - Tonhöhenempfindung ist nicht proportional zur Tonhöhe
  - mel-Skala



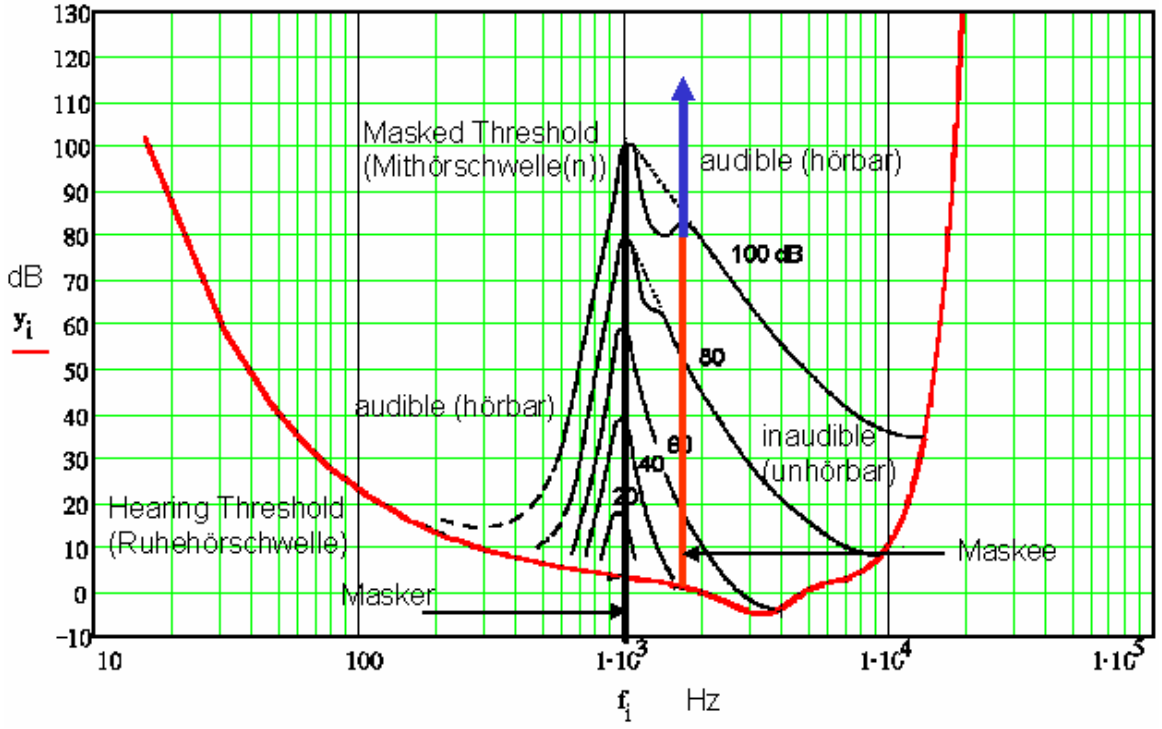
- Nachweis trotz Harmoniempfinden:
  Schmalbandrauschen, Sinustöne, keine Obertöne

Sprachperzeption

- Aggregation von Information
  - integrierte Wahrnehmung von Frequenzbändern (Frequenzgruppen)
  - 24 Frequenzgruppen füllen das Spektrum lückenlos aus
  - Frequenzgruppenbreite wächst mit steigender Frequenz
  - Frequenzgruppenbreite (in mel) ist über das gesamte Frequenzband konstant!
  - Frequenzgruppen können um jede beliebige Mittenfrequenz gebildet werden

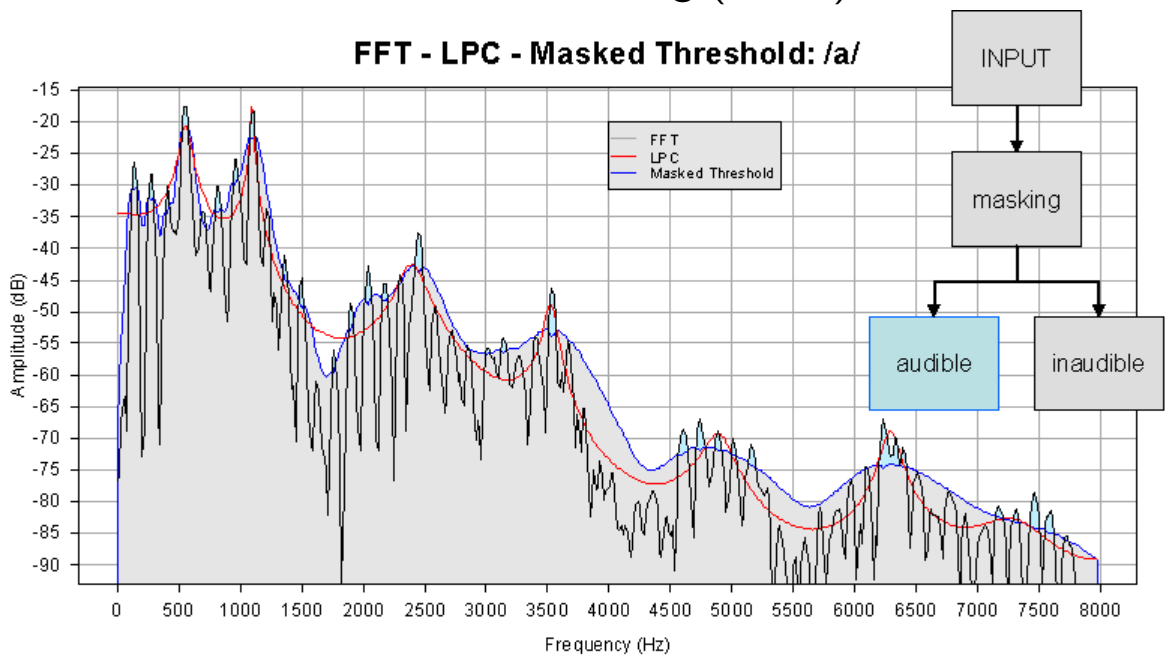
Sprachperzeption

- Verlust von Information: Maskierung



Human speech perception

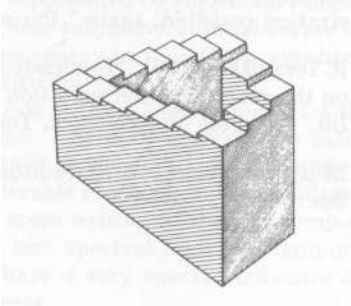
- Verlust von Information: Maskierung (Forts.)



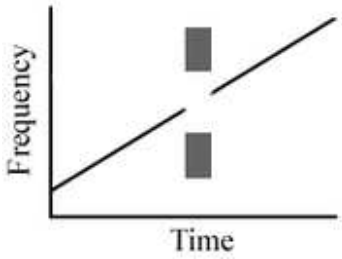
- Hinzufügen von Information:
  Residualtoneempfindung (Glocken, weibliche Stimme)

Sprachperzeption

- akustische Illusionen
  - Shepard-Töne:



  - Kontinuitätsillusion:

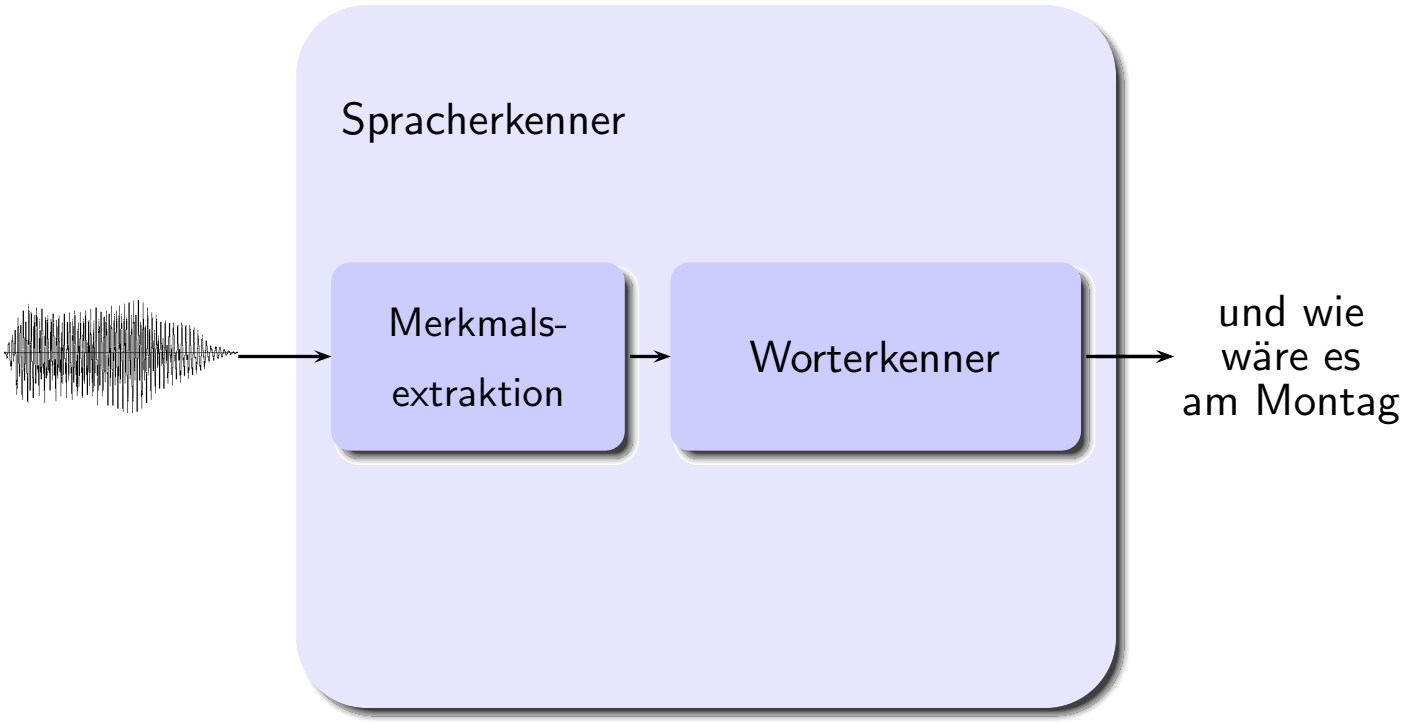


  - Segmentillusion: gesprochene Sprache wird als Sequenz von isolierten Stimuli wahrgenommen (Perlenkette)
  - frühe Idee: segmentiere das Signal und klassifiziere die Segmente in Phone bzw. artikulatorische Merkmale

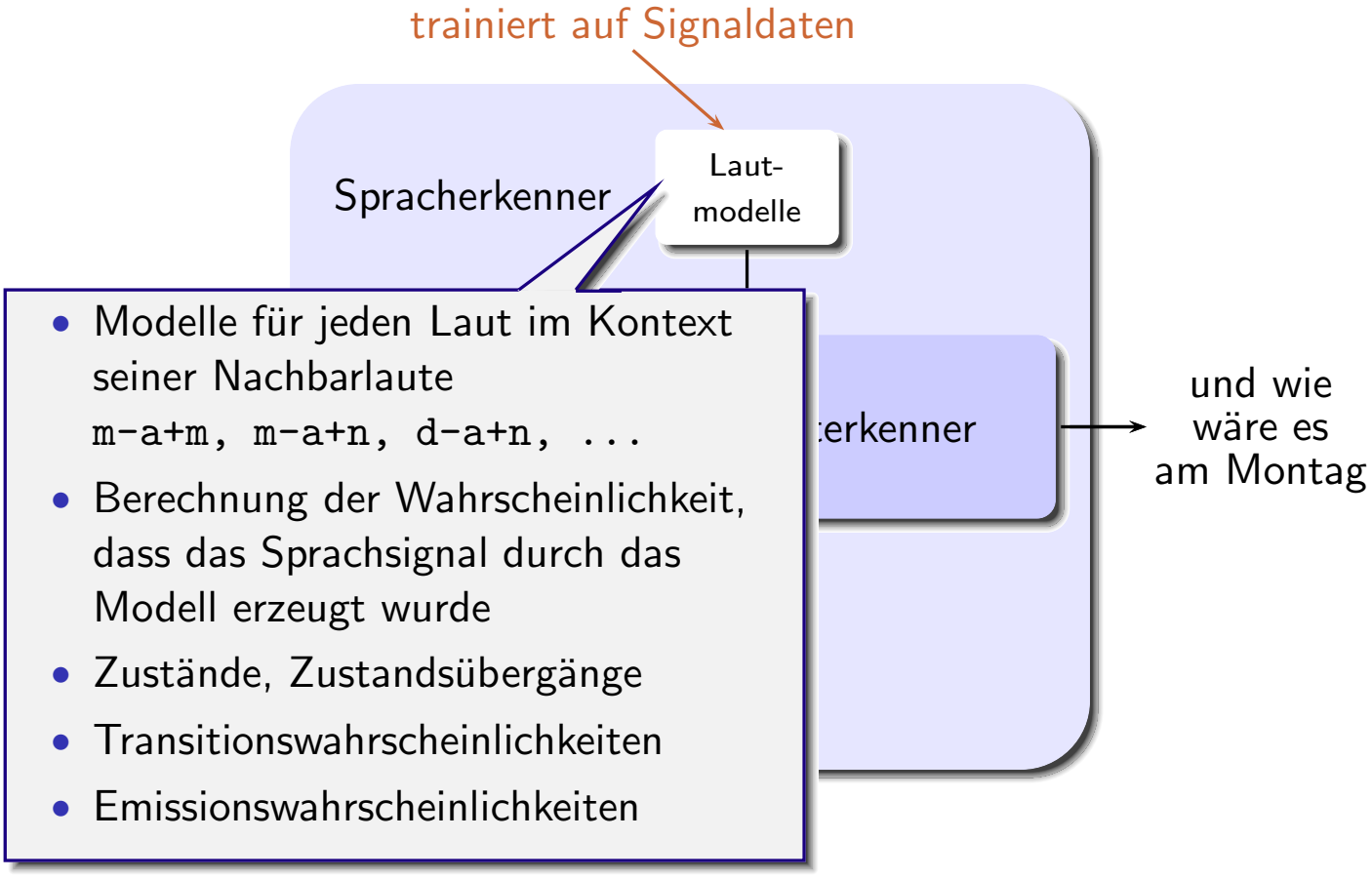
Grundlagen der Sprachsignalerkennung

- Spracherkennung als technisches und akademisches Problem
- Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung
- Merkmalsextraktion
- Worterkennung
- Phon-basierte Worterkennung
- Erkennung fließender Sprache
- Systemarchitekturen

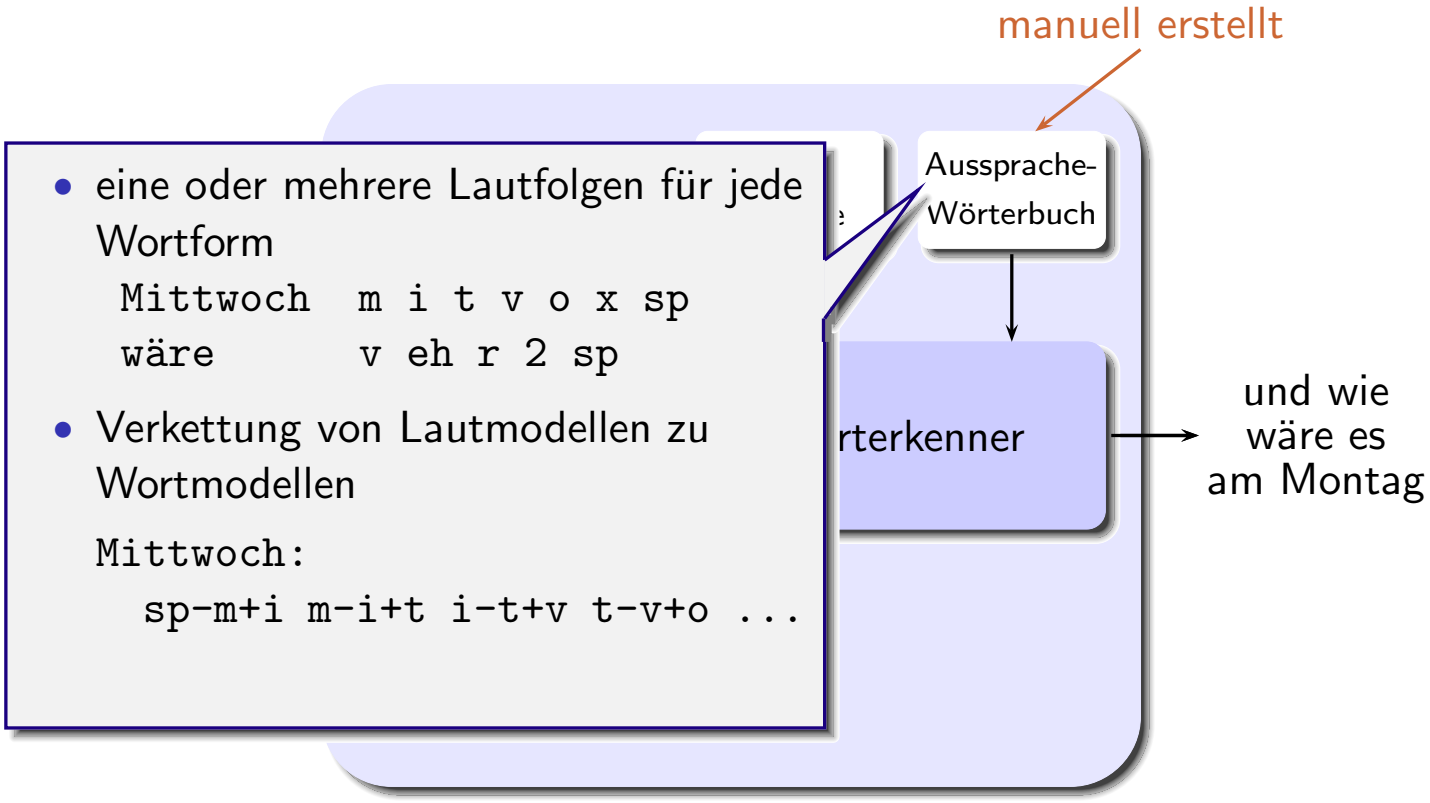
Überblick Spracherkennung



Überblick Spracherkennung



Überblick Spracherkennung



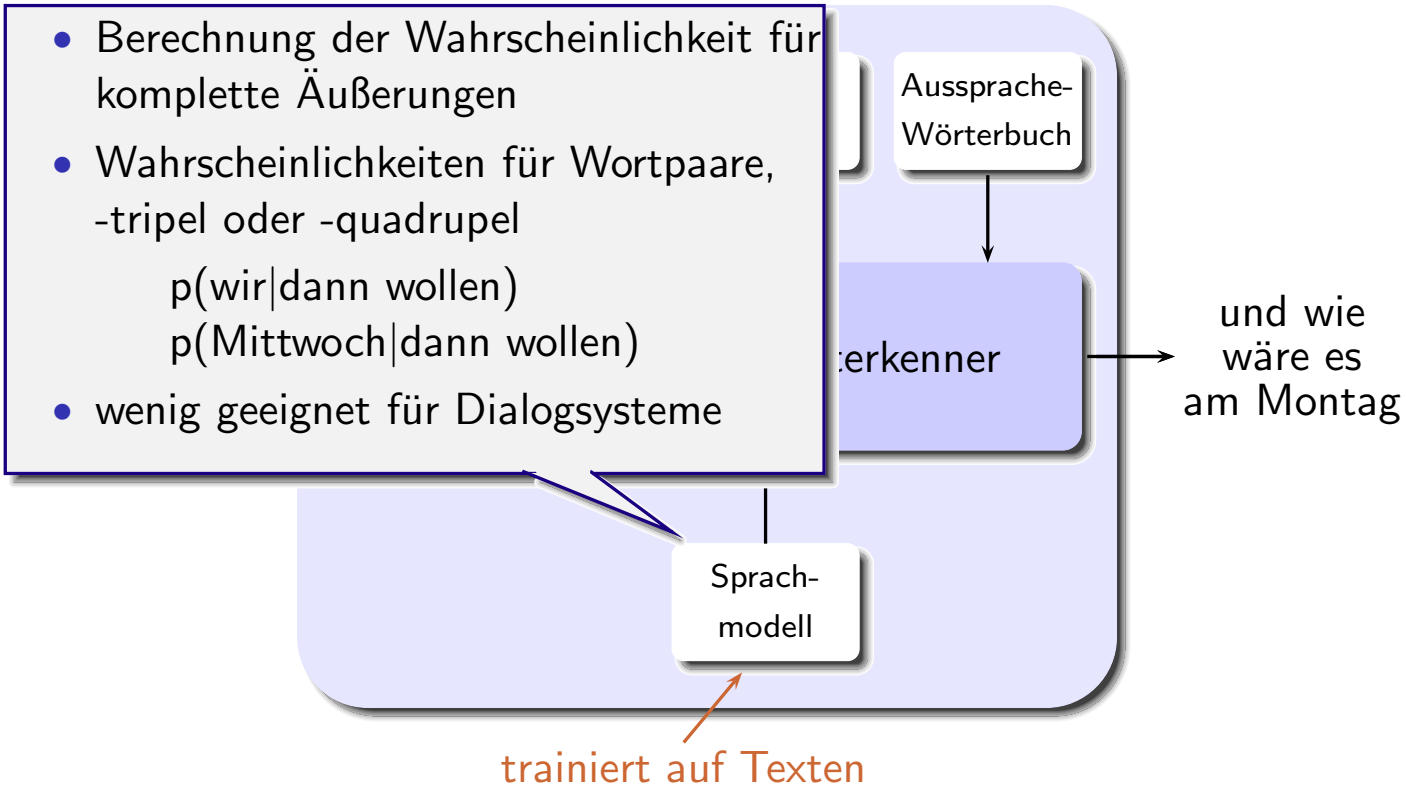
- eine oder mehrere Lautfolgen für jede Wortform

  ```
  Mittwoch   m i t v o x sp
  wäre       v eh r 2 sp
  ```

- Verkettung von Lautmodellen zu Wortmodellen

  ```
  Mittwoch:
    sp-m+i m-i+t i-t+v t-v+o ...
  ```

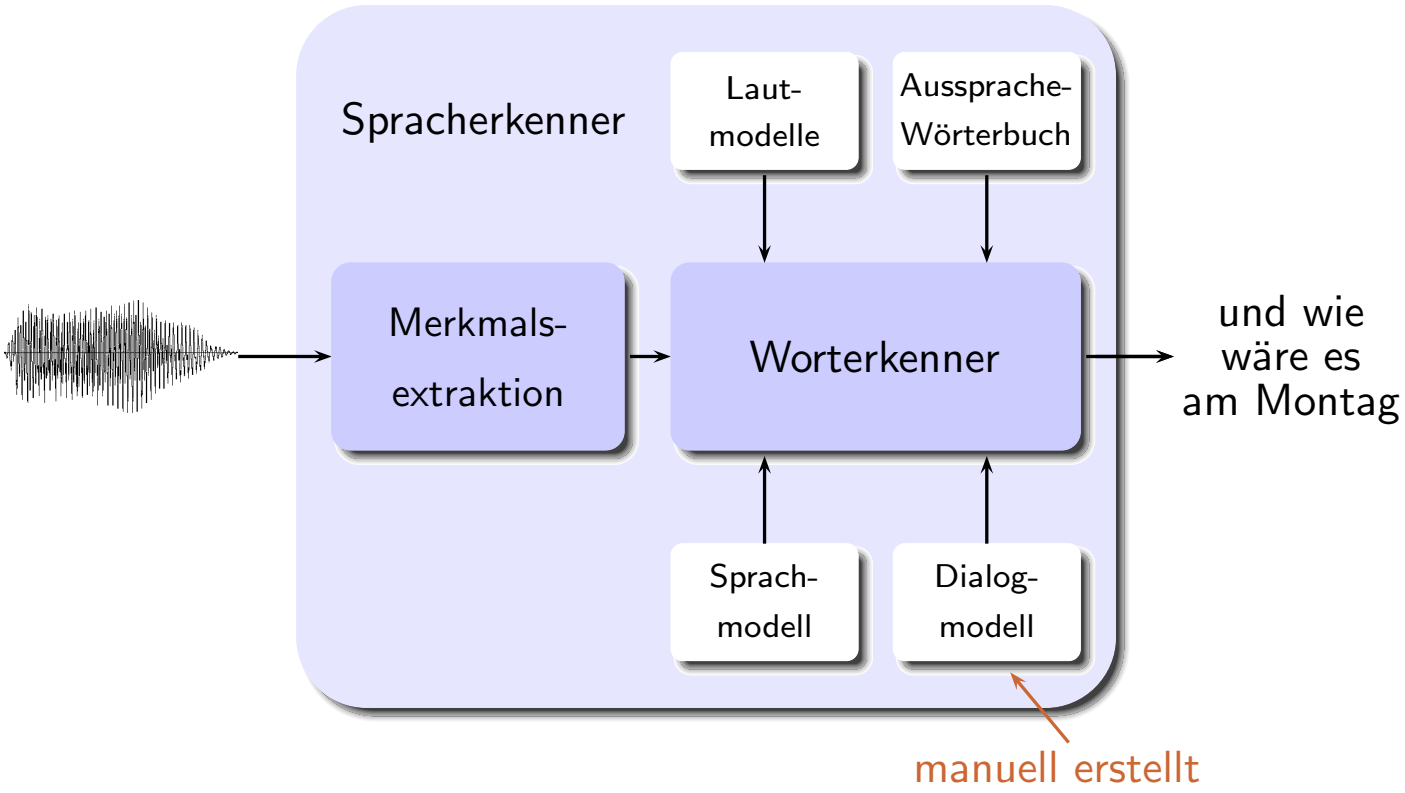

Überblick Spracherkennung



- Berechnung der Wahrscheinlichkeit für komplette Äußerungen
- Wahrscheinlichkeiten für Wortpaare, -tripel oder -quadrupel

  p(wir|dann wollen)
  p(Mittwoch|dann wollen)
- wenig geeignet für Dialogsysteme

Überblick Spracherkennung



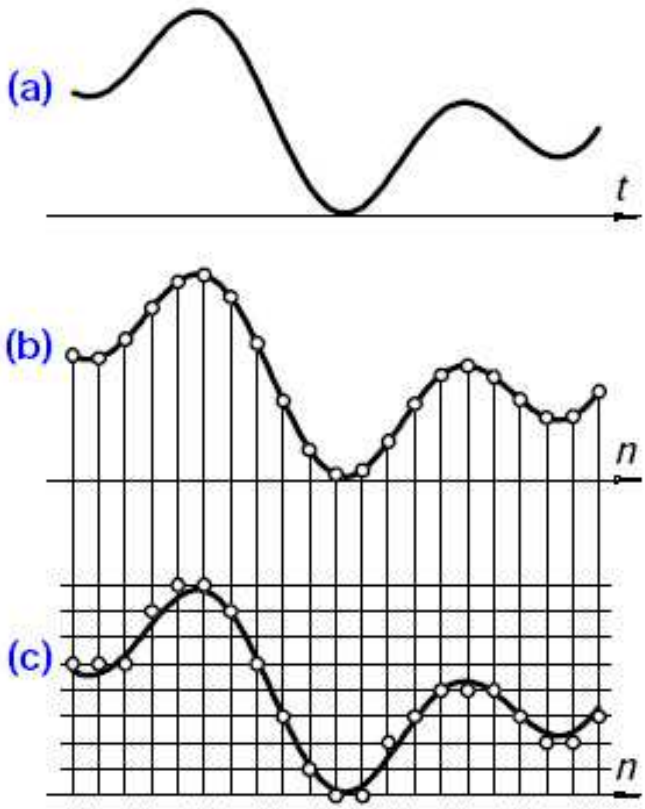
Merkmalsextraktion

- Signalwandlung
- Beschreibung im Zeitbereich
- Beschreibung im Frequenzbereich
- Cepstral-Analyse
- Linear Predictive Coding: LPC-Analyse
- Psychoakustisch inspirierte Techniken
- Gradientenmodellierung

Signalwandlung

- Mikrofon: Schalldruck → elektrische Spannung
- Funktionsprinzip
  - Kohlemikrofon (Telefon)
  - elektrodynamisches Mikrofon
  - elektrostatisches Mikrofon
- Übertragungsfaktor (Empfindlichkeit)
- Frequenzgang
- Linearität
- Richtcharakteristik
- stationär/mobil

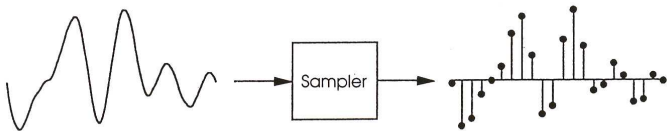
Digitalisierung



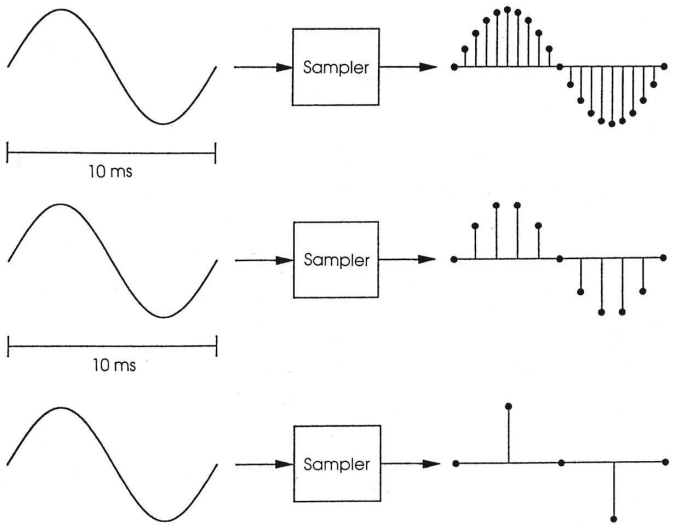
- Abtastung
- Quantisierung
- Kodierung

Abtastung

- Diskretisierung entlang der Zeitachse (Sampling)
  Multiplikation mit einer Folge von Nadelimpulsen

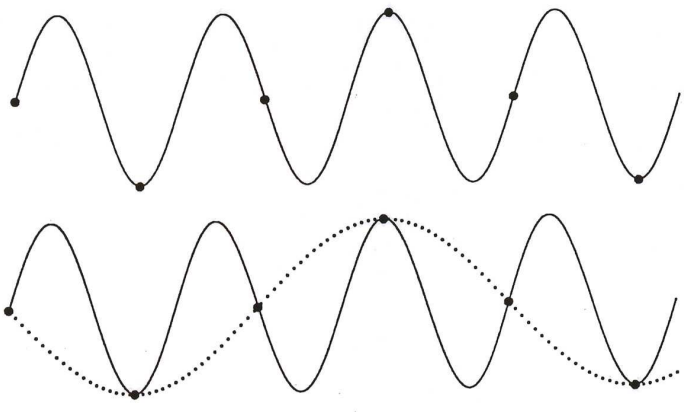


- Abtastrate



Abtastung

- Aliasbildung



- Abtasttheorem für bandbreitenbegrenzte Signale (Shannon und Weaver 1949)

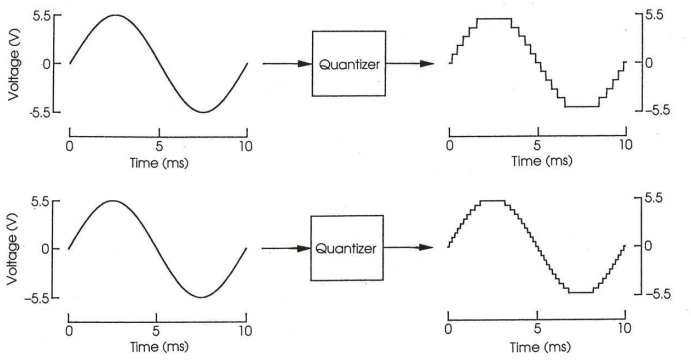
$$S = \frac{1}{T} = 2f_{max}$$

Abtastung

- typische Samplingraten
  - Telefon: 8 KHz
  - sonst: 16 ... 20 KHz

Quantisierung

- Diskretisierung entlang der Amplitudenachse



- linear / logarithmisch

Quantisierung

- Quantisierungsfehler
- Qualitätsmaß: Störabstand *SNR* (signal-to-noise ratio)
  - stationäres weißes Rauschen
  - Fehler ist unabhängig vom Inputsignal
  - Gleichverteilung für die Fehlerwahrscheinlichkeit in allen Quantisierungsintervallen

$$SNR = \frac{\sum_i {x_i}^2}{\sum_i (\hat{x}_i - x_i)^2}$$

x_i gemessener Inputwert
$\hat{x}_i$ quantisierter Inputwert

- typische Amplitudenauflösung: 6 ... 7 bit

Merkmalsextraktion

- Signalerfassung
- Beschreibung im Zeitbereich
- Beschreibung im Frequenzbereich
- Cepstral-Analyse
- Linear Predictive Coding: LPC-Analyse
- Psychoakustisch inspirierte Techniken
- Gradientenmodellierung

Signalbeschreibung im Zeitbereich

- Dilemma:
  - Sprache ist ein hochdynamischer Prozess
  - alle technischen Meßverfahren machen eine Annahme über die Stationarität

- Kurzzeitanalyse (Zeitfenster)
  - viele temporale Parameter können in kurzen Signalabschnitten (10 ... 30 ms) als konstant betrachtet werden

- Wahl der Fensterbreite
  - kurz: schlechte Glättung
  - lang: schlechte Reaktion auf plötzliche Änderungen im Signal

Signalbeschreibung im Zeitbereich

- Energie

$$E(n) = \sum_{m=0}^{M} x(n-m)^2 \qquad M: \text{ Fensterbreite}$$

  - stimmhaft: hohe Energie
  - stimmlos: niedrige Energie

- Nulldurchgangsdichte

$$Z(n) = \sum_{m=0}^{M} |\text{signum}(x(n-m)) - \text{signum}(x(n-m-1))|$$

  - stimmhaft: niedrige Nulldurchgangsdichte
  - stimmlos: hohe Nulldurchgangsdichte
  - simples Maß, für einfache Worterkenner bereits geeignet

Signalbeschreibung im Zeitbereich

- Autokorrelation
  - für deterministische diskrete Signale

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)\, x(m+k)$$

  - für zufällige und periodische Signale

$$\phi(k) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{m=-N}^{N} x(m)\, x(m+k)$$

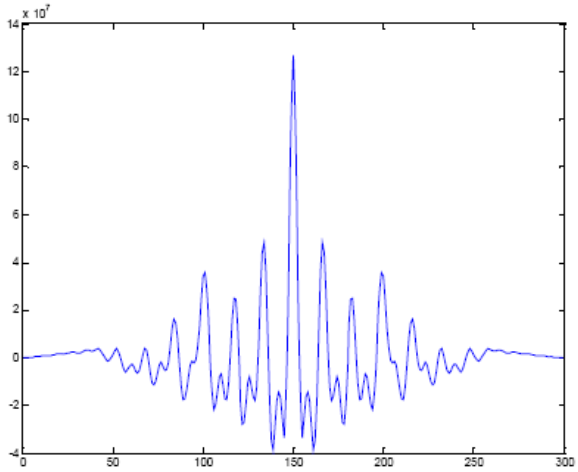
  - Kurzzeitautokorrelation

$$R(n,k) = \sum_{m=0}^{M} x(n-m)\, x(n-m-k)$$

  - die Autokorrelationsfunktion eines periodischen Signals ist ebenfalls periodisch und besitzt die gleiche Periode

Signalbeschreibung im Zeitbereich

- Nachweis der Periodizität durch Autokorrelation



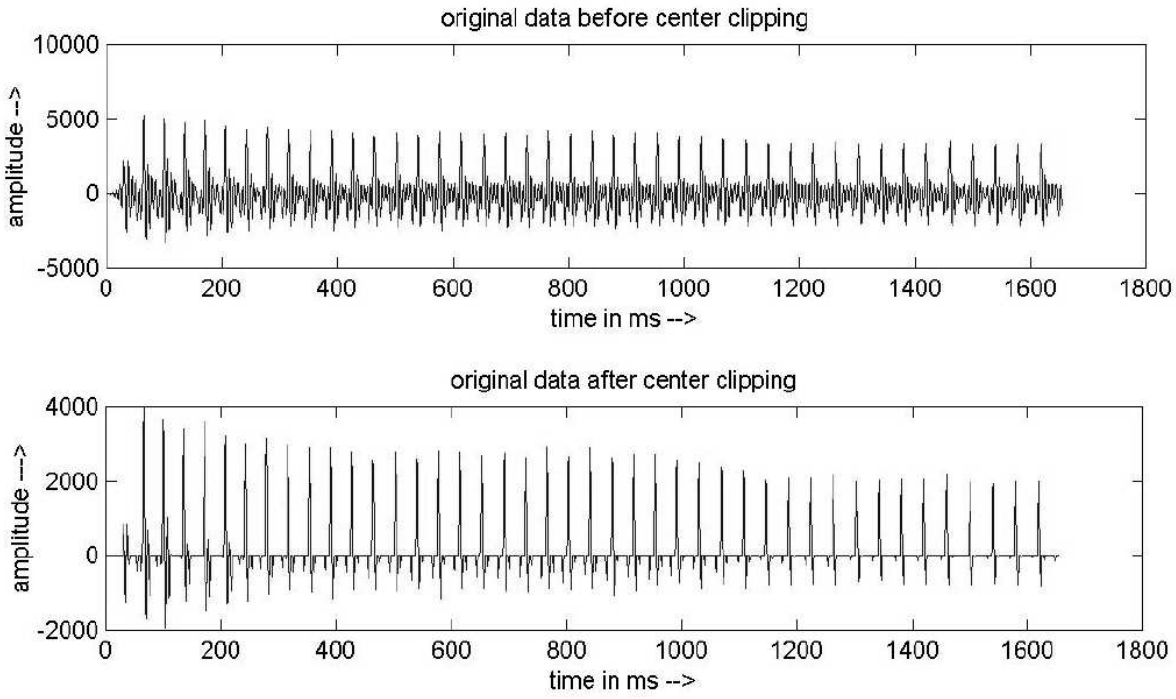
ein Frame des Vokals /i/ in *six*

Beauchaine (2003)

→ stimmhaft/stimmlos-Detektor

Signalbeschreibung im Zeitbereich

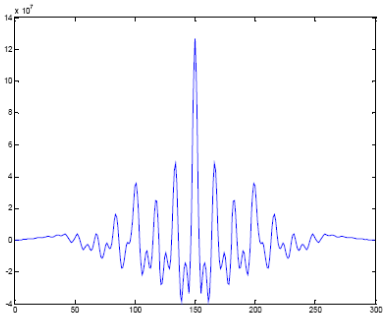
- höhere Detektionsschärfe durch Center Clipping

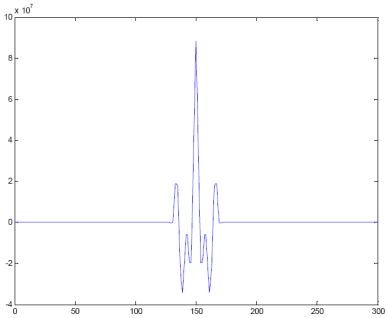


KAVITA (1999)

Signalbeschreibung im Zeitbereich

- Autokorrelation nach center clipping





Merkmalsextraktion

- Signalerfassung
- Beschreibung im Zeitbereich
- Beschreibung im Frequenzbereich
- Cepstral-Analyse
- Linear Predictive Coding: LPC-Analyse
- Psychoakustisch inspirierte Techniken
- Gradientenmodellierung

Signalbeschreibung im Frequenzbereich

- Transformationen für kontinuierliche und diskrete Zeitfunktionen

	kontinuierlich	diskret
beliebige Funktionen	Laplace-Transform.	z-Transformation
zeitbegrenzte oder periodische Funkt.	Fourier-Transform.	diskrete Fourier-Transform.

z-Transformation

- Transformation

$$X(z) = \mathcal{Z}(x(n)) = \sum_{n=-\infty}^{\infty} x(n)\, z^{-n}$$

z	komplexe Variable
$X(z)$	komplexe Funktion

  - konvergiert nur, wenn $\sum_{n=-\infty}^{\infty} |x(n)|\, |z^{-n}| < \infty$
  - Konvergenzbedingung hat immer die Form: $R_1 < |z| < R_2$
- Rücktransformation

$$x(n) = \mathcal{Z}^{-1}(X(z)) = \frac{1}{j2\pi} \oint_C X(z)\, z^{n-1} dz$$

  Kontur C im Konvergenzgebiet von X(z) (Einheitskreis)
- z-Transformierte ist eine komplexe Funktion in einer komplexen Ebene

z-Transformation

- Spezialfall: Nadelimpuls

$$\delta(n) = \begin{cases} 1 & n = 0 \\ 0 & \text{sonst} \end{cases}$$

$$X(z) = \sum_{n=-\infty}^{\infty} \delta(n)\, z^{-n} = 1$$

- Impulsantwort eines linearen Systems enthält die vollständige Information über seinen Frequenzgang

Diskrete Fourier-Transformation

- Spezialfall der z-Transformation für $z = e^{j\omega}$
- komplexe Funktion über dem Einheitskreis in einer komplexen Ebene
- Transformation

$$X(e^{j\omega}) = \mathcal{F}(x(n)) = \sum_{n=-\infty}^{\infty} x(n)\, e^{-j\omega n}$$

- Rücktransformation

$$x(n) = \mathcal{F}^{-1}(X(e^{j\omega})) = \int_{-\pi}^{\pi} X(e^{j\omega})\, e^{j\omega n} d\omega$$

- ω ist Winkel zur rellen Achse
- Konvergenzbereich: $\sum_{n=-\infty}^{\infty} |x(n)| < \infty$
- Spektrum $X(e^{j\omega})$ ist eine periodische Funktion von ω
  - Periode 2π
  - Nachweis durch Ersetzen von ω durch $\omega + 2\pi$

Kurzzeitanalyse

- z- und Fourier-Transformation sind nur für unendliche Zeitfunktionen definiert
  auch negative Zeiten!
  - Spektralanalyse über ganze Äußerungen ?
  - nur für stationäre Signalabschnitte sinnvoll
    sonst werden alle temporären Vorgänge "herausgemittelt"
  - Sprache ist ein dynamischer Prozeß
  - Ziel: spektrale Charakterisierung relativ kurzer Signalabschnitte

- Analysezeitfenster

$$w(n) \begin{cases} > 0 & n_0 \le n < n_0 + N \\ = 0 & \text{sonst} \end{cases}$$

- typische Fensterbreiten: 10 ... 20 ms (überlappend)

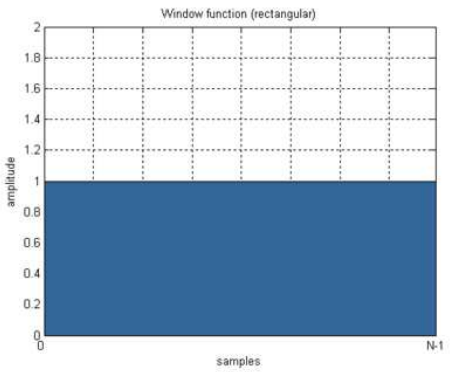
Kurzzeitanalyse

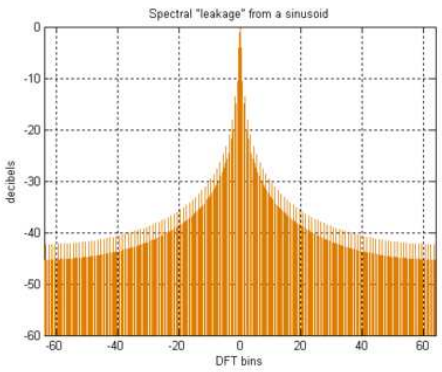
- zeitabhängige Fourier-Transformation
  - Bewegen des Zeitfensters über die Zeitachse
  - Multiplikation des Eingangssignals mit der Fensterfunktion $w(m)$
  - Frequenzcharakteristik des Zeitfensters geht mit in das Transformationsergebnis ein
  - unerwünschte "Filterung" des Sprachsignals

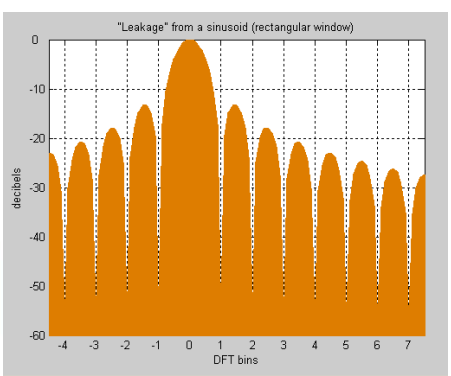
$$
\begin{aligned}
X(n, e^{j\omega}) &= \mathcal{WF}(x(n)) \\
&= \sum_{m=-\infty}^{\infty} w(n-m)\, x(m)\, e^{-j\omega m} \\
x(n) &= \mathcal{WF}^{-1}(n, X(e^{j\omega})) \\
&= \frac{1}{2\pi w(0)} \int_{-\pi}^{\pi} X(n, e^{j\omega})\, e^{j\omega n} d\omega
\end{aligned}
$$

Kurzzeitanalyse

- Fenster sind nichtlineare zeitliche Filter
  - Einfluss auf das Spektrum ist ebenfalls nichtlinear
  - leakage: Signalenergie wird auf andere Frequenzen verteilt
  - zwei Sinusschwingungen können ununterscheidbar werden
  - Fensterform kann die Wirkung abschwächen aber nicht vermeiden
- maximale Signaldeformation: Rechteckfenster

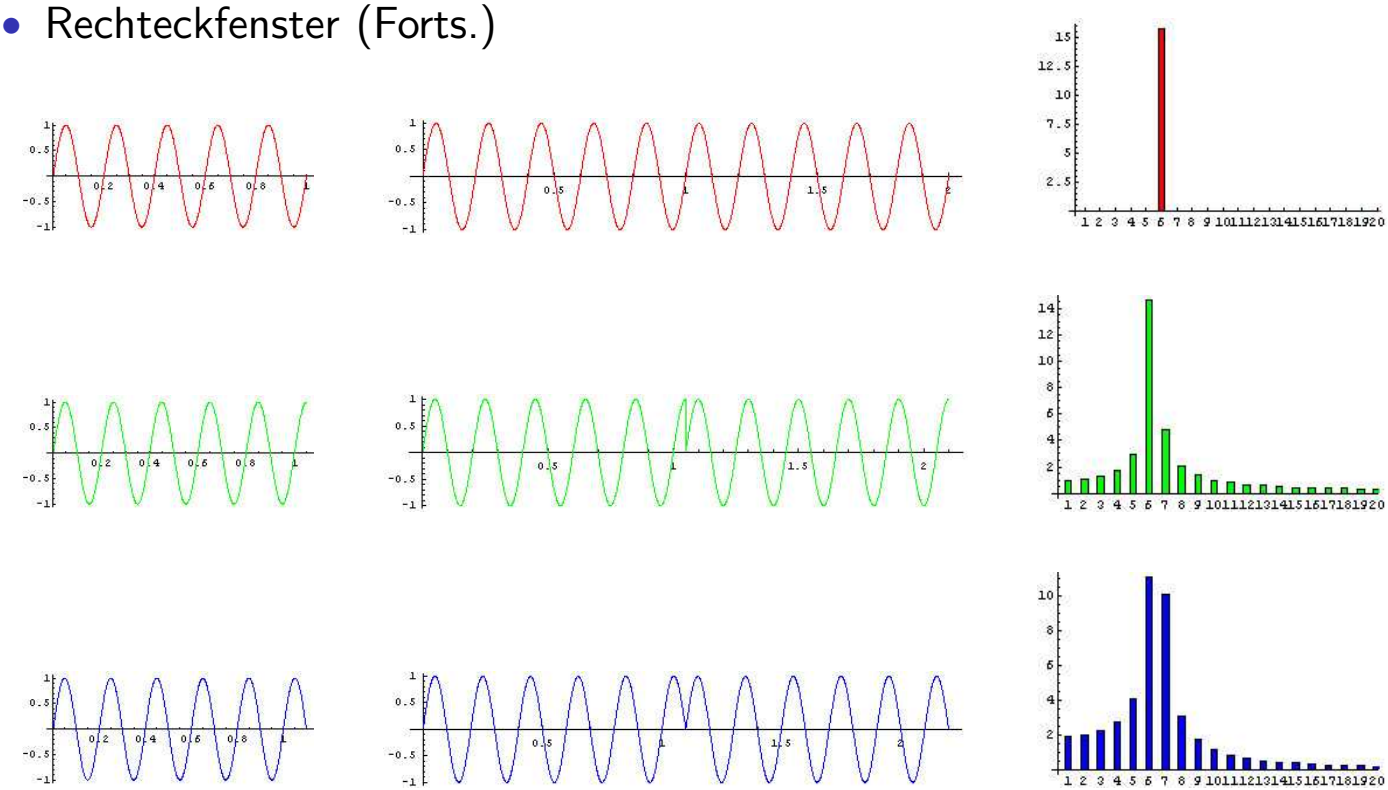






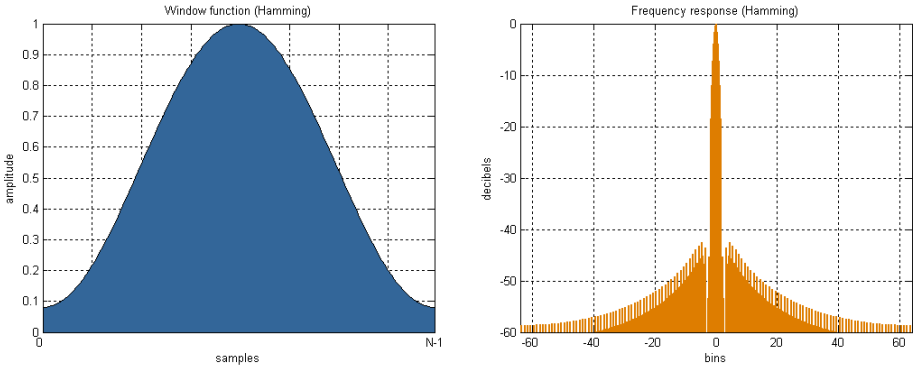
Kurzzeitanalyse

- Rechteckfenster (Forts.)

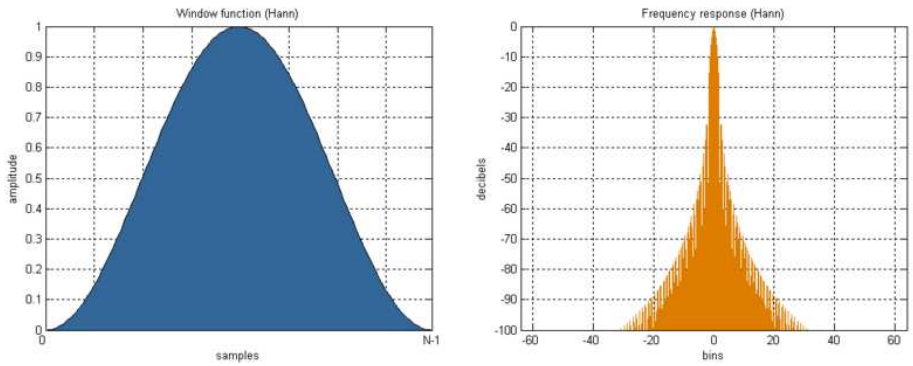


Kurzzeitanalyse

- Hamming-Fenster



- von Hann -Fenster ("Hanning")



Kurzzeitanalyse

- Unschärferelation
  - die Breite des Zeit- und des Frequenzfensters verhalten sich indirekt proportional zueinander

$$\Delta f = \frac{1}{\Delta t}$$

  - gute Zeitauflösung $\rightsquigarrow$ schlechte Frequenzauflösung
  - gute Frequenzauflösung $\rightsquigarrow$ schlechte Zeitauflösung

Diskrete Kurzzeit-Fourier-Transformation

- Annahme: Inputsignal ist periodische Fortsetzung des Fensterausschnittes

$$x(n) = x(n+N)$$

- Transformation

$$X(k) = \mathcal{KF}(x(n)) = \sum_{n=0}^{N-1} x(n)\ e^{-j\frac{2\pi}{N}kn}$$

$$0 \leq k < N$$

- Rücktransformation

$$x(n) = \mathcal{KF}^{-1}(X(k)) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)\ e^{j\frac{2\pi}{N}kn}$$

$$0 \leq n < N$$

Diskrete Kurzzeit-Fourier-Transformation

- Fast-Fourier-Transformation (FFT)
  - effiziente rechentechnische Realisierung der diskreten Kurzzeit-Fourier-Transformation
- Simulation von Bandpaß-Filterbänken durch FFT
  - typische Anzahl von Filtern: 20 ... 24

Merkmalsextraktion

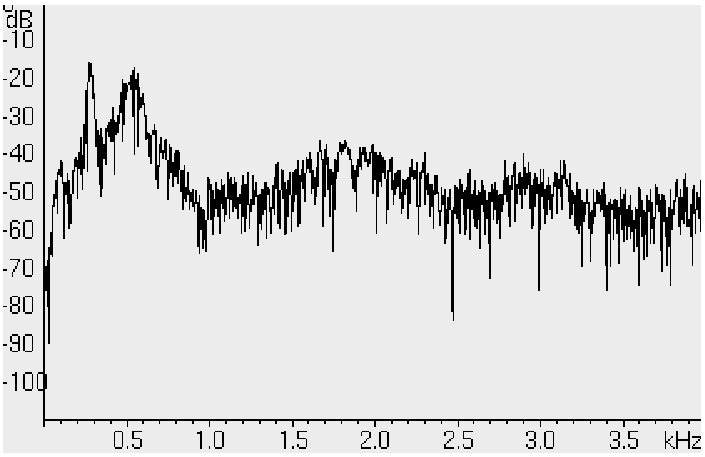
- Signalerfassung
- Beschreibung im Zeitbereich
- Beschreibung im Frequenzbereich
- Cepstral-Analyse
- Linear Predictive Coding: LPC-Analyse
- Psychoakustisch inspirierte Techniken
- Gradientenmodellierung

Cepstral-Analyse (Homomorphe Analyse)

- Sprachproduktion ist Faltung von Anregung und Impulsantwort des Vokaltraktes

$$x(n) = a(n) \otimes v(n)$$

- daher besitzt Spektrum für stimmhafte Abschnitte noch spektrale Feinstruktur



  - Spektrale Hülle: Filtercharakteristik des Vokaltraktes
  - Feinstruktur: Anregungscharakteristik

Cepstral-Analyse

- Ziele:
  - Eliminieren der Anregungscharakteristik ("Glätten" des Spektrums)
  - Reduktion der Merkmalsdimensionalität
    - $\approx$ 20 ... 24 Frequenzgruppenfilter
    - $\approx$ 6 ... 10 Cepstralkoeffizienten

Cepstral-Analyse

- Idee: Falten ist Multiplikation im Frequenzbereich

$$X(k) = A(k)\ V(k)$$

  - Logarithmus eines Produktes ist die Summe der Logarithmen der Faktoren

$$\begin{aligned} \hat{X}(k) &= \log(X(k)) \\ &= \log(A(k)\ V(k)) \\ &= \log(A(k)) + \log(V(k)) \end{aligned}$$

  - Rückführung der Multiplikation auf eine Addition

Cepstral-Analyse

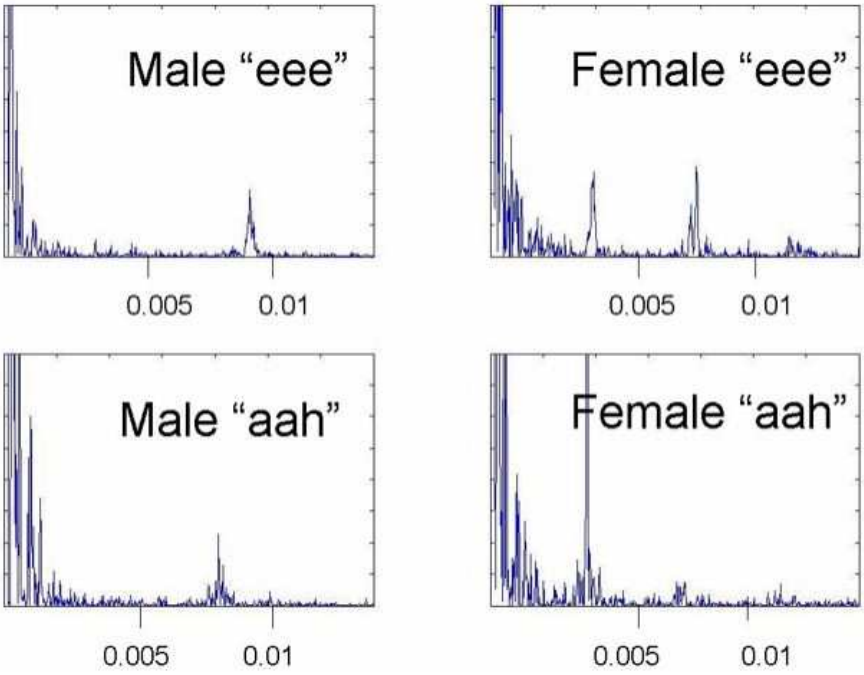
- Cepstrum ist Abbildung in einen nichtlinear transformierten Zeitbereich ("Quefrency")

$$C(m) = \mathcal{F}^{-1}(\hat{X}(k)) = \mathcal{F}^{-1}(\log(\mathcal{F}(x(n))))$$

- drei Schritte
  - Fourier-Transformation
  - Logarithmieren
  - inverse Fourier-Transformation

Cepstral-Analyse

- Anregungs- und Vokaltraktanteil liegen im Cepstrum weit auseinander und können daher einfach getrennt werden ("liftering")



Brian van Osdol

- noch schärfere Ausprägung der Grundfrequenzspitze durch Quadrieren des Cepstrums

Merkmalsextraktion

- Signalerfassung
- Beschreibung im Zeitbereich
- Beschreibung im Frequenzbereich
- Cepstral-Analyse
- Linear Predictive Coding: LPC-Analyse
- Psychoakustisch inspirierte Techniken
- Gradientenmodellierung

Linear Predictive Coding: LPC-Analyse

- autoregressives Modell: Vorhersage des aktuellen Abtastwertes $\tilde{x}(n)$ aufgrund der vorangegangenen K Abtastwerte
  - Linearkombination

$$\tilde{x}(n) = \sum_{k=1}^{K} \alpha_k \, x(n-k)$$

  - zugehörige Systemfunktion (komplexes Allpol-Modell)

$$P(z) = \sum_{k=1}^{K} \alpha_k \, z^{-k}$$

  - Vorhersagefehler

$$e(n) = x(n) - \tilde{x}(n) = x(n) - \sum_{k=1}^{K} \alpha_k \, x(n-k)$$

LPC-Analyse

- Fehlersequenz ist Ausgabe eines Systems mit der Systemfunktion

$$A(z) = 1 - \sum_{k=1}^{K} \alpha_k \, z^{-k}$$

- entspricht unter der Annahme eines idealisierten Source-Filter-Models genau der Übertragungsfunktion des Vokaltraktes $V(z)$ (inklusive Abstrahlungscharakteristik)!

LPC-Hypothese

- Falls der mittlere quadratische Fehler minimal ist, dann kann *angenommen* werden, dass die geschätzten Parameter des Prädiktors $P(z)$ genau den *Spektralkoeffizienten* von $V(z)$ entsprechen.
  - intuitive Stützung:
    länger zurückliegende Signalabschnitte werden vorrangig durch die tieferfrequenten Spektralanteile beeinflusst

LPC-Analyse

- Finde einen Parametersatz $\alpha_1 ... \alpha_K$, der den quadratischen Fehler $e(n)^2$ minimiert ($K \approx s[kHz] + 4 \approx 10...15$)
- Internes Modell der Sprachproduktion wird kontinuierlich an die Signalrealität adaptiert.
- Modellparameter dienen als (spektrale) Signalbeschreibung $\rightarrow$ Motor-Theory

- verschiedene Schätzverfahren
  - Kovarianzmethode
  - Autokorrelationsmethode
  - Lattice-Methode

- LPC-Analyse erlaubt zusätzlich auch die Ermittlung der Anregungscharakteristik aus dem Fehlerresiduum $e(n)$

- auch Cepstral-Transformation für LPC-Koeffizienten

Merkmalsextraktion

- Signalerfassung
- Beschreibung im Zeitbereich
- Beschreibung im Frequenzbereich
- Cepstral-Analyse
- Linear Predictive Coding: LPC-Analyse
- Psychoakustisch inspirierte Techniken
- Gradientenmodellierung

Gehörbezogene Parameter

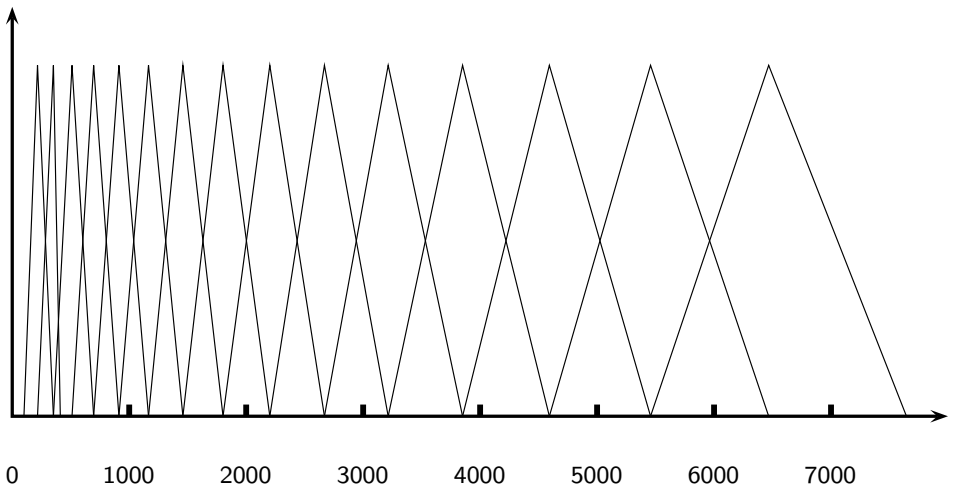
- Transformation der Spektralparameter
  - lautheitsbasierte Filterung
  - tonheitsbezogene Verzerrung der Frequenzachse
    mel-Skala: logarithmische Approximation der Bark-Skala:

$$mel(f) = 2595 log(1 + \frac{f}{700})$$

  - tonheitsbezogene Wichtung der Spektralparameter

Gehörbezogene Parameter

- z.B. Filterung des Signals mit einer äquidistanten mel-Skalen-Filterbank



- mel-Cepstrum: Berechnung der Cepstralkoeffizienten auf der Basis der Parameter einer mel-Skalen-Filterbank

Perzeptive LP-Analyse

- perzeptuelle Distanz zwischen Formanten ist viel kleiner als die Cepstral-Distanz der Maxima
  - Perzeption ist durch den Frontalbereich des Artikulationstrakts dominiert
  - Kindersprache ist verständlich, trotz erheblicher Abweichungen in der Spektralcharakteristik
  - zweiter Formant
    - enthält die meiste sprecherneutrale Information
    - korrespondiert mit dem Bereich der größten Empfindlichkeit der akustischen Wahrnehmung

Perzeptive LP-Analyse

- starke integrative Wahrnehmung beim Sprachverstehen
  - Integration über 3.5 Bark
  - nur grobe Form des Spektrums wird unterschieden
  - Newton (1665): Vokalsynthese durch Anblasen einer unterschiedlich weit gefüllten Bierflasche
  - Helmholtz (1885): Vokale mit einem (/a/,/o/,/u/) bzw. zwei Maxima (/e/,/i/,/y/)
  - Chistovich (1985): zwei Formanten werden perzeptuell zusammengefasst, wenn sie weniger als 3.5 Bark auseinanderliegen

Perzeptive LP-Analyse

- PLP: Perceptive Linear Prediction analysis (Hermansky 1991)
- Grundlagen
  - Bandbreitenwahl in Übereinstimmung mit der Frequenzgruppenwahrnehmung
  - Lautheitsbezogene Wichtung der Frequenzanteile
  - Berücksichtigung der Intensitäts-Lautheits-Beziehung

Perzeptive LP-Analyse

- Teilschritte
  - Spektraltransformation
  - Verzerren der Frequenzachse entsprechend der Bark-Skala
  - Faltung mit der Frequenzgruppenkurve und Abtastung
  - Korrektur der Intensität auf gleiche Lautheit
  - Verzerrung der Intensität entsprechend einer linearen Lautheitswahrnehmung
- Allpol-Modell 5. Grades ist optimal
  - max. 2 Formanten darstellbar!
- starke Korrelation von F_2 für PLP und Perzeption

Perzeptive LP-Analyse

- Vorteile
  - Approximation des perzeptuell wichtigen, sprecherinvarianten 2. Formanten
  - geringere Wichtung der hohen Frequenzen
  - Verringerung der Ungleichheit zwischen stimmhaften und stimmlosen Lautabschnitten
  - relativ gute Unabhängigkeit von der Länge des Vokaltrakts
  - größere Empfindlichkeit für Änderungen von F_1 und F_2

RASTA-Analyse

- Ziel: Reduktion der Abhängigkeit von der Spektralcharakteristik des Übertragungskanals (Mikrofon, ...)
- Subtraktion des spektralen Langzeitmittels
  - aber: in Echtzeitanwendungen schwierig zu ermitteln
- RASTA (Hermansky und Morgan 1994): Auswertung relativer spektraler Eigenschaften
  - Filtern langsamer Änderungen im Signal
- perzeptionspsychologische Fundierung:
  - Lautwahrnehmung verbleibt, auch wenn seine Formantstruktur eingeebnet wurde
  - Vokalperzeption wird durch die spektrale Differenz zum vorangegangenen Laut determiniert

RASTA-Analyse

- RASTA-PLP: vor Schätzung der Allpol-Parameter:
  - nichtlinare Verzerrung (logarithmisch oder linear-logarithmisch)
  - Herausfiltern aller konstanten bzw. niedrigfrequenten Änderungen der Ausgangswerte der Frequenzgruppenfilterbank
  - inverse Verzerrung
  - Verwendung von Filtern mit hohem zeitlichen Integrationsintervall
- höhere Robustheit gegenüber
  - Änderungen in der Übertragungscharakteristik (Faltungsstörungen)
  - Nebengeräuschen (additive Störungen)
- starke Abhängigkeit vom vorangegangenen Signalkontext
  - reduzierte Erkennungsgenauigkeit bei kurzen, kontextunabhängigen Lautmodellen

Innenohrmodelle

Merkmalsextraktion

- Signalerfassung
- Beschreibung im Zeitbereich
- Beschreibung im Frequenzbereich
- Cepstral-Analyse
- Linear Predictive Coding: LPC-Analyse
- Psychoakustisch inspirierte Techniken
- Gradientenmodellierung

Gradientenmodellierung

- Δ-Koeffizienten: Anstieg des Parameterverlaufs zum Zeitpunkt t
- ideal:

$$\frac{dx(n,t)}{dt}$$

- Approximation 1: Regressionskoeffizienten

$$\frac{dx(n,t)}{dt} \simeq \Delta x(n,t) = \frac{\sum_{k=-K}^{K} k\, w(k)\, x(n,t+k)}{\sum_{k=-K}^{K} k^2\, w(k)}$$

  - Fenster $w(k)$, meist Rechteckfenster

$$w(k) = 1 \text{ für } -K \leq k \leq K$$

  - typische Fensterbreite: 50 ms

Gradientenmodellierung

- Approximation 2: komponentenweise Differenzbildung zwischen benachbarten Merkmalvektoren

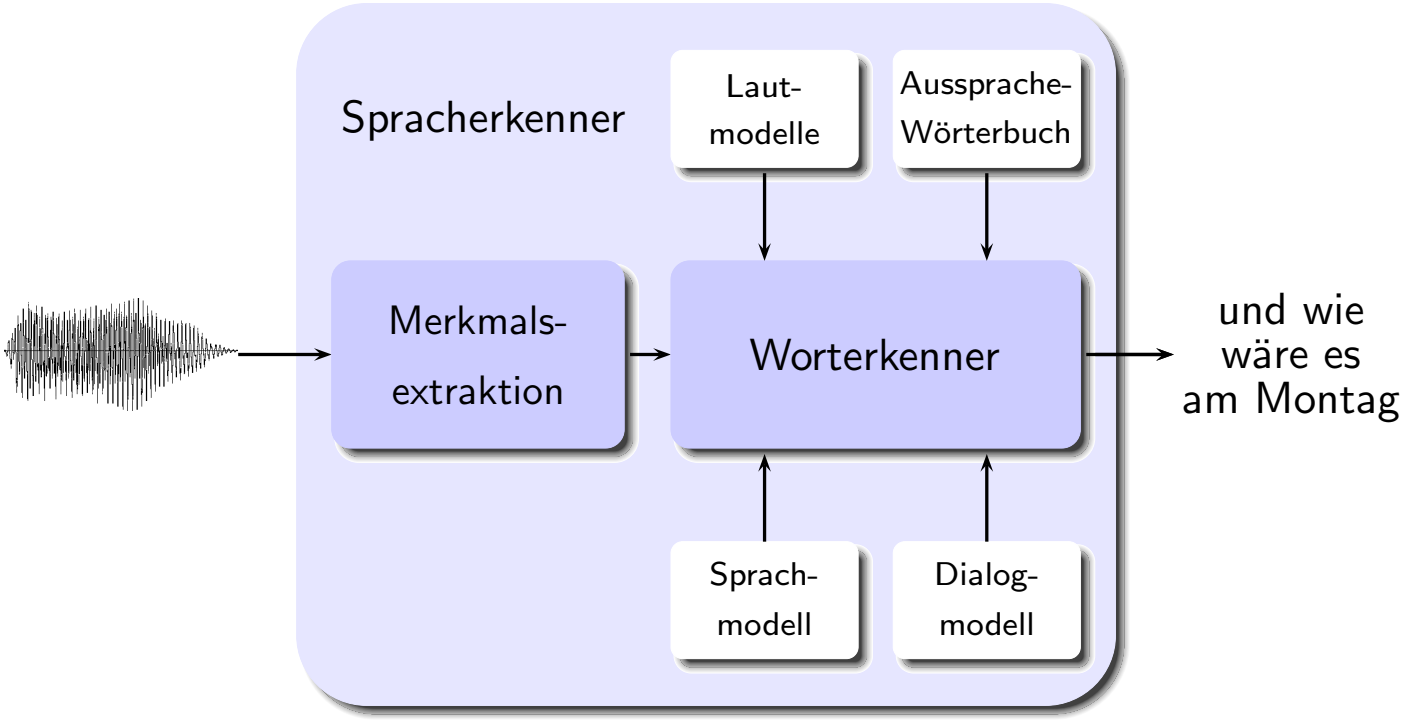
$$\frac{dx(n,t)}{dt} \simeq \Delta x(n,t) = x(n,t) - x(n,t-1)$$

- auch Δ^2-Parameter

Grundlagen der Sprachsignalerkennung

- Spracherkennung als technisches und akademisches Problem
- Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung
- Merkmalsextraktion
- Worterkennung
- Phon-basierte Worterkennung
- Erkennung fließender Sprache
- Systemarchitekturen

Überblick Spracherkennung



Worterkennung

- Klassifikation von atomaren Objekten
- Wahrscheinlichkeitsverteilungen
- Vektorquantisierung
- Klassifikation von Objektsequenzen
- Nichtlineare Zeitnormierung
- Hidden-Markov-Modellierung
- Erkennung mit Hidden-Markov-Modellen
- Training von Hidden-Markov-Modellen
- Kontinuierliche Hidden-Markov-Modelle

Klassifikation von atomaren Objekten

- subsymbolische $\rightarrow$ symbolische Beschreibung
- Rückführung auf eine Extremwertentscheidung
  - Unterscheidungsfunktion $D(k, \vec{x})$

$$k(\vec{x}) = \arg\max_k D(k, \vec{x})$$

$$k(\vec{x}) = \arg\min_k D(k, \vec{x})$$

Klassifikation von atomaren Objekten

- Training der Modellparameter auf Datensammlungen
  - Lernstichprobe
  - Paare aus Merkmalsvektor und Klassenzugehörigkeit
  - Schätzung der Klassifikatorparameter
  - Optimierungskriterium
    - minimaler Erwartungswert für Fehlentscheidungen
    - bei sehr unausgewogenen Verteilungen aber problematisch

Instanzenbasierte Klassifikatoren

- Nearest-Neighbor-Klassifikator
  - Training: Vollständiges Abspeichern der Lernstichprobe

    $$S = \{S_i | i = 1, \ldots, I\}$$

  - jedes Stichprobenelement besteht aus Merkmalvektor $\vec{x_i}$ und seiner Klassenzugehörigkeit $k(s_i)$

    $$s_i = (\vec{x_i}, k(s_i)) \qquad i = 1, ..., I$$

  - Klassifikation:
    - Berechnung der Distanz zu allen Elementen der Lernstichprobe
      $d(\vec{x}, \vec{x_i}) \qquad i = 1, ..., I$
    - Entscheidung für die Klasse des Stichprobenelementes mit der minimalen Distanz
      $k(\vec{x}) = k(s_i)$ mit $i = \arg\min_i d(\vec{x}, \vec{x_i})$

Distanzmaße

- City-Block-Distanz

$$d(\vec{x}, \vec{y}) = \sum_{i=1}^{I} |x_i - y_i|$$

- Euklidische Metrik

$$d(\vec{x}, \vec{y}) = \sum_{i=1}^{I} (x_i - y_i)^2$$

- gewichtete Euklidische Metrik

$$d(\vec{x}, \vec{y}) = \sum_{i=1}^{I} w(i)(x_i - y_i)^2$$

- Gewichtsfunktion w(i)
  - inverse Varianz: günstig insbesondere für Cepstralkoeffizienten
  - peak-weighted distance: Betonung der spektralen Maxima

Distanzmaße

- LPC-basierte Abstandsmaße
  - Maximum-Likelihood-Distanz (Itakura-Saito-Distanz)
  - cosh-Maß (symmetrische Maximum-Likelihood-Distanz)
  - Vorhersage-Residuum
  - Rücktransformiertes Vorhersage-Residuum

Instanzenbasierte Klassifikatoren

- Potentialfunktionenklassifikator
  - Unterscheidungsfunktion als Summe von Potentialfunktionen über den Stichprobenwerten

$$D(k, \vec{x}) = \sum_{i=1}^{I} f_{pot}(i, k)$$

  - Wahl von f_{pot}
    - Maximum an der Stelle des Stichprobenelementes
    - monoton fallend mit wachsender Distanz zum Stichprobenelement
    - Nearest-Neighbor-Klassifikator ist ein Spezialfall
  - keine Generalisierung über der Lernstichprobe
  - Klassifikationsaufwand wächst mit der Stichprobengröße

Probabilistische Klassifikatoren

- Naiver BAYES-Klassifikator
- Annahme einer Verteilungsfunktion
- Training: Schätzen der Parameter einer statistischen Verteilung für alle Stichprobenelemente gleicher Klassenzugehörigkeit $p(x|k)$
- Entscheidung nach maximaler *a posteriori*-Wahrscheinlichkeit für die Klassenzugehörigkeit
- BAYES-Entscheidungsregel

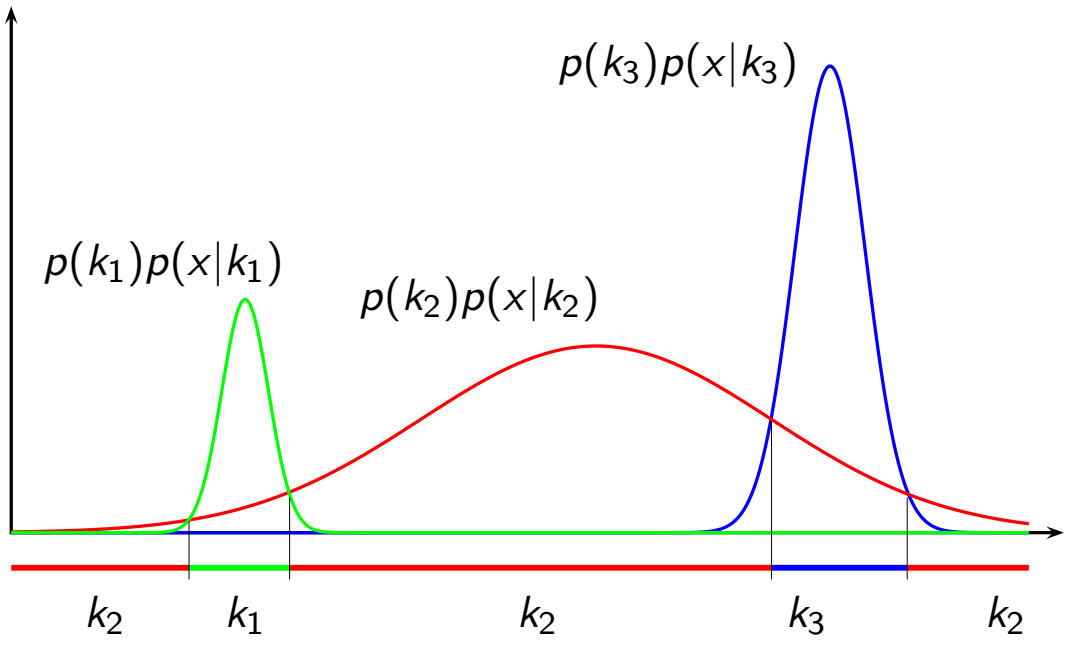
$$p(k|x) = \frac{p(k)\ p(x|k)}{p(x)}$$

x	Beobachtung
k	Klassenzugehörigkeit
$p(k)$	*a priori*-Wahrscheinlichkeit der Klassenzugehörigkeit

Probabilistische Klassifikatoren

- Annahme: $p(x)$ ist konstant für alle x

$$D(k, \vec{x}) = p(k|\vec{x}) = p(k)\ p(\vec{x}|k)$$



Probabilistische Klassifikatoren

- Klassifikatoren mit Rückweisung
  - zusätzliche neutrale Klasse
  - Senkung der Fehlerrate
  - Annahme: Rückweisung ist weniger problematisch als eine Fehlentscheidung

Probabilistische Klassifikatoren

- kostenbehaftete Klassifikation: Kostenfunktion c_{ki}
  - Kosten (penalty) für alle Entscheidungsmöglichkeiten des Klassifikators
  - Optimierung auf mittlere minimale Kosten
  - BAYES-Klassifikator: (0,1)-Kostenfunktion

$$c_{ki} = \begin{cases} 0 & \text{für } k = i \\ 1 & \text{für } k \neq i \end{cases}$$

  - Kostenfunktion für Klassifikatoren mit Rückweisung

$$c_{kk} < c_{0i} < c_{ki} \text{ mit } i, k = 1...K, k \neq i$$

$$c_{ki} = c_{ik}$$

Probabilistische Klassifikatoren

- Klassifikatoren mit unsymmetrischer Kostenfunktion
- unerwünschte Klassifikationsfehler werden mit besonders hohen Kosten belegt
  - Fehldiagnose beim medizinischen Screening
  - Fehlalarm in der Anlagenüberwachung

  $\longrightarrow$ geringere Kosten als ein irrtümlich nicht ausgelöster Alarm
- Ereignisdetektion für unterspezifizierte phonologische Beschreibungen

  $\longrightarrow$ fehlendes Merkmal kann eventuell durch Kontextinformation kompensiert werden (Prinzip der Informationsanreicherung)
- Detektion von Aussprachefehlern (Fremdsprachenlernen)
  - Fehler sind deutlich seltener als korrekte Lautrealisierungen
  - andererseits: unberechtigte Kritik ist problematisch; Ignorieren von Fehlern nicht so sehr
- auch für Worterkennung
  - domänenspezifische Relevanz von Wortformen

Probabilistische Klassifikatoren

- Unterscheidungsfunktion für Klassifikator mit Kostenfunktion

$$D(k, \vec{x}) = \sum_{i=1}^{K} (1 - c_{ki})\ p(i)\ p(\vec{x}|i)$$

$$k(\vec{x}) = \arg\max_{k} D(k, \vec{x})$$

Verteilungsfreie Klassifikatoren

- Approximation eines BAYES-Klassifikators durch einen äquivalenten Klassifikator mit (mathematisch) einfacherer Unterscheidungsfunktion
  - Linearklassifikator
  - stückweise-linearer Klassifikator
    (→ Perzeptron, Neuronale Netze)

Worterkennung

- Klassifikation von atomaren Objekten
- Wahrscheinlichkeitsverteilungen
- Vektorquantisierung
- Klassifikation von Objektsequenzen
- Nichtlineare Zeitnormierung
- Hidden-Markov-Modellierung
- Erkennung mit Hidden-Markov-Modellen
- Training von Hidden-Markov-Modellen
- Kontinuierliche Hidden-Markov-Modelle

Wahtscheinlichkeitsverteilungen

- diskrete Verteilungen
  - Verteilungsvektoren
  - Gleichverteilung
  - POISSON-Verteilung
- kontinuierliche Verteilungen
  - Gauss-Verteilung
  - Multivariate Gauss-Verteilung
  - Mischverteilungen

Diskrete Verteilungen

- Verteilungsvektoren (Aufzählung der Verteilungswerte)

  $$(p(x_1|k), p(x_2|k), \ldots, p(x_n|k))$$

  - Parameter: $p(x_1|k), p(x_2|k), \ldots, p(x_n|k)$
  - nur für endliche Mengen von Beobachtungen
  - auch für nichtnumerische Daten geeignet
  - hohe Parameterzahl, aber dennoch oft genutzt
- Gleichverteilung (für M Ereignisse)

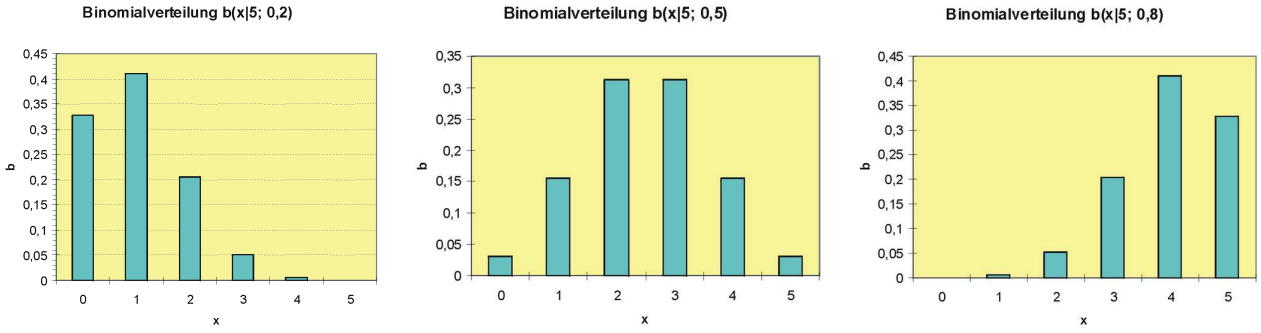
  $$p(x|k) = \frac{1}{M}$$

  - als Diskriminanzfunktion ungeeignet

Diskrete Verteilungen

- z.B. Binomialverteilung
  - 2-Klassen-Auswahl mit Zurücklegen
  - n: Anzahl der Versuche
  - k: Anzahl der ausgewählten Objekte einer Klasse

$$p(k) = \left(\begin{array}{c} n \\ k \end{array} \right) p(k)^k (1 - p(k))^{(n-k)}$$



- "Normalverteilung" für diskrete Ereignisse bei hinreichend großem n

Diskrete Verteilungen

- alternative Verteilungen
  - hypergeometrische Verteilung
  - Poisson-Verteilung
- erfordern eine numerische Ordnung auf den Beobachtungsdaten
- nur für Spezialfälle geeignet (z.B. Längenmodellierung)

Kontinuierliche Verteilungen

- Wahrscheinlichkeitsdichten statt Wahrscheinlichkeiten
- Normalverteilung (GAUSS-Verteilung, GAUSS'sche Glockenkurve)

$$p(x|k) = \mathcal{N}[x, \mu, \sigma] = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- LAPLACE-Verteilung

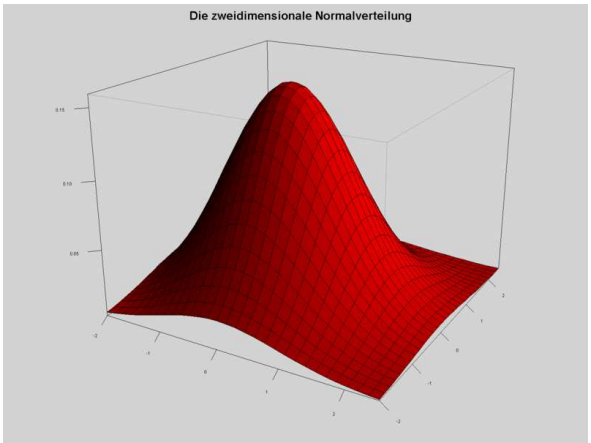
$$p(x|k) = \mathcal{L}[x, \mu, \sigma] = \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}}$$

- Parameter:
  - Mittelwert μ
  - Varianz σ

Kontinuierliche Verteilungen

- Verteilungen für mehrdimensionale Ereignisse
- z.B. multivariate Normalverteilung

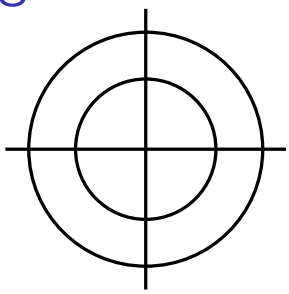
$$p(\vec{x}|k) = \mathcal{N}[\vec{x}, \vec{\mu}, \Sigma]$$



- Parameter:
  - Mittelwertvektor $\vec{\mu}$
  - Kovarianzmatrix Σ

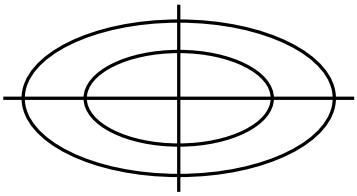
Kontinuierliche Verteilungen

diagonale Kovarianzmatrix
uniform besetzt
(rotationssymmetrisch zum Mittelpunktvektor)



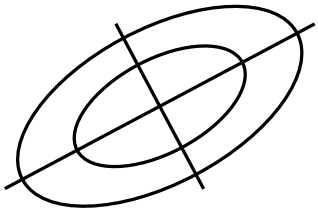
$$\sigma_{ij} = \begin{cases} n & \text{für } i = j \\ 0 & \text{sonst} \end{cases}$$

diagonale Kovarianzmatrix
beliebig besetzt
(axialsymmetrisch)



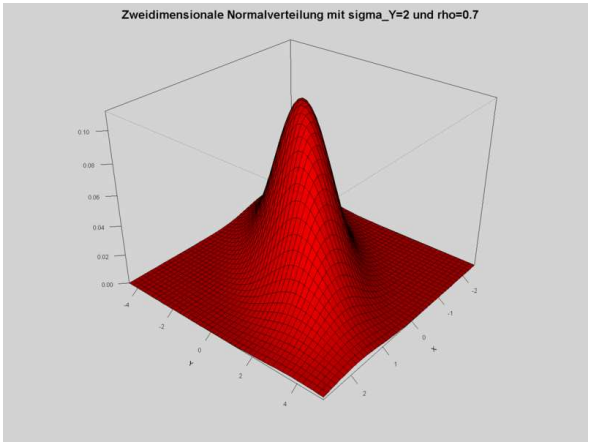
$$\sigma_{ij} = \begin{cases} n_i & \text{für } i = j \\ 0 & \text{sonst} \end{cases}$$

vollständig besetzte Kovarianzmatrix



Kontinuierliche Verteilungen

- diagonal besetzte Kovarianzmatrix: unkorrelierte Merkmale
  relativ geringe (zu trainierende) Parameterzahl
- voll besetzte Kovarianzmatrix: korrelierte Merkmale
  hohe (zu trainierende) Parameterzahl



- Dekorrelation der Merkmale: Transformation des Koordinatensystems
  Principal Component Analysis, Karhunen-Loève-Transformation

Kontinuierliche Verteilungen

- Abstandsklassifikator ist Spezialfall eines BAYES-Klassifikators mit
  - rotationssymmetrischer multivariater Normalverteilung und
  - gleichen a priori Wahrscheinlichkeiten

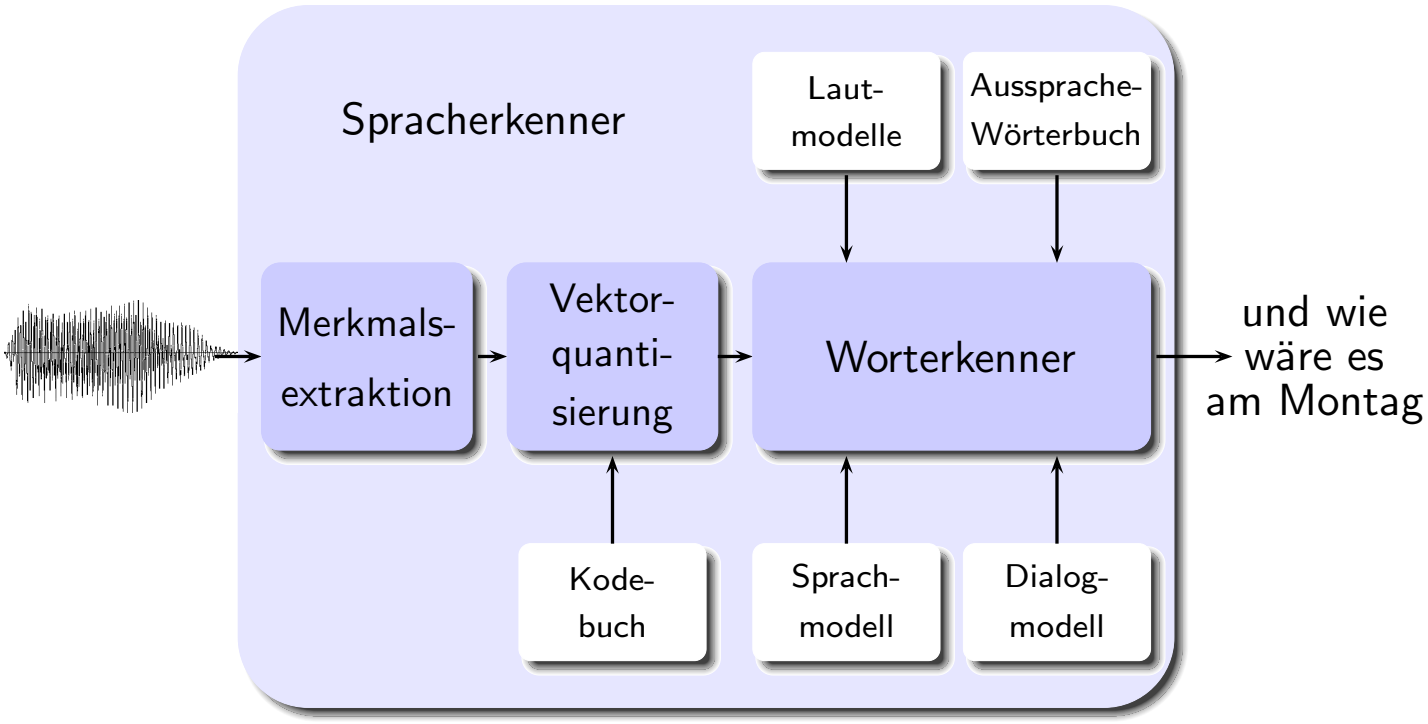
Verteilungen

- Verteilungen sind immer nur Modellannahmen über die Realität
  - Modellparameter werden geschätzt
  - Maß für die Güte des Modells: Konfidenz
  - Abschätzung der Modellgüte über die Varianz
  - Modellevaluation über die Erkennungsergebnisse

Worterkennung

- Klassifikation von atomaren Objekten
- Wahrscheinlichkeitsverteilungen
- Vektorquantisierung
- Klassifikation von Objektsequenzen
- Nichtlineare Zeitnormierung
- Hidden-Markov-Modellierung
- Erkennung mit Hidden-Markov-Modellen
- Training von Hidden-Markov-Modellen
- Kontinuierliche Hidden-Markov-Modelle

Vektorquantisierung



Vektorquantisierung

- Clusterung der Merkmalsvektoren zur Datenreduktion
  - Ersetzen der gemessenen (hochdimensionalen) Merkmalsvektoren durch ein Symbol aus einer endlichen Symbolmenge S

$$\vec{x} \mapsto k \text{ mit } k \in S = \{s_1, \ldots, s_K\}$$

  - jedes Klassensymbol $s \in S$ steht für einen Prototypvektor $\vec{x_i}$ (Kodebuch)
  - keine extern vorgegebene Interpretation für die Klassen
  - nur Anzahl der Klassen vorgegeben
  - Lernstichprobe besteht aus Merkmalsvektoren *ohne* Klassenzuordnung
  - unüberwachtes Lernen ("Lernen ohne Lehrer")
- Elemente von S meist Binärzahlen: $card(S) = 2^R$
  R: Quantisierungsrate in bit/Vektor

Vektorquantisierung

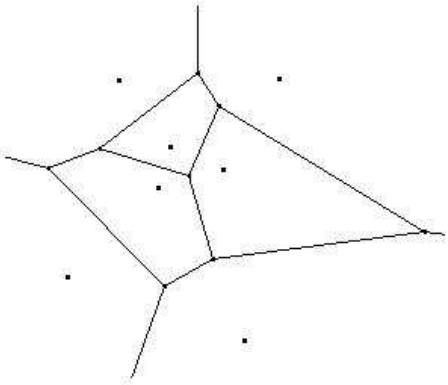
- Nearest-Neigbor-Mapping

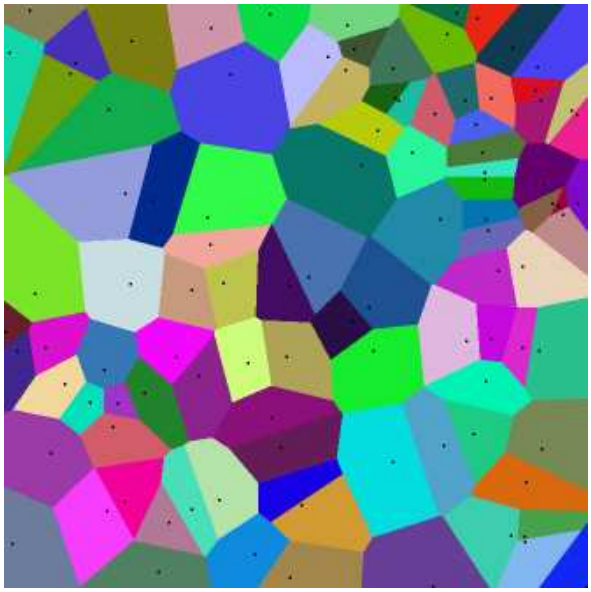
$$k(\vec{x}) = k(\vec{x_i} \mid i = \arg\min_i \; d(\vec{x}, \vec{x_i}))$$

- Distanzmaß
  - Euklidische Distanz
  - Itakura-Saito-Distanz (für LPC-Koeffizienten)

Vektorquantisierung

- vollständige Aufteilung des Merkmalsraumes in Teilbereiche





VORONOI- oder DIRICHLET-Partitionierung

Vektorquantisierung

- Ermittlung der Prototypvektoren $\vec{x_i}$

$$\vec{x_i} = cent(i) = \arg\min_i E(d(\vec{x}, \vec{x_i})|k(\vec{x}) = i)$$

  Zuordnung des Zentroiden jeder Klasse als Prototypvektor
  Mittelwertschätzung
- Teufelskreis:
  - die Zentroidenermittlung setzt Klasseneinteilung voraus
  - die Klasseneinteilung erfolgt aufgrund der Kenntnis der Zentroiden

  $\rightarrow$ rekursive Verfeinerung

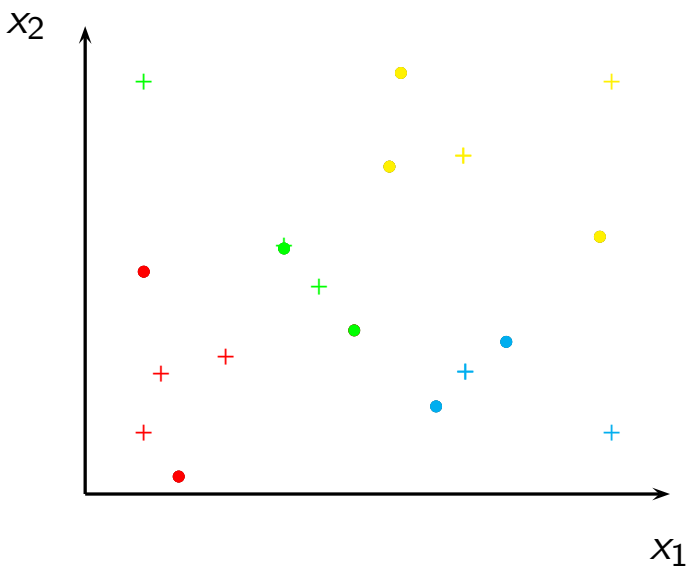
Vektorquantisierung

- Partitionierung des Merkmalsraumes
  Ermitteln der optimalen Zentroiden für eine gegebene Klassenanzahl

  Algorithmus:

  1. Festlegen eines initialen Kodebuches (beliebige Wahl von K Zentroiden)
  2. Klassifizierung der Stichprobe nach der Nearest-Neighbor-Regel
  3. Berechnung der mittleren Distanz für jede Klasse
  4. Berechnung neuer Zentroiden für jede Klasse
  5. Klassifizierung der Stichprobe nach der Nearest-Neighbor-Regel
  6. Ermittlung der mittleren Distanz für jede Klasse
     - Unterschreitet die Änderung der mittleren Distanz einen vorgegeben Schwellwert $\longrightarrow$ Abbruch
     - sonst $\longrightarrow$ 4

Vektorquantisierung



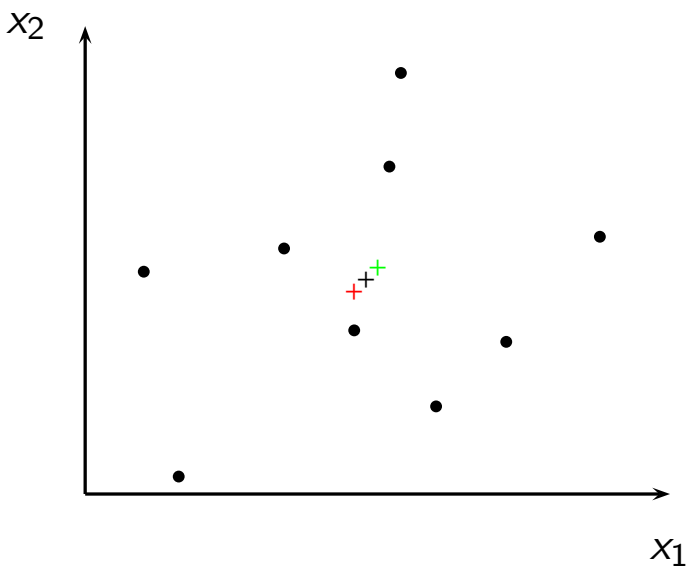
Vektorquantisierung

- Festlegen eines initialen Kodebuches durch rekursives Aufspalten der Kodeklassen

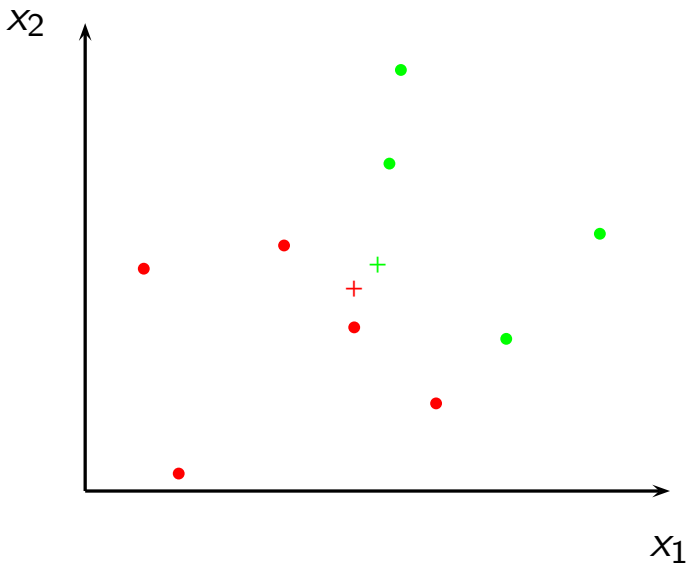
  Algorithmus:

  1. Start mit einem Zentroiden
     (Mittelwertvektor aller Stichprobenelemente)
  2. Aufspalten des / der Zentroiden
     - Zufällige Störung oder
     - Wahl von zwei Punkten mit maximaler Distanz
  3. Ermittlung der optimalen Zentroiden (s.o.)
     - Gewünschte Zahl von Kodeklassen erreicht $\longrightarrow$ Abbruch
     - sonst: $\longrightarrow$ 2

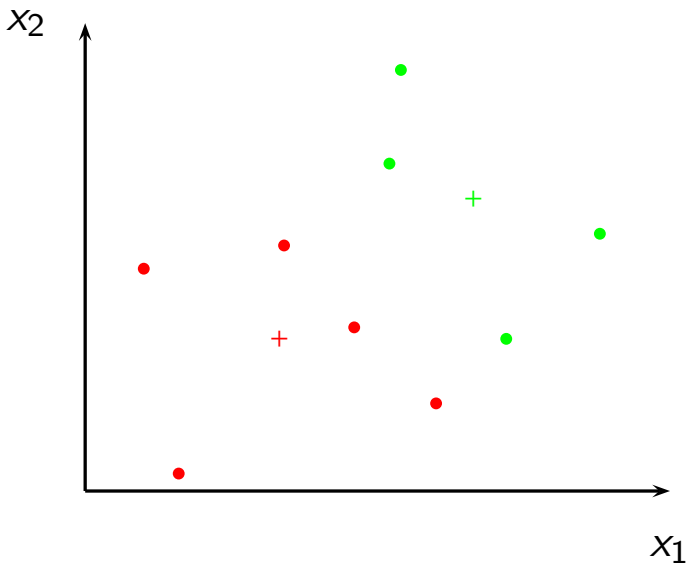
Vektorquantisierung



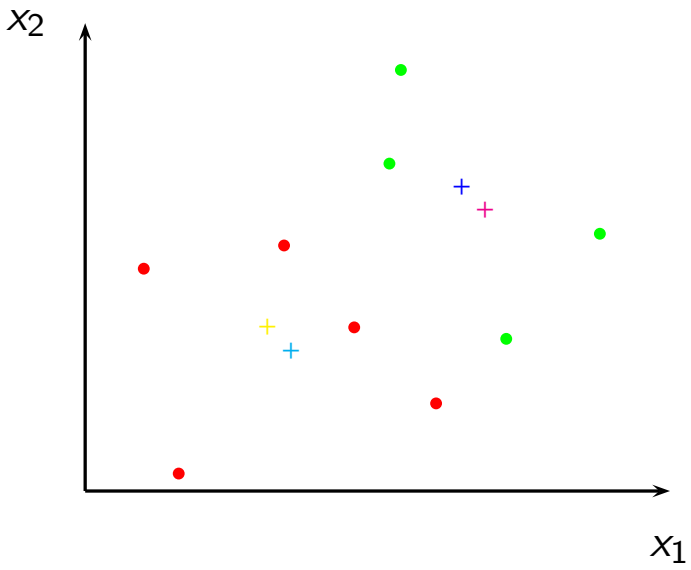
Vektorquantisierung



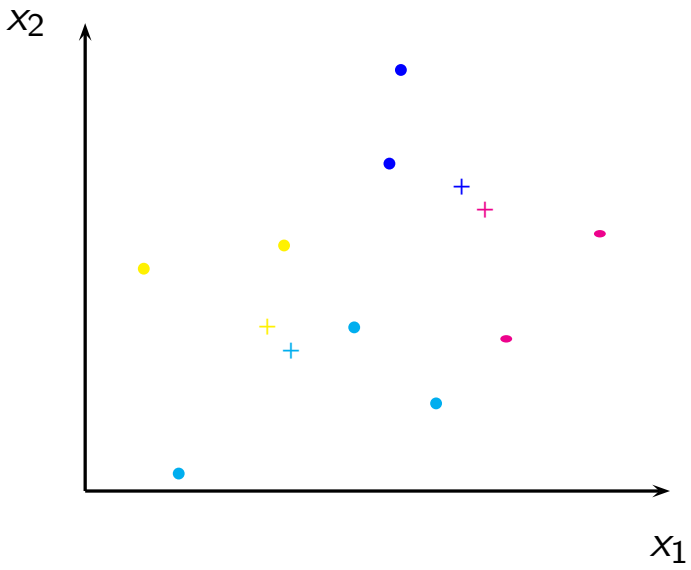
Vektorquantisierung



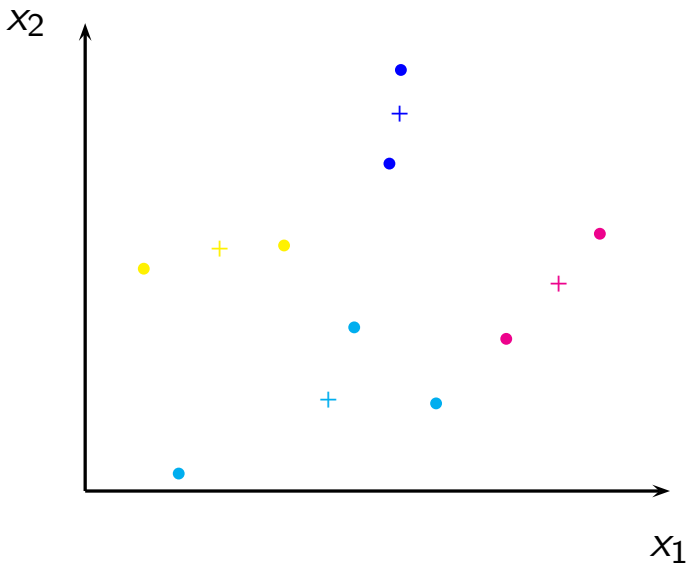
Vektorquantisierung



Vektorquantisierung



Vektorquantisierung



Vektorquantisierung

- Zusätzliche Partitionierungsheuristiken
  - Wenn eine Klasse zu viele Stichprobenelemente enthält, zerlege sie.
  - Wenn zwei Klassen zu dicht benachbart sind, vereinige sie.

Vektorquantisierung

- Quantisierungsfehler
- gemessen am dekodierten Sprachsignal
- Störabstand

$$SNR = 10 \log_{10} \frac{\sum_{\vec{x}} ||\vec{x}||^2}{\sum_{(\vec{x},\hat{\vec{x}})} d(\vec{x},\hat{\vec{x}})}$$

$\vec{x}$ gemessener Inputvektor
$\hat{\vec{x}}$ quantisierter Outputvektor

- Störabstand für Zeitsignalkodierung (Quantisierungsrate: 1 bit/Abtastwert, Abtastfrequenz 6.5 kHz)

Kodebuchgröße	2	4	8	16	32	64	128	256
Störabstand	2.1	5.3	6.0	7.0	7.6	8.1	8.4	8.8

Vektorquantisierung

- Störabstand für die Kodierung von LPC-Koeffizienten Modell 10. Ordnung, untertrainiert

Kodebuchgröße	2	4	8	16	32	64	128	256
Quantisierungsr.	.008	.016	.023	.031	.039	.047	.055	.062
Störabstand	2.9	5.2	6.2	7.9	8.8	9.5	10.1	10.7

- weitere Verbesserungen
  - Vektorquantisierung mit Gedächtnis
  - adaptive Vektorquantisierung
  - Zerlegung des Merkmalsvektors und Verwendung mehrerer Kodebücher

Worterkennung

- Klassifikation von atomaren Objekten
- Wahrscheinlichkeitsverteilungen
- Vektorquantisierung
- Klassifikation von Objektsequenzen
- Nichtlineare Zeitnormierung
- Hidden-Markov-Modellierung
- Erkennung mit Hidden-Markov-Modellen
- Training von Hidden-Markov-Modellen
- Kontinuierliche Hidden-Markov-Modelle

Klassifikation von Objektsequenzen

- Sprache als Folge von atomaren Objekten?
  - frameweise Klassifikation: Mikrophoneme
  - mit anschließender Glättung
  - nicht sehr erfolgreich
- Vergleich zwischen Sequenzen
  - Mustersequenzen $\vec{x_k}[1{:}N_k]$ mit $i = 1, \ldots, I$
  - Inputsequenz $\vec{x}[1{:}N]$
  - Ziel: Abstandsklassifikator
  - ab sofort Vektornotation auch ohne Pfeil: $x[m]$, $x_k[n]$

Klassifikation von Objektsequenzen

- "naiver" Ansatz: lineare Distanz
  - Ermitteln eines Distanzmaßes für jedes Vektorpaar

$$d(x[m], x_k[n]) \quad \text{mit} \quad n = m$$

  - Summierung über die Sequenz

$$d(x[1{:}M], x_k[1{:}N]) = \sum_{j=1}^{J} d(x[j], x_k[j])$$

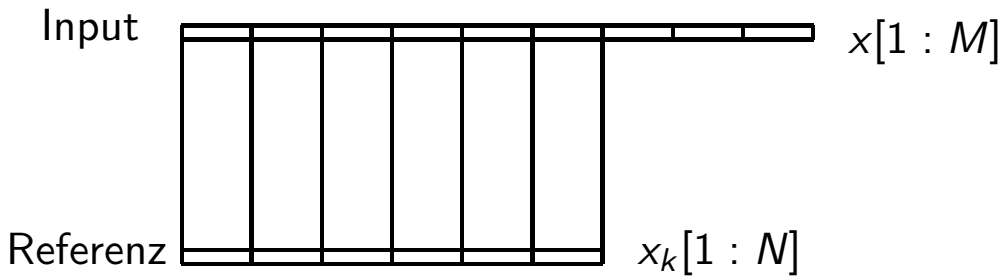
$$\text{mit } J = min\{M, N\}$$

  - Auswahl desjenigen Musters mit der minimalen Distanzsumme

$$k(x[1{:}M]) = \arg\min_k d(x[1{:}M], x_k[1{:}N_k])$$

Klassifikation von Objektsequenzen

- zeitliche Länge von Wortrealisierungen ist variabel: $M \neq N_k$

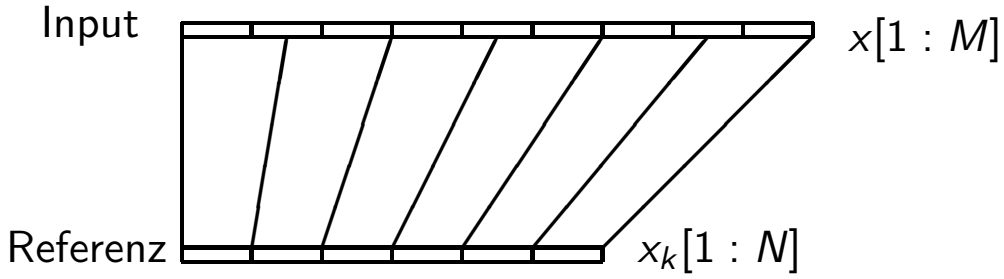


  - äquidistante Zuordnung

→ Zeitnormierung ist erforderlich

Klassifikation von Objektsequenzen

- normierte lineare Distanz
  - alle Vektorsequenzen werden auf eine normierte Länge gestreckt oder gestaucht (Wortmuster und Inputsequenzen)



- lineare Streckung
  - Aufteilung der Vektorsequenz in eine vorgegebene Anzahl von Zeitintervallen
  - bei Bedarf lineare Interpolation

Klassifikation von Objektsequenzen

- in praktischen Realisierungen auch mehrere Referenzmuster für ein Wort
  $\rightarrow$ Nearest-Neighbor-Klassifikator

$$k(x[1{:}M]) = k(x_i[1{:}N_i])$$

$$\text{mit } i = \arg\min_i d(x[1{:}M], x_i[1{:}N_i])$$

Worterkennung

- Klassifikation von atomaren Objekten
- Wahrscheinlichkeitsverteilungen
- Vektorquantisierung
- Klassifikation von Objektsequenzen
- Nichtlineare Zeitnormierung
- Hidden-Markov-Modellierung
- Erkennung mit Hidden-Markov-Modellen
- Training von Hidden-Markov-Modellen
- Kontinuierliche Hidden-Markov-Modelle

Nichtlineare Zeitnormierung

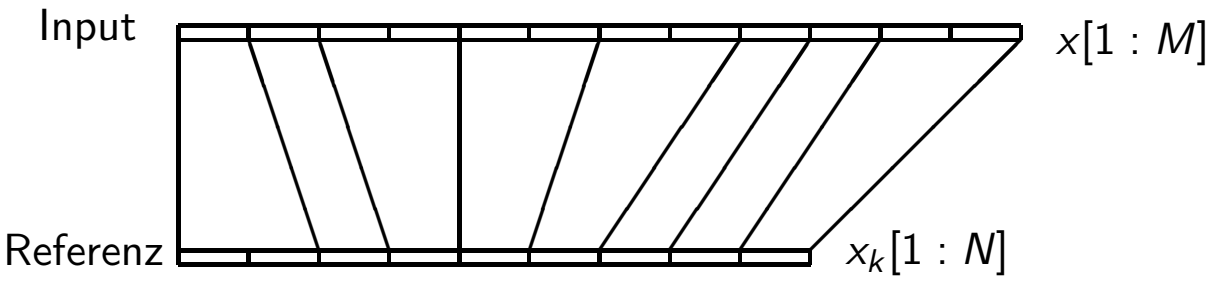
- lineare zeitliche Streckung des Sprachsignals ist starke Idealisierung
  - Sprechgeschwindigkeit ist auch innerhalb eines Wortes variabel
  - Schnellsprache verkürzt nur die stationären Phasen

- Ziel: dynamische Zuordnung (dynamic time warping, DTW), erste Erfolgsstory der Spracherkennung

Nichtlineare Zeitnormierung

- diskrete Signalrepräsentation
- jede Zeitverzerrung läßt sich auf eine kleine Menge zulässiger Elementarzuordnungen zurückführen
  - einem Element im Input entspricht ein Element im Muster
  - einem (oder mehreren) Element(en) im Input entspricht kein Element im Muster
    - Überspringen von Merkmalvektoren im Muster
    - Strecken der Inputsequenz
  - einem (oder mehreren) Elementen im Muster entspricht kein Element im Input
    - Überspringen von Merkmalvektoren im Input
    - Stauchen der Inputsequenz

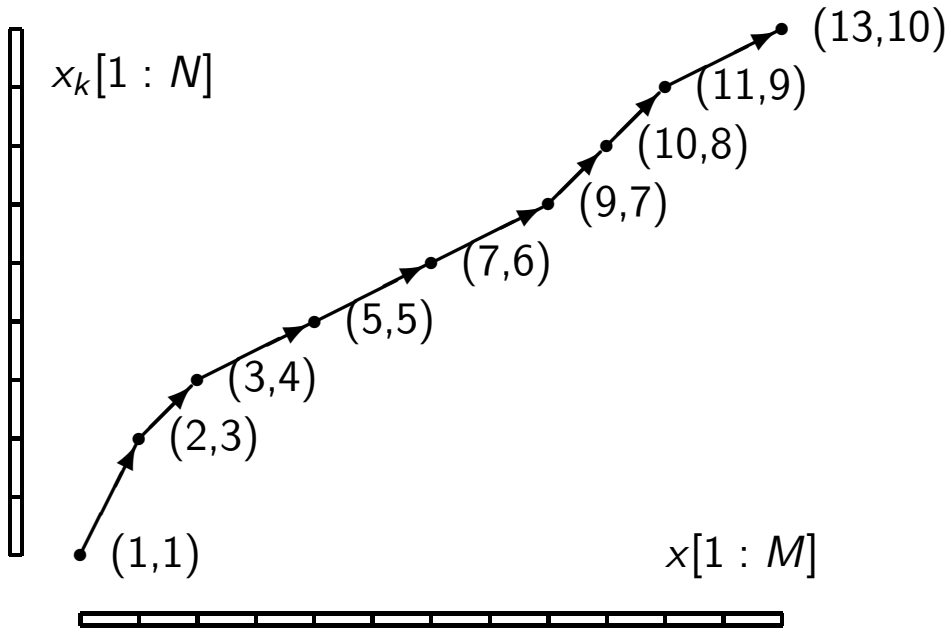
Nichtlineare Zeitnormierung



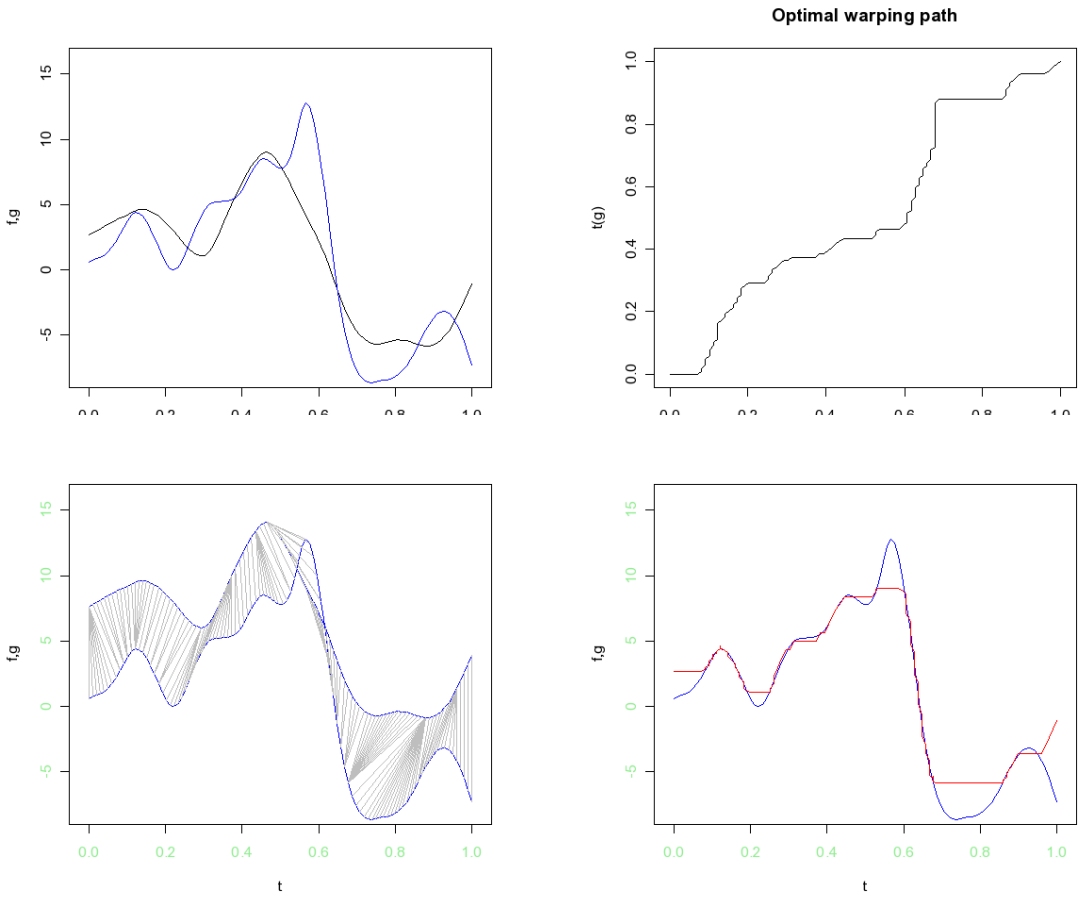
Nichtlineare Zeitnormierung

- Verzerrungs-Funktion (Warping-Funktion)

$$V = v_1 \dots v_I \quad \text{mit} \quad v_i = (m_i, n_i)$$
$$d(v_i) = d(x[m_i], x_k[n_i])$$

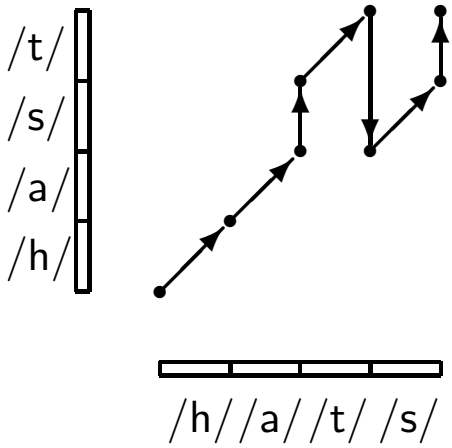


Nichtlineare Zeitnormierung



Nichtlineare Zeitnormierung

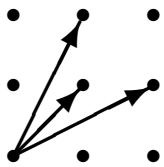
- keine beliebigen Verzerrungsfunktionen zugelassen
  - z.B. Monotonitätsbedingung



Nichtlineare Zeitnormierung

- slope-Constraint für dei Warping-Funktion
- z.B. SAKOE-CHIBA mit Auslassungen

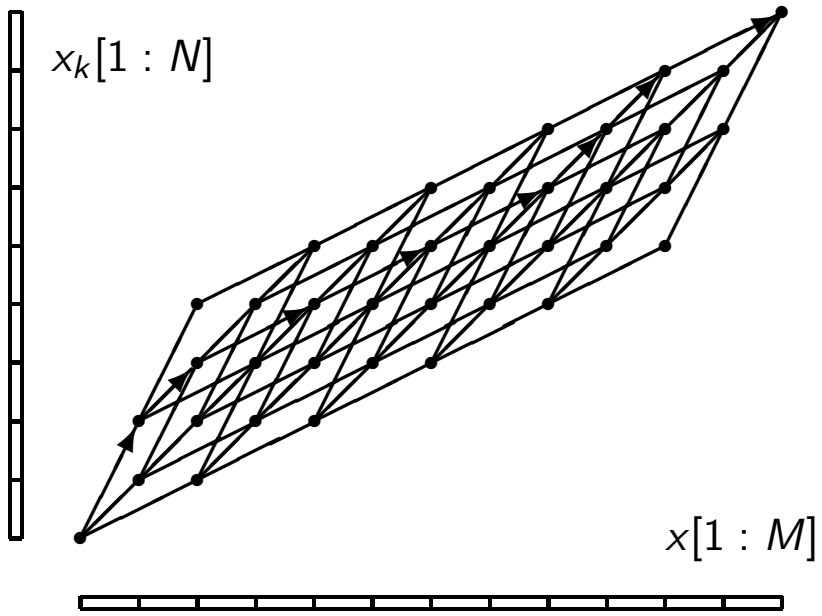
$$v_{i-1} = \begin{cases} (m_i - 1, n_i - 1) \\ (m_i - 2, n_i - 1) \\ (m_i - 1, n_i - 2) \end{cases}$$



  - symmetrisches slope-Constraint

Nichtlineare Zeitnormierung

- Trellis



Nichtlineare Zeitnormierung

- Distanz zweier Vektorsequenzen

$$d(x[1{:}M], x_k[1{:}N]) = \min_{\forall V} \sum_{i=1}^{I} d(v_i)$$

V: Warping-Funktionen

- mit Normierungskoeffizienten w_i

$$d(x[1{:}M], x_k[1{:}N]) = \min_{\forall V} \frac{\sum_{i=1}^{I} d(v_i)\, w_i}{\sum_{i=1}^{I} w_i}$$

mit $\quad d(v_i) = d(x[m_i], x_k[n_i])$

Nichtlineare Zeitnormierung

- symmetrische Normierung

$$w_i = (m_i - m_{i-1}) + (n_i - n_{i-1}) \quad \text{mit} \quad \sum_{i=1}^{I} w_i = M + N$$

Normierung auf die gemeinsame Länge von Input und Muster

- asymmetrische Normierung

$$w_i = (m_i - m_{i-1}) \quad \text{mit} \quad \sum_{i=1}^{I} w_i = M$$

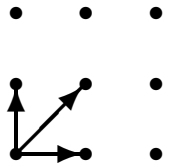
$$d(x[1{:}M], x_k[1{:}N_k]) = \frac{1}{M} \min_{\forall V} \sum_{i=1}^{I} d(v_i)\, w_i$$

Normierung auf die Länge der Inputsequenz

Nichtlineare Zeitnormierung

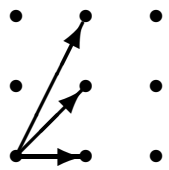
- alternative slope-Constraints
  - SAKOE-CHIBA ohne Auslassungen

$$v_{i-1} = \begin{cases} (m_i - 1, n_i - 1) \\ (m_i, n_i - 1) \\ (m_i - 1, n_i) \end{cases}$$



  - ITAKURA (asymmetrisch)

$$v_{i-1} = \begin{cases} (m_i - 1, n_i) \\ (m_i - 1, n_i - 1) \\ (m_i - 1, n_i - 2) \end{cases}$$



    - erfordert zusätzliche globale Constraints
    - Vorteil: bei asymmetrischer Normierung gilt immer $w_i = 1$

$$d(x[1{:}M], x_k[1{:}N]) = \frac{1}{M} \min_{\forall V} \sum_{i=1}^{I} d(v_i)$$

Nichtlineare Zeitnormierung

- algorithmische Realisierung: Dynamische Programmierung
  - Optimierungsverfahren, Suche in Graphen
  - Graph ist Suchraum aus verschiedenen Zuordnungsvarianten zwischen Muster- und Inputsequenz
  - Suchraum durch die slope-Constraints definiert
  - Wege im Graphen sind bewertet (Distanzmaß an den Knoten)

Prinzip der dynamischen Programmierung (BELLMANN)

Führen zwei Wege zu einem Knoten, so wird der partiell optimale Weg Bestandteil des global optimalen Weges sein, wenn der Knoten Bestandteil des globalen Optimums ist.

Nichtlineare Zeitnormierung

- Rückführung eines globalen Optimierungsproblems durch Rekursion auf lokale Optimalitätsentscheidungen
- für ITAKURA-Constraint:

$$d(x[1{:}i], x_k[1{:}j])$$

$$= \min \left\{ \begin{array}{l} d(x[1{:}i-1], x_k[1{:}j]) \\ d(x[1{:}i-1], x_k[1{:}j-1]) \\ d(x[1{:}i-1], x_k[1{:}j-2]) \end{array} \right\} + d(x[i], x_k[j])$$

Nichtlineare Zeitnormierung

- Training: Abspeichern von Wortmustern
  - rein instanzenbasiert
  - keine Generalisierungen über Klasseneigenschaften

- Alternativen
  - mehrere Muster pro Klasse
    $\rightsquigarrow$ Nearest-Neighbor-Klassifikator für sequenzielle Objekte
  - nichtsequentielle Muster (Phongraphen)
    - Training geht nicht mehr
    - manuell erarbeiten
    - zu aufwendig

DTW-basierte Systeme

- Vintsjuk 1969, Sakoe und Chiba 1970
- Itakura 1975
  - isolierte Wortformen, sprecherabhängig
  - 200 Wortformen, 1 Muster pro Wortform
  - Worterkennungsrate: 97%
- Rabiner et al. 1979
  - isolierte Wortformen, mehrere Sprecher (nicht sprecherunabhängig!)
  - 39 Wortformen, 100 Sprecher, 1 Muster pro Wortform und Sprecher
  - Worterkennungsrate: 79%
- Sakoe 1979
  - einfache Form verbundener Sprache, sprecherabhängig
  - 10 Zahlworte, bis zu vier Zahlworte in Folge
  - Worterkennungsrate: 99.6%

Worterkennung

- Klassifikation von atomaren Objekten
- Wahrscheinlichkeitsverteilungen
- Vektorquantisierung
- Klassifikation von Objektsequenzen
- Nichtlineare Zeitnormierung
- Hidden-Markov-Modellierung
- Erkennung mit Hidden-Markov-Modellen
- Training von Hidden-Markov-Modellen
- Kontinuierliche Hidden-Markov-Modelle

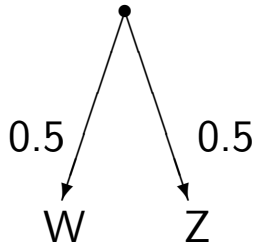
Hidden-Markov-Modellierung

- Ziel: stochastische Verteilungen als Diskriminanzfunktion
  - diskrete Verteilungen
  - kontinuierliche Verteilungen
- generalisierte Beschreibungen für Klassen, nicht für Instanzen

Hidden-Markov-Modellierung

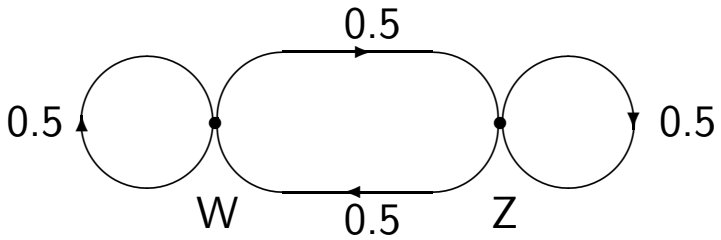
- diskrete stochastische Quellen ohne "Gedächtnis"
  - Beispiel: Münzwurf
  - gleichverteiltes Zweiklassenproblem

$$p_e(Z) = p_e(W) = 0.5 \quad \text{(Emissionswahrscheinlichkeiten)}$$



Hidden-Markov-Modellierung

- Markov-Modelle: diskrete stochastische Quellen mit "Gedächtnis"
  - Zustände z_i
  - Übergangswahrscheinlichkeiten $p_t(z_i|z_j, \ldots, z_k)$
  - Markov-Modelle erster Ordnung: Übergangswahrscheinlichkeit hängt nur von dem aktuellen Zustand ab: $p_t(z_i|z_j)$
  - "Gedächtnis" der Länge eins
  - Münzwurf: jedem Zustand wird eine Münzseite zugeordnet

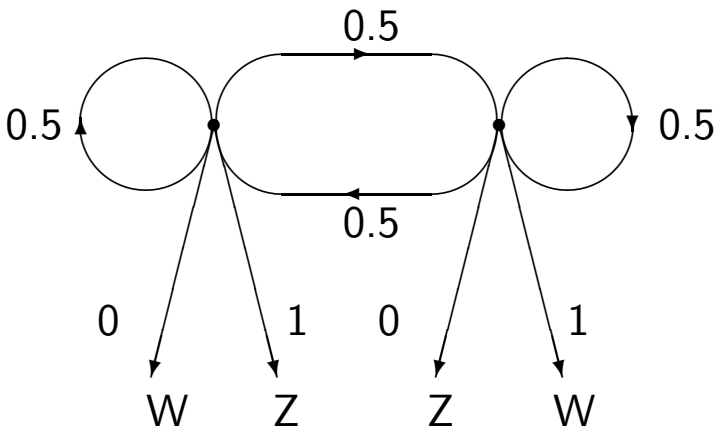


$$p(W|Z) = p(Z|W) = p(W|W) = p(Z|Z) = 0.5$$

$\longrightarrow$ unterschiedliche Modelle für den gleichen stochastischen Prozeß

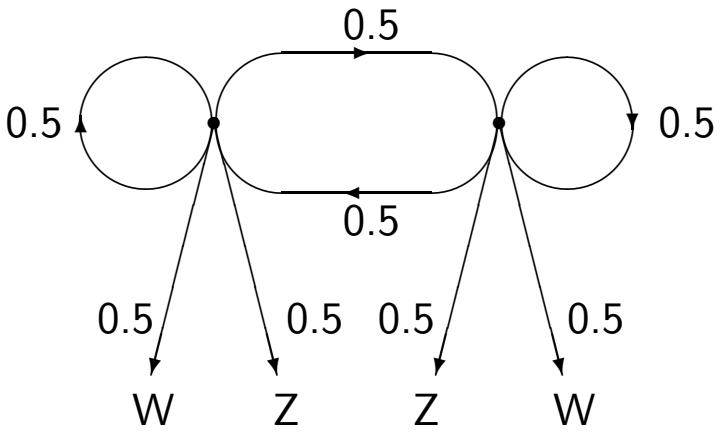
Hidden-Markov-Modellierung

- Hidden-Markov-Modelle
  - Trennung von Ereignissen und Zuständen
  - Zustände z_i
  - Übergangswahrscheinlichkeiten $p_t(z_i|z_j)$
  - Emissionswahrscheinlichkeiten $p_e(x|z_j)$



Hidden-Markov-Modellierung

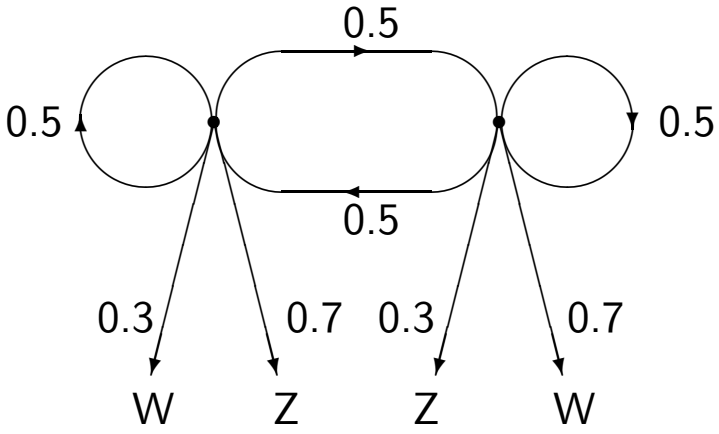
- unterschiedliche Modelle können die gleiche Beobachtung beschreiben
- zwei Münzen, zufällig gewählt



- externe Beobachtbarkeit ist stark eingeschränkt
  In welchen Zustand befindet sich das Modell?
  Mit welcher Münze wurde das Ergebnis ermittelt?

Hidden-Markov-Modellierung

- zwei "gezinkte" Münzen, zufällig ausgewählt



- Langzeitbeobachtung ist unverändert

$$p(Z) = p(W) = 0.5$$

- auch Modelle mit mehr als zwei Zuständen möglich

Hidden-Markov-Modellierung

- Hidden-Markov-Modelle bieten einen extrem flexiblen Rahmen, um Modellstrukturen an Beobachtungsdaten anzupassen.
  - Modellstrukturen
    - verborgene Zustände
    - Übergangswahrscheinlichkeiten
    - Emissionswahrscheinlichkeiten
  - Beobachtungsdaten: emittierte Symbole

Hidden-Markov-Modellierung

- Modellierung
  - Wahl der Modellstruktur
  - Schätzen der Modellparameter

- Modellstrukturen
  - ergodische Markov-Modelle: Übergang zwischen beliebigen Zuständen möglich

$$P_t = \begin{bmatrix} p(1|1) & p(2|1) & \dots & p(m|1) \\ p(1|2) & p(2|2) & \dots & p(m|2) \\ \vdots & \vdots & & \vdots \\ p(1|m) & p(2|m) & \dots & p(m|m) \end{bmatrix}$$

Hidden-Markov-Modellierung

- Sprachsignal ist zeitlich strikt geordnete Ereignissequenz
- Rückkehr zu früheren Zuständen ist kann ausgeschlossen werden
- nur "obere" Dreiecksmatrix besetzt

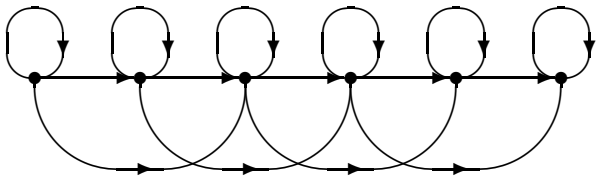
$$P_t = \begin{bmatrix} p(1|1) & p(2|1) & \dots & p(m|1) \\ 0 & p(2|2) & \dots & p(m|2) \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & p(m|m) \end{bmatrix}$$

- zusätzliche Einschränkung auf "kurze Sprünge"

$$P_t = \begin{bmatrix} p(1|1) & p(2|1) & p(3|1) & 0 & 0 & \dots \\ 0 & p(2|2) & p(3|2) & p(4|2) & 0 & \dots \\ 0 & 0 & p(3|3) & p(4|3) & p(5|3) & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

Hidden-Markov-Modellierung

- Bakis-Modell (Bakis 1976)



- vgl. Itakura-Constraint
- weitere Vereinfachung: Weglassen der Sprünge
- Symbolvorrat
  - Signalbeobachtungen, z.B. quantisierte Merkmalsvektoren

Hidden-Markov-Modellierung

- Bewerten der Modellgüte:
  Gegeben: HMM $\mathcal{M} = (\mathcal{Z}, P_i, P_t, P_e)$ und eine Signalbeobachtung $x[1{:}m]$
  Gesucht: $p(x[1{:}m]|\mathcal{M})$
  - Wahrscheinlichkeit, daß die Signalbeobachtung $x[1{:}m]$ durch das Modell $\mathcal{M}$ erzeugt wurde
  - Bewertung der "Güte" des Modells
  - $\rightarrow$ Distanzmaß $d(x[1{:}m], \mathcal{M}) = p(x[1{:}m]|\mathcal{M})$
  - Vorwärts-Algorithmus
    (manchmal auch Vorwärts-Rückwärts-Algorithmus)
- Suche der optimalen Zustandsfolge
  Gegeben: HMM $\mathcal{M} = (\mathcal{Z}, P_i, P_t, P_e)$ und eine Signalbeobachtung $x[1{:}m]$
  Gesucht: "optimale" Zustandsfolge, die das Signal erzeugen würde
  - Viterbi-Algorithmus

Hidden-Markov-Modellierung

- Schätzen der Modellparameter
  Gegeben: eine Zustandsmenge $\mathcal{Z}$ und eine Signalbeobachtung $x[1{:}m]$
  Gesucht: Parametervektoren und -matrizen P_i, P_t und P_e
  - Vorwärts-Rückwärts-Algorithmus, Baum-Welch-Algorithmus

Worterkennung

- Klassifikation von atomaren Objekten
- Wahrscheinlichkeitsverteilungen
- Vektorquantisierung
- Klassifikation von Objektsequenzen
- Nichtlineare Zeitnormierung
- Hidden-Markov-Modellierung
- Erkennung mit Hidden-Markov-Modellen
- Training von Hidden-Markov-Modellen
- Kontinuierliche Hidden-Markov-Modelle

Erkennung mit Hidden-Markov-Modellen

- Vorwärts-Algorithmus
  - Wahrscheinlichkeit dafür, daß eine Beobachtung $x[1{:}n]$ durch ein Modell $\mathcal{M}$ erzeugt wurde

$$p(x[1{:}N]|\mathcal{M}) = \sum_{\forall L} p(x[1{:}n]|L,\mathcal{M})\, p(L|\mathcal{M})$$

$L = l_1, l_2, \ldots, l_l$	beliebiger Pfad in $\mathcal{M}$
$p(L\|\mathcal{M})$	Wahrscheinlichkeit von L in $\mathcal{M}$
$p(x[1{:}N]\|L,\mathcal{M})$	Wahrscheinlichkeit, daß die Beobachtung $x[1{:}n]$ durch den Pfad L in $\mathcal{M}$ erzeugt wurde

Erkennung mit Hidden-Markov-Modellen

$$p(x[1{:}N]|\mathcal{M}) = \sum_{\forall L} p(x[1{:}n]|L,\mathcal{M})\, p(L|\mathcal{M})$$

$$p(x[1{:}N]|L,\mathcal{M}) = \prod_{n=1}^{N} p_e(x[n]|l_n)$$

$$p(L|\mathcal{M}) = p_i(l_1) \prod_{i=1}^{l-1} p_t(l_{i+1}|l_i)$$

- Problem: Summierung über alle Pfade
- $2\,N\,I^N$ Multiplikationen

Erkennung mit Hidden-Markov-Modellen

- Umwandlung in rekursive Rekombinationsgleichung
- Wahrscheinlichkeit für die Erzeugung der Gesamtsequenz am Endknoten
  - Wahrscheinlichkeit für die Erzeugung der um ein Element verringerten Sequenz an allen Vorgängerknoten
  - Wahrscheinlichkeit für den Übergang von den Vorgängerknoten
  - Emissionswahrscheinlichkeit für das letzte Element der Sequenz
- Vorwärtskoeffizienten $\alpha_n(i)$: Wahrscheinlichkeit für die Erzeugung einer Teilsequenz $x[1{:}n]$ durch alle Pfade zum Zustand z_i

Erkennung mit Hidden-Markov-Modellen

- Vorwärtskoeffizienten

$$\alpha_n(i) = p(x[1{:}n], l_n = z_i|\mathcal{M})$$

- Induktionsanfang

$$\alpha_1(i) = p_i(z_i)\; p_e(x[1]|z_i)$$

- Induktionsschritt (Summation über alle eingehenden Kanten)

$$\alpha_{n+1}(j) = p_e(x[n{+}1]|z_j) \sum_{i=1}^{I} \alpha_n(i)\; p_t(z_j|z_i)$$

- gesuchte Bewertung der Inputsequenz

$$p(x[1{:}N]|\mathcal{M}) = \sum_{i=1}^{I} \alpha_N(i)$$

- nur noch Summe über alle (End-)Zustände
- $I^2\, N$ Multiplikationen im allgemeinen Fall
- $k\, I\, N$ Multiplikationen bei Beschränkung auf Bakis-Modelle

Erkennung mit Hidden-Markov-Modellen

- Ermittlung der wahrscheinlichsten Zustandsfolge
- Voraussetzung: Rückwärtskoeffizienten
  analog zu den Vorwärtskoeffizienten

$$\beta_n(i) = p(x[n{:}N]|z_i = l_n, \mathcal{M})$$

$$\beta_N(i) = 1$$

$$\beta_n(j) = \sum_{i=1}^{I} p_t(z_i|z_j)\ p_e(x[n{+}1]|z_i)\ \beta_{n+1}(i)$$

- $\gamma_n(i)$: Wahrscheinlichkeit dafür, daß sich das Modell $\mathcal{M}$
  zur Zeit n im Zustand z_i befindet

$$\gamma_n(i) = p(l_n = z_i|x[1{:}N], \mathcal{M})$$

$$\gamma_n(i) = \frac{\alpha_n(i)\ \beta_n(i)}{p(x[1{:}N]|\mathcal{M})}$$

$$i(n) = \arg\max_i \gamma_n(i)$$

Erkennung mit Hidden-Markov-Modellen

- VITERBI-Algorithmus

$$\delta_n(j) = p_e(x[n]|z_j) \max_i(\delta_{n-1}(i)\ p_t(zj|zi))$$

  - Dynamische Programmierung
  - Verwandschaft zum Vorwärts-Algorithmus
- Suchraumbeschränkungen
  - Einbeziehung des slope-Constraints:
    nur die tatsächlichen Vorgänger / Nachfolger werden
    berücksichtigt
  - Beam-Search

Training von Hidden-Markov-Modellen

- Baum-Welch-Algorithmus
  - rekursive Verfeinerung der Schätzwerte
  - $\xi_n(i,j)$: Wahrscheinlichkeit für einen Übergang von z_i zu z_j bezüglich der Trainingssequenz

$$\xi_n(i,j) = p(l_l = z_i, l_{l+1} = z_j | x[1{:}N], \mathcal{M})$$

$$\xi_n(i,j) = \frac{\alpha_n(i)\ p_t(z_j|z_i)\ p_e(x[n{+}1]|z_j)\ \beta_{n+1}(j)}{p(x[1{:}N]|\mathcal{M})}$$

Worterkennung

- Klassifikation von atomaren Objekten
- Wahrscheinlichkeitsverteilungen
- Vektorquantisierung
- Klassifikation von Objektsequenzen
- Nichtlineare Zeitnormierung
- Hidden-Markov-Modellierung
- Erkennung mit Hidden-Markov-Modellen
- Training von Hidden-Markov-Modellen
- Kontinuierliche Hidden-Markov-Modelle

Training von Hidden-Markov-Modellen

- Neuschätzen der Modellparameter

$$\bar{p}_i(z_i) = \gamma_1(i)$$

$$\bar{p}_t(z_j|z_i) = \frac{\sum_{n=1}^{N-1} \xi_n(i,j)}{\sum_{n=1}^{N-1} \gamma_n(i)}$$

$$\bar{p}_e(x|z_i) = \frac{\left[\sum_{n=1}^{N} \gamma_n(i)\right]_{x[n]=x}}{\sum_{n=1}^{N} \gamma_n(i)}$$

Training von Hidden-Markov-Modellen

- Parameteranzahl pro Zustand:

$$p = t + k$$

  mit t: Anzahl Transitionen pro Zustand, k: Kodebuchgröße
- mehrere Realisierungen pro Wort erforderlich
- Trainingsaufwand ist proportional zur Wortschatzgröße

Worterkennung

- Klassifikation von atomaren Objekten
- Wahrscheinlichkeitsverteilungen
- Vektorquantisierung
- Klassifikation von Objektsequenzen
- Nichtlineare Zeitnormierung
- Hidden-Markov-Modellierung
- Erkennung mit Hidden-Markov-Modellen
- Training von Hidden-Markov-Modellen
- Kontinuierliche Hidden-Markov-Modelle

Kontinuierliche Hidden-Markov-Modelle

- Verwendung kontinuierlicher Wahrscheinlichkeitsdichten für die Emission

$$p(x|z_i) = \mathcal{N}[x, \mu_m, \Sigma_m]$$

- Vorteil: Vektorquantisierung entfällt (Quantisierungsfehler)
- Nachteile:
  - Festlegung auf einen Verteilungstyp
    Diskrete Emissionswahrscheinlichkeiten können beliebige Verteilungstypen approximieren
  - hohe Anzahl an Verteilungsparametern
    pro Modell-Zustand:

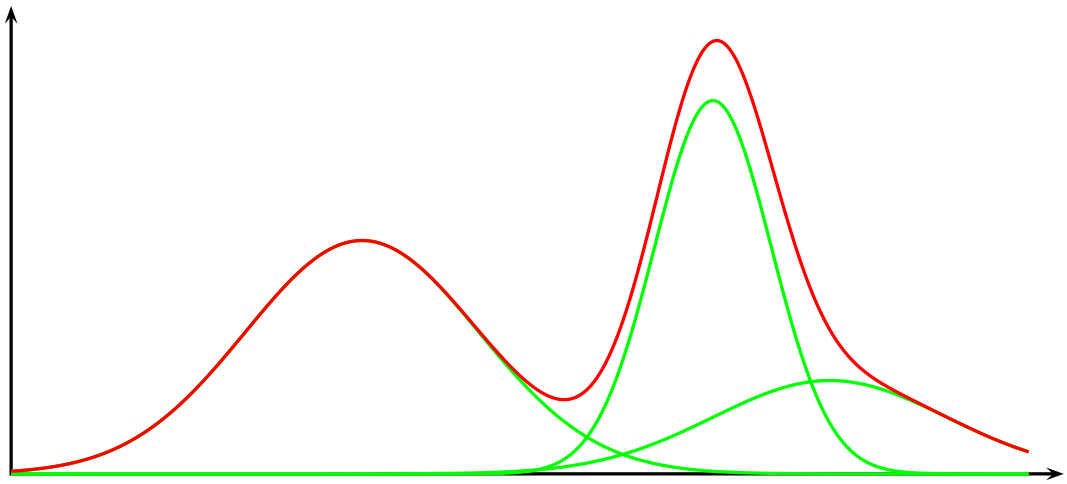
$$p = t + n + n^2$$

    mit n: Dimensionalität der Merkmalvektoren
    $\rightarrow$ mehr Trainingsdaten erforderlich

Kontinuierliche Hidden-Markov-Modelle

- kontinuierliche Mischverteilungen ((Gaussian) mixtures)
  - Konstruktion beliebiger Verteilungsdichten als Summe von elementaren Verteilungsdichten

$$p(x|z_i) = \sum_{m=1}^{M} c_m \, \mathcal{N}[x, \mu_m, \Sigma_m]$$



Kontinuierliche Hidden-Markov-Modelle

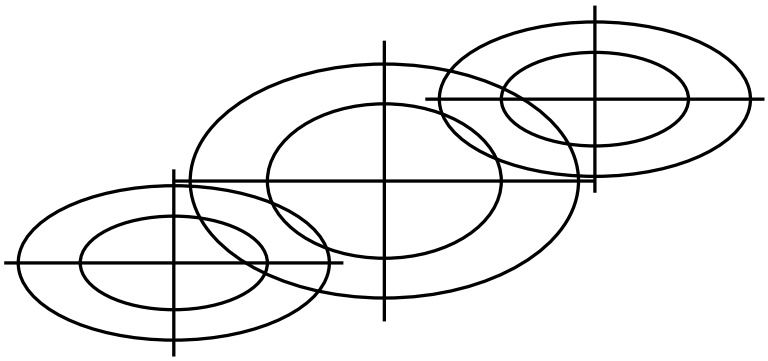
- kontinuierliche Mischverteilungen
  - Vorteil: multimodale Modellierung von Sprache (Sprecherinvarianz)
  - Nachteil: noch größere Anzahl von zu schätzenden Parametern pro Modell-Zustand:

$$p = t + m(n + n^2)$$

    mit m: Anzahl der Verteilungsmodi
    $\rightarrow$ noch mehr Trainingsdaten erforderlich
- Anzahl der Modi
  - uniform: gleich Anzahl pro Modellzustand
  - dynamisch: in Abhängigkeit von den Beobachtungen pro Zustand

Kontinuierliche Hidden-Markov-Modelle

- Reduktion der Parameteranzahl durch Beschränkung auf die Diagonale der Kovarianzmatrix möglich
  - Korrelationen können durch Überlagerung verschiedener Moden approximiert werden.



  - Parameteranzahl pro Zustand

$$p = t + 2mn$$

Kontinuierliche Hidden-Markov-Modelle

- Mischverteilungen ergeben einen zweistufigen stochastischen Prozess
  - Zustand $\rightarrow$ Verteilung $\rightarrow$ Merkmalsvektor
- welche Verteilung hat die Beobachtung erzeugt?
- Training erfordert ebenfalls EM Algorithmus
- Anzahl der Moden pro Zustand kann dynamisch in Abhängigkeit von der Anzahl der Beoachtungen variiert werden

Kontinuierliche Hidden-Markov-Modelle

- tied mixture models: Verwendung kontinuierlicher Verteilungen im Vektorquantisierer
- ebenfalls zweistufige Zuordnung:
  Modellzustand $\mapsto$ diskretes Symbol $\mapsto$ Verteilungsfunktion
- Verteilungsdichtefunktion statt Abstandsmaß

$$p(x|v_i) = \sum_{m=1}^{M} c_m \, \mathcal{N}[x, \mu_m, \Sigma_m]$$

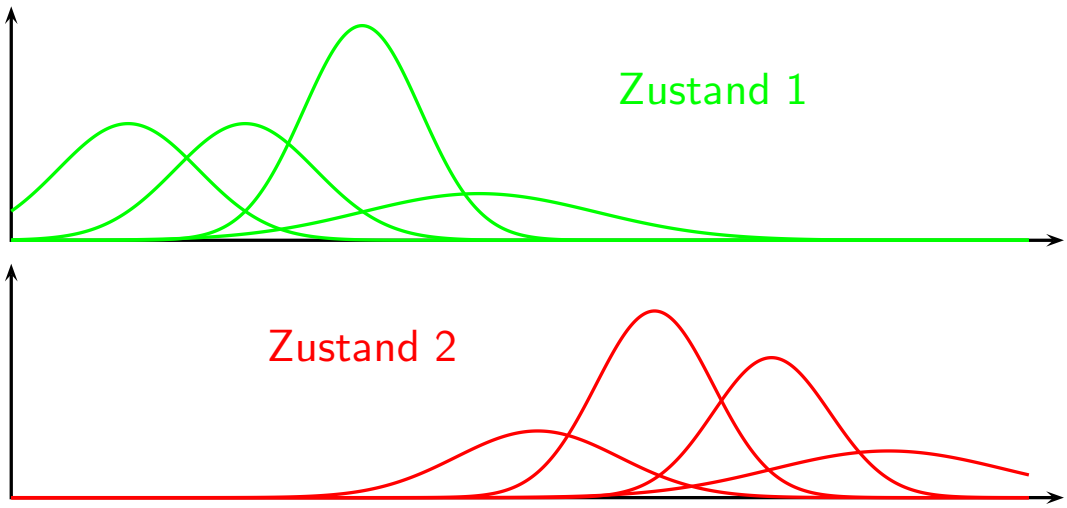
- Verteilungsdichtefunktion eines Zustandes als gewichtete Summe der Verteilungsfunktionen für die emittierten Vektorklassen

$$p(x|z_i) = \sum_{k=1}^{K} p(x|v_k) p_e(v_k|z_i)$$

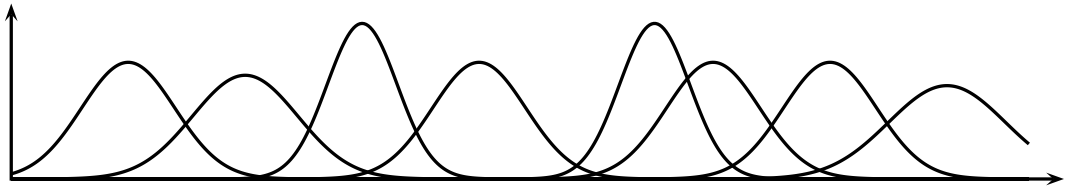
- gemeinsames Inventar von Verteilungen für alle Zustände
- HMM-Training schätzt nur noch die zustandsabhängigen Gewichtsfaktoren für die einzelnen Moden

Semi-kontinuierliche Hidden-Markov-Modelle

- kontinuierliche Mischverteilungen

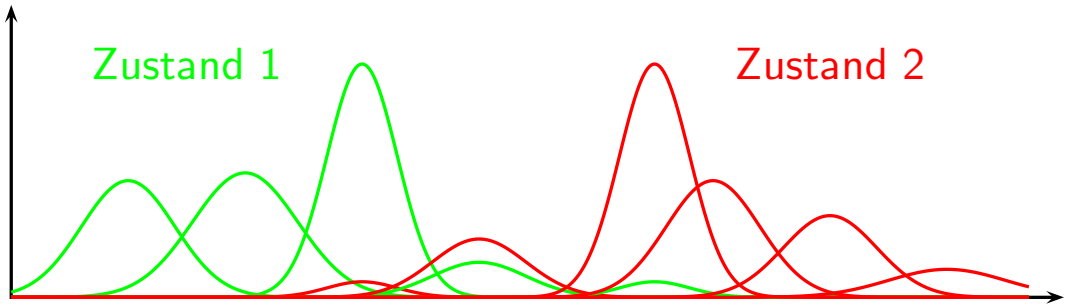


- Kodebuch für semi-kontinuierliche Verteilungen



Semi-kontinuierliche Hidden-Markov-Modelle

- semi-kontinuierliche Verteilung



- Zahl der Vektorklassen $<$ Zahl der Modellzustände
  $\rightarrow$ weniger Parameter zu trainieren

$$p = t + k(d + d^2)$$

- weitere Reduzierung: Summierung über die 2 ... 8 besten Vektorklassen pro Zustand reicht aus

Kontinuierliche Hidden-Markov-Modelle

- diskrete und kontinierliche Modelle sind Spezialfälle der semi-kontinuierlichen Hidden-Markov-Modelle:
  1. Grenzfall 1: alle Vektorklasse emittieren identische rotationssymmetrische Verteilungsfunktionen (Parameterverklebung/tying)
     $\rightarrow$ diskretes HMM
  2. Grenzfall 2: Kodebuch wird so groß gewählt, daß jedem Modellzustand eine eigene Vektorklasse zugeordnet werden kann
     $\rightarrow$ kontinuierliches HMM

Grundlagen der Sprachsignalerkennung

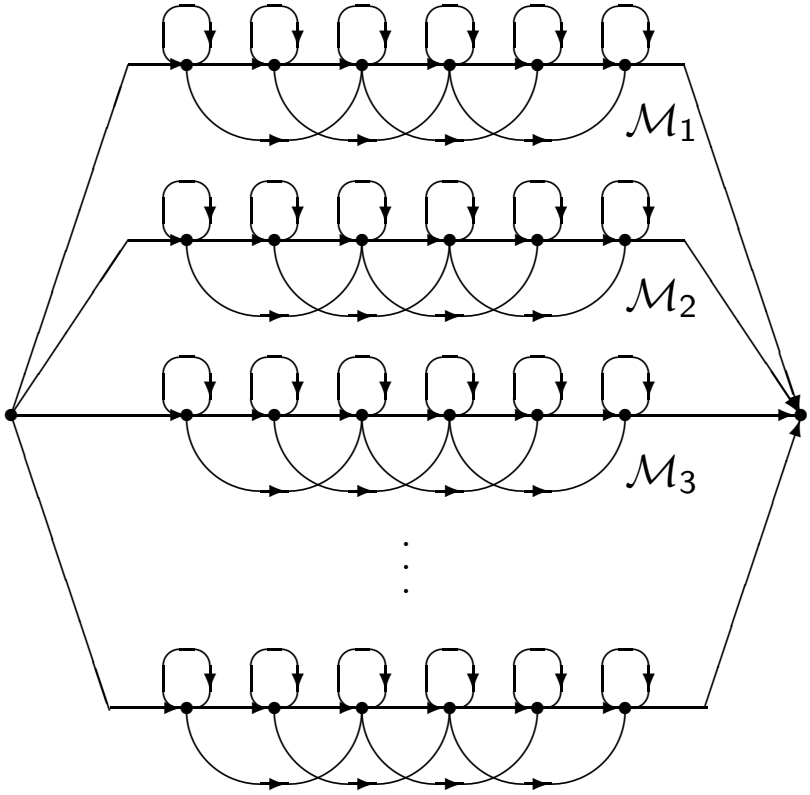
- Spracherkennung als technisches und akademisches Problem
- Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung
- Merkmalsextraktion
- Worterkennung
- Phon-basierte Worterkennung
- Erkennung fließender Sprache
- Systemarchitekturen

Phonbasierte Worterkennung

- Worterkennung bei großem Wortschatz
- Phonmodelle
- Kompositionelle Wortmodelle
- Modellierung der Koartikulation
- Parameterreduktion
- Wörterbucherweiterung

Worterkennung bei großem Wortschatz

- Worterkennung mit n Wortmodellen
  - Parallele Suche durch alle Wortmodelle



Worterkennung bei großem Wortschatz

- Erkennung
  - Vorwärtsalgorithmus

$$w = \arg\max_{w} p(x[1:n]|\mathcal{M}_{\sqsupseteq})$$

  - Viterbi-Algorithmus
    - nur auf isolierte Modelle anwenden
    - ermittelt den optimalen Pfad
    - nicht das optimale Modell

Worterkennung bei großem Wortschatz

- individuelles Training der Wortmodelle
  - abhängig von der Wortschatzgröße
    - Beispiel: Diktiersystem mit 20000 Einheiten
    - 10 Realisierungen pro Wort
    - 100 ... 200 Stunden Training
    - sprecherabhängig!

- Suche nach praktikablen Wortuntereinheiten
  - Silben / Halbsilben
  - Phoncluster
  - Phone

Phonbasierte Worterkennung

- Worterkennung bei großem Wortschatz
- Phonmodelle
- Kompositionelle Wortmodelle
- Modellierung der Koartikulation
- Parameterreduktion
- Wörterbucherweiterung

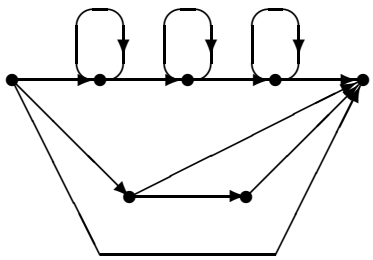
Phonmodelle

- Phone - kleinste segmentale Einheit zur Sprachbeschreibung

- Phoninventar
  - Vokale
  - Konsonanten
  - Wortpause
  - allphonische Varianten, insbesondere für die Plosive

Phonmodelle

- mögliche Modellstruktur (Lee 89)



- hier: Emissionswahrscheinlichkeiten an den Kanten!

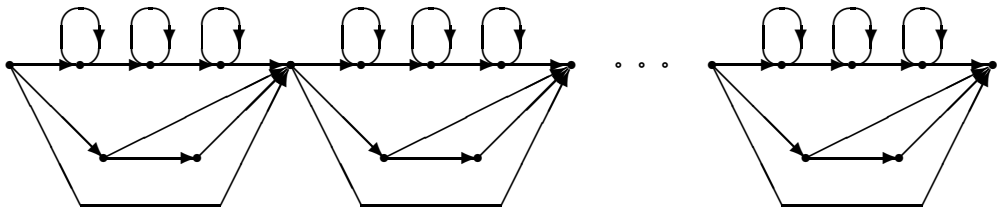
Phonmodelle

- Phonspezifische Topologien
- Modifikation der Zustandsanzahl
  - Pausen und Frikative mit weniger Zuständen
  - Diphthonge mit mehr Zuständen

  $\longrightarrow$ Verschlechterung der Erkennung
- Beseitigen nutzloser Knoten
  - Entfernen aller Kanten mit $p_t(i,j) = 0$
  - Entfernen aller nicht mehr erreichbarer Knoten
    5% Reduktion der Knotenzahl
    10% Reduktion der Kantenzahl
  - nur für ausreichend gut trainierte Modelle!

Phonmodelle

- Kopplung von Phonmodellen zu Wortmodellen

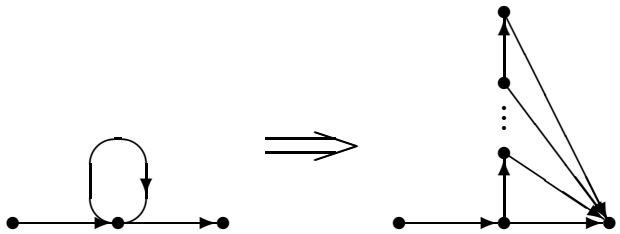


  - beliebige Sequenzen aus Phonmodellen
    - sehr unsichere Erkennungsergebnisse: 60 . . . 70%
    - Beschränkung auf die durch das Wörterbuch liziensierten Phonsequenzen

    $\longrightarrow$ Worterkennung auf der Basis von Wortuntereinheiten
  - unrestringiert verkoppelte Phonmodelle: Modellierung unbekannter Wörter

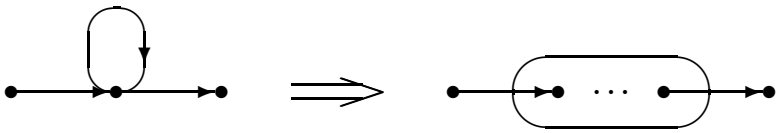
Dauermodellierung

- Minimal- und Maximaldauern für Phoneme
- Semi-Markov-Modelle (Russell 1985)
  - Schätzen der Verweildauer in den Zyklen (modifizierter Vorwärts-Rückwärts-Algorithmus)
  - Integration in die normale Viterbi-Suche
  - D neue Parameter pro Zustand (D maximale Zyklenzahl)
    $\rightarrow$ D^2-fache Rechenzeit
    $\rightarrow$ D-facher Speicherbedarf



Dauermodellierung

- separate Verwaltung der Verweilwahrscheinlichkeiten (Rabiner 1985)
  - z.B. als Postprozessor zur Wichtung von Pfaden
  - suboptimal
  - aber deutlich weniger Rechenzeitaufwand
- Segment-Level-Integration (Lee 1989)
  - Komplexe Zustände mit variabler Länge
  - Übergangs- und Emissionswahrscheinlichkeiten werden erst beim Verlassen eines Zustandes berechnet
  - Multiplikation mit der Verweilwahrscheinlichkeit



Phonbasierte Worterkennung

- Worterkennung bei großem Wortschatz
- Phonmodelle
- Kompositionelle Wortmodelle
- Modellierung der Koartikulation
- Parameterreduktion
- Wörterbucherweiterung

Kompositionelle Wortmodelle

- Aussprachewörterbuch: Zuordnung von Phonsequenzen zu Wortformen
- einfachster Fall: nur kanonische Aussprache pro Wortform
  - keine Berücksichtigung von Aussprachevarianten
  - Trainig erzwingt Zuordnung von Aussprachevarianten zu den Modellen
    - schwa-Ausfall im Wortauslaut: *lesn*, *hörn*
    - Aspiration von Plosiven im Wortauslaut: *Hut*, *Park*
    - Palatalisierung: *jut*, *abjegeben*
    - Dialektale Vokalfärbungen: sächs. *klar*, berl. *mein*
    - Lautreduktion bei Schnellsprache: *national*

  → "verschmutzte" Phonmodelle

Kompositionelle Wortmodelle

- Aussprachealternativen
  - Kombinatorik führt zu zahlreichen Wörterbucheinträgen
  - steigender Erkennungsaufwand
  - Verteilung von Wahrscheinlichkeitsmasse auf die Alternativen
    → Erkennungssicherheit sinkt

Kompositionelle Wortmodelle

- Phonnetze
- manuelle Erarbeitung ist
  - aufwendig und unzuverlässig
- Konstruktion von Aussprachenetzen
  - ausgehend von der kanonischen Aussprache
  - Anwendung phonologische Regeln
- Vorteil: "saubere" Phonmodelle
- Probleme:
  - Viterbi-Algorithmus: Verteilung der Wahrscheinlichkeitsmasse auf zahlreiche alternative Pfade
  - Schätzen der Wahrscheinlichkeiten
    Atlantik mit 6912 Aussprachevarianten (Lee 1988)
    nicht genügend Trainingsmaterial vorhanden

Kompositionelle Wortmodelle

- Vereinfachung: Substitutionsfreie Netze
  - Substitutionen sind weniger kritisch: werden bereits durch die Emissionswahrscheinlichkeiten modelliert
  - Einfügungen und Weglassungen betreffen weniger Varianten: weniger Parameter
- Modelle mit optionalen Phonen
  - nur Weglassungen werden modelliert

Kompositionelle Wortmodelle

- kompositionelle Phone
  - besonders zur Modellierung der nichtstationären Laute (Plosive)
  - optionale Verschlußpause / obligate Verschlußlösung
  - optionale Verschlußpause / optionale Verschlußlösung
    - p_1: nur Verschlußpause
    - p_2: Verschlußpause und -lösung
    - p_3: nur Verschlußlösung
    - $p_1 > p_2 > p_3$
- obligatorische Modifikationen
  - flapping
    better $\rightarrow$ *bedder*
- kompositionelle Modelle für die Affrikaten (/ts/)

Phonbasierte Worterkennung

- Worterkennung bei großem Wortschatz
- Phonmodelle
- Kompositionelle Wortmodelle
- Modellierung der Koartikulation
- Parameterreduktion
- Wörterbucherweiterung

Modellierung der Koartikulation

- Phonbasierte Worterkennung: doppelte Fehlerrate gegenüber der Erkennung mit Ganzwortmodellen
- Ursache: Nichtbeachten der Koartikulation
  - physikalische Beschränkungen für die Bewegung der Artikulatoren
- alternative Wortuntereinheiten
- Silben / Halbsilben (Phonsequenzmodelle)
  - zu viele $\rightarrow$ Trainingsprobleme
- Diphone, Transitionsmodelle
  - zu viele $\rightarrow$ Trainingsprobleme

Modellierung der Koartikulation

- Wortabhängige Phonmodelle
  - Phonmodelle im Kontext des Wortes, in dem sie auftreten
  - sehr viele
  - aber interpolationsfähig
    - Nutzung von Parametern für besser trainierbare Modelle
    - z.B. Interpolation mit kontextunabhängigen Phonmodellen
  - Wortschatz- und damit auch anwendungsspezifisch
  - nur für häufige Wortformen praktikabel
    $\rightarrow$ funktionswortabhängige Phonmodelle

Modellierung der Koartikulation

- kontextabhängige Phonmodelle ("Triphone")
  - Phonmodelle im Kontext ihrer unmittelbaren Nachbarphone
  - ebenfalls interpolationsfähig
  - Interpolation mit
    - links-kontextabhängigen Modellen
    - rechts-kontextabhängigen Modellen
    - kontextunabhängigen Modellen
  - Generalisierung: Polyphone, auch auf Graphembasis
    - Berücksichtigung des Kontexts in genau dem Maße, wie ausreichend Trainingsdaten zur Verfügung stehen
  - großer Speicheraufwand (24MByte)
  - keine Generalisierung über Kontextklassen
    - z.B. ist der Kontexteinfluß von stimmhaften und stimmlosen Plosiven im wesentlichen gleich

Modellierung der Koartikulation

- Clustern von Kontexten
  - datengetriebenes Ranking von Entscheidungsfragen (Expertenwissen)
    - aufgrund ihres Informationsgewinn
    - nach ihrem Einfluss auf die Erkennungsqualität
- nur die Zusammenfassungen mit geringen Auswirkungen werden berücksichtigt
- gemeinsames Training der Kontextklassen oder nachträgliches Mischen der Modelle

Modellierung der Koartikulation

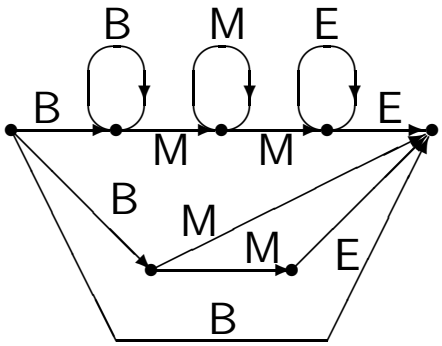
Erkennungseinheit	Modellierung der Koartikulation	Trainier-barkeit	Interpolier-barkeit
Wortmodelle	gut	schlecht	unmöglich
Phonmodelle	schlecht	gut	unnötig
Phonsequenzmodelle	gut	schwierig	schlecht
Transitionsmodelle	gut	schwierig	schlecht
Wortabhängige Phonmodelle	gut	schwierig	gut
Kontextabhängige Phonmodelle	gut	schwierig	gut

Phonbasierte Worterkennung

- Worterkennung bei großem Wortschatz
- Phonmodelle
- Kompositionelle Wortmodelle
- Modellierung der Koartikulation
- Parameterreduktion
- Wörterbucherweiterung

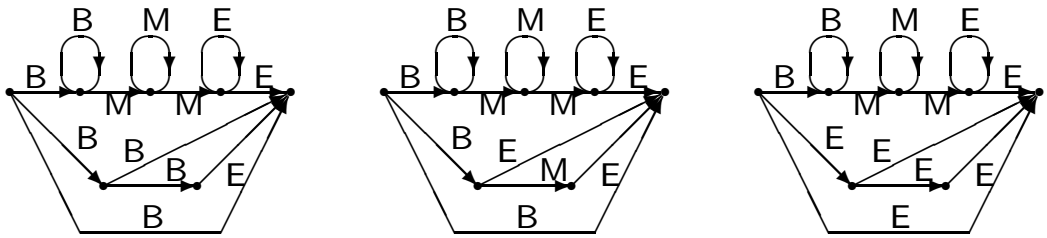
Parameterreduktion

- Tying (Verkleben)
  - Zusammenfassen von Emissionswahrscheinlichkeiten



Parameterreduktion

- Phonabhängiges Tying



1. Vokale (a, e, o), Diphthonge, Frikative
2. Vokale (schwa, i, u), initiale Konsonanten
3. Plosive

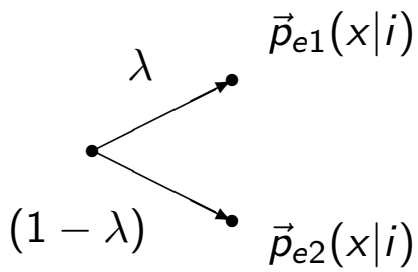
- Vernachlässigung der Transitionswahrscheinlichkeiten (Ney 1993)
  - keine nennenswerte Verschlechterung der Erkennungsergebnisse

Parameterreduktion

- Interpolation
- Kombination zweier unabhängig ermittelter Verteilungen
  z.B. kontextunabhängige / kontextabhängige Phone

$$\vec{p}_e(x|i) = \lambda\, \vec{p}_{e1}(x|i) + (1-\lambda)\, \vec{p}_{e2}(x|i)$$

- Schätzen von λ



  auf der Basis von Daten, die noch nicht für die Berechnung von $\vec{p}_{e1}(x|i)$ und $\vec{p}_{e2}(x|i)$ verwendet wurden
- Ziel: gute Prädiktionsfähigkeit des interpolierten Modells für *neue* Beobachtungen

Parameterreduktion

- deleted interpolation
  - Abtrennen eines für die λ-Schätzung benötigten Datenblocks (deleted data block)
  - zyklisches Austauschen des reservierten Datenblocks
  - erheblich höherer Trainingsaufwand

- Smoothing
  - untertrainierte Modelle enthalten viele Singularitäten
    - zufällig nie aufgetretene Beobachtungen
    - numerische Probleme
  - $p = 0$ bedeutet aber: Ereignis ist unmöglich!

Parameterreduktion

- Smoothing
  1. Floor-Methode
     - jeder Parameter $p < \varepsilon$ wird auf ε angehoben
     - Ausgleich der Wahrscheinlichkeiten erforderlich $\sum p_i = 1$
     - entspricht Interpolation mit einer Gleichverteilung
  2. Abstandsmethode
     - ist die Emissionswahrscheinlichkeit für einen Prototypvektor $p_e(i) \approx 0$ und es gibt einen benachbarten Prototypvektor mit $p_e(i) >> 0$ und gilt $d(i,j) < \varepsilon$, so erfolgt ein Ausgleich
  3. Co-occurence-Methode
     - Ausgleich erfolgt vorrangig zwischen Prototypvektoren, die in gleichen Kontexten auftreten

Phonbasierte Worterkennung

- Worterkennung bei großem Wortschatz
- Phonmodelle
- Kompositionelle Wortmodelle
- Modellierung der Koartikulation
- Parameterreduktion
- Wörterbucherweiterung

Wörterbucherweiterung

- Wörterbucherweiterungen erfordern Ergänzung des Aussprachewörterbuchs

- nur sinnvoll, wenn Systemreaktion auch problemlos definiert werden kann
  → Diktiermaschine
  - manuelle Eingabe einer orthographischen Repräsentation
  - Ermitteln der Normaussprache
    - Graphem-Phonem-Umsetzung aus der Schriftversion
    - manuelle Eingabe einer pseudophonetischen Umschrift: *Äktschn*

Grundlagen der Sprachsignalerkennung

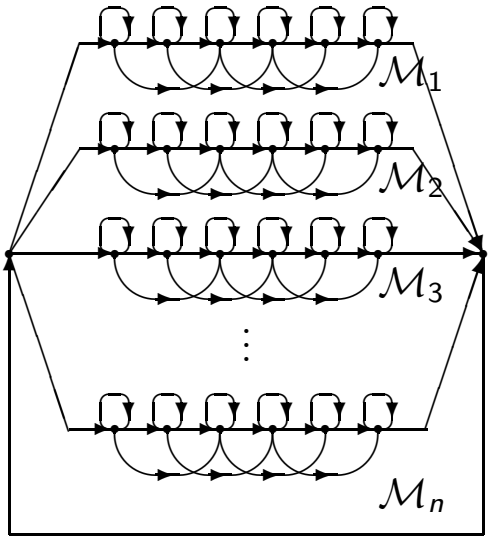
- Spracherkennung als technisches und akademisches Problem
- Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung
- Merkmalsextraktion
- Worterkennung
- Phon-basierte Worterkennung
- Erkennung fließender Sprache
- Systemarchitekturen

Erkennung fließender Sprache

- Fließende Sprache
- Sprachmodelle
- Dialogmodelle
- Fehlermaße
- Training und Erkennung
- Behandlung unbekannter Wörter
- Sprecherunabhängigkeit

Fließende Sprache

- Verbundene Sprache (connected speech)
  - Aneinanderreihung von wohlformulierten Wortformen
- unrestringierte Verkettung von Wortmodellen
  - Ganzwortmodelle oder phonbasierte Wortmodelle



Erkennung fließender Sprache

- Entscheidung über Wortsequenzen, nicht Einzelwörter
  - variable *a priori*-Wahrscheinlichkeiten für Wortformen

$$p(w|x) = \frac{p(x|w)\ p(w)}{p(x)}$$

$p(w)$ ist kontextabhängig

  - unterschiedliche *a priori*-Wahrscheinlichkeiten für Wortformsequenzen

$$p(w[1:n]|x) = \frac{p(x|w[1:n])\ p(w[1:n])}{p(x)}$$

$$w = \arg\max_{w}\ p(x|w[1:n])\ p(w[1:n])$$

Erkennung fließender Sprache

- fließende Sprache (continuous speech)
  - Koartikulation an Wortübergängen
  - Modellierung nur für phonbasierte Modelle realistisch (cross-word triphones)
  - Reduktion der Fehlerrate um 15 % (relativ) Huang et al. 1990
  - aber Einfluss des Testmaterials: unterschiedliche Sprechstile haben unterschiedlichen Grad an Koartikulation

Erkennung fließender Sprache

- Fließende Sprache
- Sprachmodelle
- Fehlermaße
- Dialogmodelle
- Training und Erkennung
- Behandlung unbekannter Wörter
- Sprecherunabhängigkeit

Sprachmodelle

- Modellierung der zulässigen bzw. typischen Wortsequenzen innerhalb des konkreten Anwendungsgebietes

- binärer Ansatz: formale Grammatik G

$$p(w[1:n]) = \begin{cases} \text{const} \neq 0 & \text{für } w[1:n] \in L(G) \\ 0 & \text{sonst} \end{cases}$$

  - Gleichverteilung für alle durch die Grammatik lizensierten Wortformsequenzen
  - Approximation durch Wortpaar-Grammatik
  - einfache Realisierung für reguläre Grammatiken: dynamische Einschränkung der Wortschatzgröße auf die möglichen Nachfolger: Wortnetze

Sprachmodelle

- dynamische Beschränkung des aktuellen Wortschatzes
- statischer Verzweigungsfaktor b

$$b = \frac{e}{n - n_f}$$

e	Anzahl der Transitionen
n	Anzahl der Zustände
n_f	Anzahl der Endzustände

  beschreibt die restriktive Kapazität der Grammatik
- reguläre Grammatiken erlauben keine (beschreibungs-) adäquate Modellierung für die Syntax natürlicher Sprache
  - Ableitung aus einer kontextfreien (Unifikations-) Grammatik
  - "obere" Approximation $L(G_{reg}) \supset L(G_{cf})$

Sprachmodelle

- stochastische Modelle

$$p(x_1 ... x_n) = p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1 x_2) \cdot ... \cdot p(x_n|x_1 ... x_{n-1})$$

- Approximation durch (Unigram-) / Bigram- / Trigram-Statistiken

$$p(x_1 ... x_n) \approx p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_2) \cdot ... \cdot p(x_n|x_{n-1})$$

  - (m,) m^2 bzw. m^3 Parameter (mit m: Wortschatzgröße)
  - wachsender Speicherbedarf
  - zunehmende Datenknappheit
  - Spezialfall: Bigram-Statistik mit Gleichverteilung: Wortpaargrammatik
- bei großem Wortschatz auch kategorienbasierte n-gram-Statistiken
  - Clustering oder
  - automatisches Tagging
- stochastische Phrasenstrukturgrammatiken
  - Modellierung längerer Abhängigkkeiten

Sprachmodelle

- Beispiele für Markov-Ketten
  - Unigramme (0. Ordnung) $p(s_i)$

    aiobnin*tarsfneonlpiitdregedcoa*ds*e*dbieastnreleeucdkeaitb* dnurlarsls*omn*keu**svdleeoieei* ...
  - Bigramme (1.Ordnung) $p(s_i|s_{i-1})$

    er*agepteprteiningeit*gerelen*re*unk*ves*mterone*hin*d*an* nzerurbom* ...
  - Trigramme (2.Ordnung) $p(s_i|s_{i-1}s_{i-2})$

    billunten*zugen*die*hin*se*sch*wel*war*gen*man*nicheleblant* diertunderstim* ...
  - Quadrogramme (3. Ordnung) $p(s_i|s_{i-1}s_{i-2}s_{i-3})$

    eist*des*nich*in*den*plassen*kann*tragen*was*wiese*zufahr* ...

Perplexität

- Systeme mit Sprachmodell sind nicht mehr ohne weiteres vergleichbar
  $\rightarrow$ Maß für die restriktive Kraft eines Sprachmodells erforderlich
- Sprachmodell beschreibt einen stochastischen Prozeß zur Spracherzeugung
- Entropie H (in bit) ist ein Maß für den Informationsgehalt einer stochastischen Quelle (aus der Sicht des Empfängers)
  - $H(S)$ entspricht dem mittleren (minimalen) Kodieraufwand für die Symbole einer Quelle
  - Perplexität ist mittlere Mehrdeutigkeit der Quelle

$$Q(S) = 2^{H(L)}$$

- Informationsgehalt eines stochastischen Ereignisses

$$I(w) = \log_2 \frac{1}{p(w)}$$

Perplexität

- Entropie einer Quelle S, die Symbole $w \in W$ emittiert

$$\begin{aligned} H(S) &= \sum_w p(w) I(w) \\ &= \sum_w p(w) \log_2 \frac{1}{p(w)} \\ &= -\sum_w p(w) \log_2 p(w) \end{aligned}$$

- Perplexität: "Mehrdeutigkeit" einer Quelle

$$Q(S) = 2^{H(S)}$$

Perplexität

- Testkorpusperplexität (nach JELINEK 1975)
  - wenn Berechnung über Grammatik nicht möglich
- Testkorpus als stochastische Quelle

$$H(S) = -\sum_{w} p(w) \log_2 p(w)$$

$$H(S) = -\lim_{n\to\infty} \frac{1}{n} \sum_{w[1:n]} p(w[1:n]) \log_2 p(w[1:n])$$

Test: für unabhängige Ereignisse gilt ("it can be easily checked")

$$p(w[1:n]) = p(w_1) \cdot p(w_2) \cdot \ldots \cdot p(w_n)$$

Perplexität

- für ergodische Quellen kann gezeigt werden, dass

$$H(S) = -\lim_{n\to\infty} \frac{1}{n} \log_2 p(w[1:n])$$

$\rightarrow$ Entropie kann auf der Basis einer hinreichend langen Wortsequenz $T = w[1:n]$ geschätzt werden ($n \rightarrow \infty$)

$$H(T) = -\frac{1}{n} \log_2 p(w[1:n])$$

$$\begin{aligned} Q(T) &= 2^{H(T)} \\ &= p(w[1:n])^{-\frac{1}{n}} \end{aligned}$$

Perplexität

- Approximation der Wahrscheinlichkeit für das Testkorpus über eine Bigram(-Trigram)-Statistik
- Satzgrenzenmarkierungen s gelten als eigenständige Wortformen

$$p(w[1:n]) \approx p(w_1|s)p(w_2|w_1)p(w_3|w_2)\ldots p(s|w_n)$$

  - untere Approximation für die Wahrscheinlichkeit
  - obere Approximation für die Entropie / Perplexität

Perplexität

- Beipielperplexitäten

Aufgabe / System	Wortschatz	Wortpaar-grammatik	Bigram-modell	Trigram-modell
HARPY	1011	4.5		
RM	997	60	20	
ATIS	1401		20	
WSJ	5000		113	147
VM	3000		85	
Switchboard				68
Broadcast News				217

- Faustregel: $Q/10 \rightarrow$ Fehlerrate/2

Sprachmodelle

- die überwiegende Zahl von n-grammen wird im Training nie gesehen
  → Backoff-Modelle
  trigram → bigram → unigram
- lange Abhängigkeiten werden durch n-gramme nicht erfasst
  - distance bigrams: $p(w_n|w_{n-d-1})$
  - für Klassen und Wörter
  - Kombination mit n-gram durch (log-)lineare Interpolation
- stochastische CFG-Parser
  - Berechnung von Präfix-Wahrscheinlichkeiten
  - lexikalisierte Modelle
  - Modellierung der Kopfabhängigkeiten (exposed head word)
    $p(w|w_{exposed\ head})$

Erkennung fließender Sprache

- Fließende Sprache
- Sprachmodelle
- Fehlermaße
- Dialogmodelle
- Training und Erkennung
- Behandlung unbekannter Wörter
- Sprecherunabhängigkeit

Fehlermaße

- Satzerkennungsrate
  - oftmals recht klein
  - abhängig von der mittleren Satzlänge
  - schlechte Vergleichbarkeit unterschiedlicher Lösungsansätze
- Generalisierung der Worterkennungsrate auch für Wortsequenzen
- Probleme mit der Behandlung von Einfügungen und Weglassungen
  - Sind Einfügungen Erkennungsfehler?
  - *recognize* $\rightarrow$ *wreck a nice* = 3 Fehler?
  - Zuordnung der fehlerhaften Wortkette zur korrekten

    A B C A B C
    | | |
    A C A C
    2 Fehler 1 Fehler

    $\rightarrow$ Suche nach einer optimalen Zuordnung (minimale Fehleranzahl)

Fehlermaße

- Wortkettenzuordnung
  - elementare Wortpaarzuordnungen
    - Identität (I)
    - Substitution (S)
    - Einfügung (E)
    - Weglassung (W)

Fehlermaße

- Distanzmaß für elementare Wortpaarzuordnungen

  LEVENSTEIN-Metrik

Identität	0
sonst	1

- Einschränkungen für die Fehlerkombinierbarkeit (z.B. W S = S W)
- ermitteln der optimalen Zuordnung durch dynamische Programmierung

A	B	C	D	E	F	
I	W	I	W	I	S	E
A		C		E	G	H

- Normierung der Fehlermaße auf die Länge L der (korrekten) Referenzkette

Fehlermaße

- Korrektheit (correctness in %)

$$K = \frac{I}{L}$$

- Fehlerrate (error rate in %)

$$F = \frac{S + E + W}{L}$$

  Fehlerrate kann größer als 100% sein

- Genauigkeit (accuracy in %)

$$G = 1 - F = \frac{I - E}{L} \qquad \text{wegen } L = I + S + W$$

Fehlermaße

- Beispiele für Erkennungsraten

System	Jahr	Sprecher-	Wortsch.	Perpl.	K	G
HARPY	1978	abhängig	1011	4.5	97%	
BYBLOS	1987	abhängig	997	60	94.8%	92.5%
BYBLOS	1987	abhängig	997	997	70.1%	67.6%
SPHINX	1988	unabh.	997	20	96.2%	95.8%
SPHINX	1988	unabh.	997	60	94.7%	93.7%
SPHINX	1988	unabh.	997	997	73.6%	70.6%
Philips	1992	abhängig	997	997	95.0%	92.3%
Philips	1993	abhängig	12 073 ... 15 188	42 ... 267	91.3% ... 96.6%	87.9% ... 94.3%
VM	1995	unabh.	3000	75		72%
VM	2000	spontan	10000	87		75%

Erkennung fließender Sprache

- Fließende Sprache
- Sprachmodelle
- Fehlermaße
- Dialogmodelle
- Training und Erkennung
- Behandlung unbekannter Wörter
- Sprecherunabhängigkeit

Dialogsysteme

- Typischer Einsatzfall: Füllen von Slots in einer Informationsstruktur
  - Bestellungen, Reservierungen, Auskünfte
- pragmatische Annahmen
  - Nutzer ist an einer Problemlösung interessiert
  - Nutzer ist kooperativ
  - Nutzer ist Muttersprachler
- Designaufgabe unter Einbeziehung potentieller Nutzer
  - Wizard-of-Oz-Experimente
  - Prototyping
  - nutzerbezogene Evaluation (Dialogerfolgsrate)

Ergonomische Designkriterien

- Flexibilität und Liberalität
  - maximalen Formulierungsspielraum für den Nutzer anstreben
- Verhandlungsmöglichkeiten:
  - Nutzer kann Vorschläge des Systems annehmen oder ablehnen
  - Nutzer kann Constraints im Problemraum zurücknehmen
- Navigationsfähigkeit
  - System muss Veränderungen in der aktuellen Zielstellung erkennen können
  - Fragen zum Leistungsumfang des Systems müssen möglich sein

Ergonomische Designkriterien

- Verteilung der Initiative:
  - systemgesteuerter (geschlossener) Dialog
    - Nutzer beantwortet Fragen nach strikt vorgegebenem Schema ("Verhör")
  - Dialog mit verteilter Initiative (mixed-initiative, offener Dialog)
    - Nutzer kann die Initiative ergreifen
    - (relativ) freie Spracheingabe
    - stark komprimierte Nutzeranfragen
    - barge in: Unterbrechen des Systems
  - Mischformen
    - z.B. Initiative im allgemeinen Fall beim Nutzer, System übernimmt bei erkennbaren Problemen die Steuerung

Ergonomische Designkriterien

- Kontakt mit dem Nutzer
  - ständig Orientierung über Dialogzustand und Systemannahmen gewährleisten
  - direkte Antworten geben
  - Bestätigung von Erkennungsergebnissen bei geringer Konfidenz
  - Klärungsdialoge
  - Entscheidung über Abbruch $\rightarrow$ Weiterleitung an den Operator

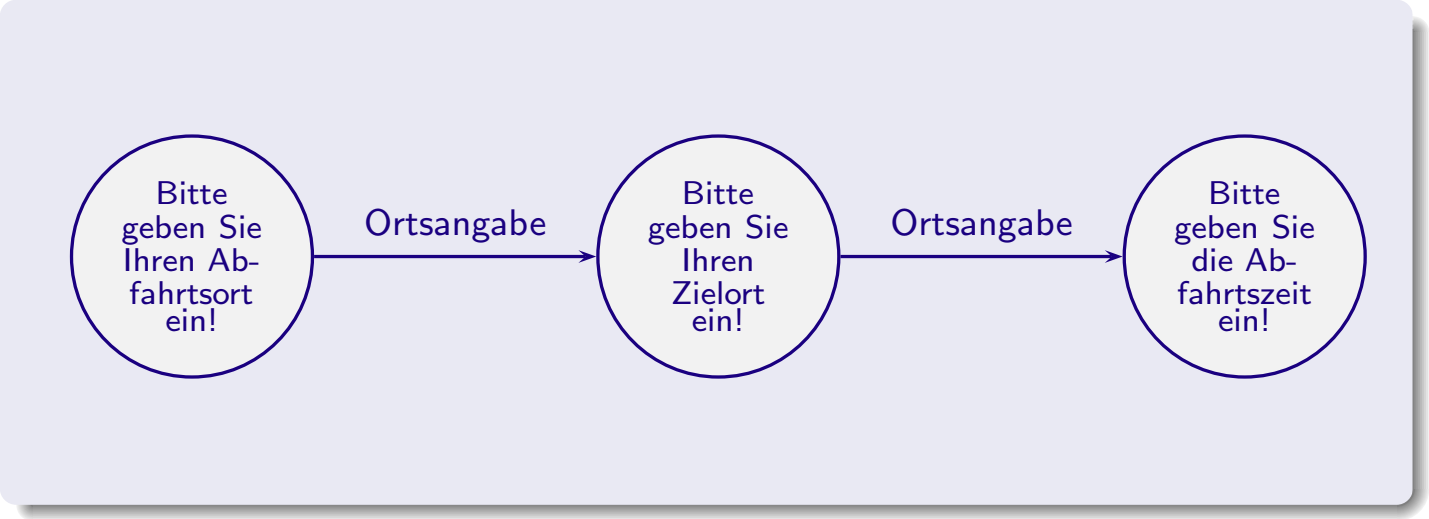
Geschlossene Dialoge

- Prädiktion möglicher (zulässiger) Nutzerreaktionen → temporäre Reduzierung des Wortschatzes
- Modellierung durch endliche Automaten
  - Dialogzustände: Aktionen, z.B. Aufforderung zur Eingabe (Prompt)
  - Übergänge zwischen Dialogzuständen: Erkennung von Nutzeräußerungen
- Beispiel: VoiceXML

Dialogmodellierung

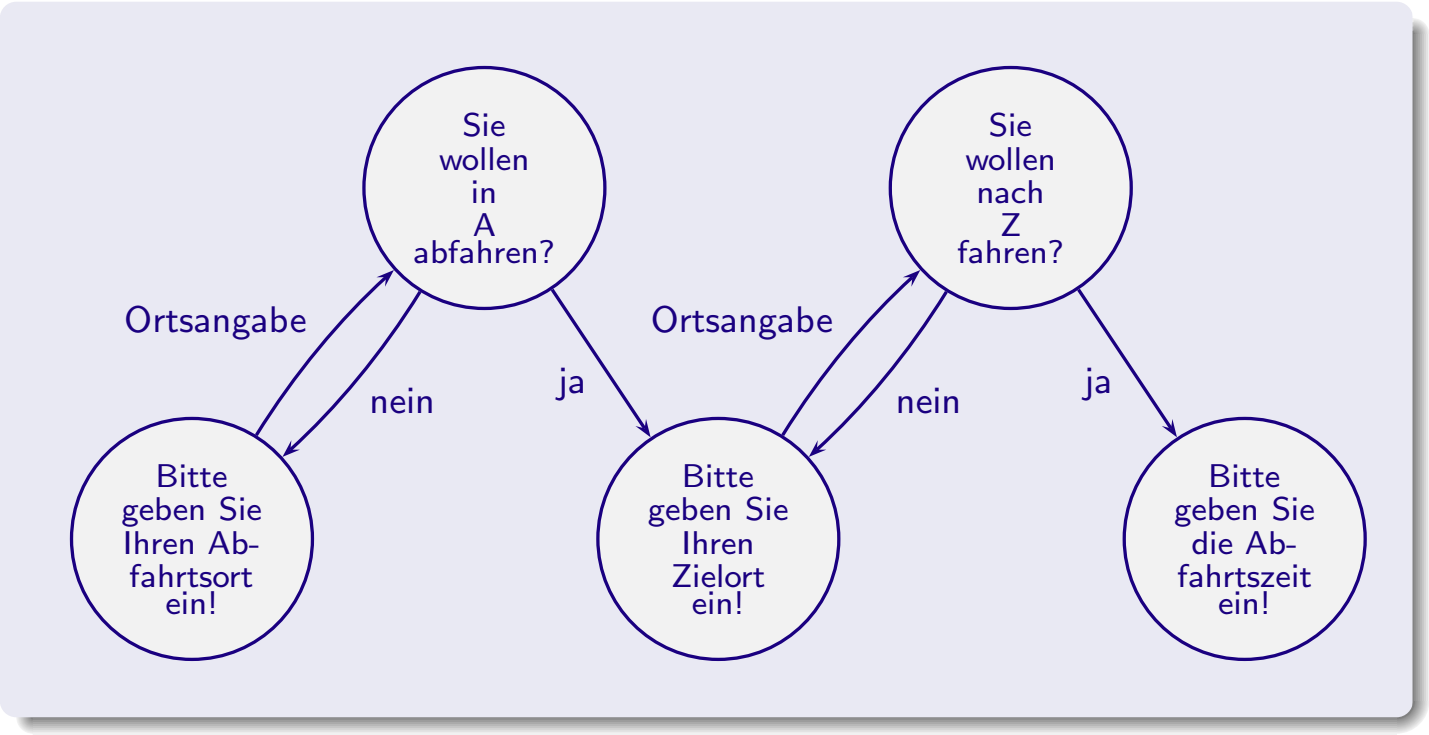
Dialogmodellierung

- Mehrfachverwendung von Teilnetzen



Dialogmodellierung

- Erhöhen der Zuverlässigkeit durch Rückfragen



VoiceXML

- Abspielen von Audiosamples

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<vxml version="2.0" lang="en">
<form>
 <block>
  <prompt bargein="false">Welcome to Travel Planner!
   <audio src="http://www.adline.com/mobile?code=12s4"/>
  </prompt>
 </block>
</form>
</vxml>
```


VoiceXML

- akustisches Auswahlmenü

```
<?xml version="1.0"?>
<vxml version="2.0">
<menu>
  <prompt>
    Say one of: <enumerate/>
  </prompt>
  <choice next="http://www.sports.example/start.vxml">
     Sports
  </choice>
  <choice next="http://www.weather.example/intro.vxml">
     Weather
  </choice>
  <choice next="http://www.news.example/news.vxml">
     News
  </choice>
  <noinput>Please say one of <enumerate/></noinput>
</menu>
</vxml>
```

VoiceXML

- Dialogbeispiel

System:	Say one of: Sports; Weather; News.
Human:	Astrology
System:	I did not understand what you said.
	(a platform-specific default message.)
System:	Say one of: Sports; Weather; News.
Human:	Sports
System:	*proceeds to* `http://www.sports.example/start.vxml`

VoiceXML

- Forms: Slotstrukturen, die mit Werten zu füllen sind

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<vxml version="2.0" lang="en">
<form>
  <field name="city">
  <prompt>Where do you want to travel to?</prompt>
   <option>London</option>
   <option>Paris</option>
   <option>Stockholm</option>
 </field>
 <field name="travellers" type="number">
  <prompt>
   How many persons are travelling to <value expr="city"/>?
  </prompt>
 </field>
 <block>
  <submit next="http://localhost/handler"
      namelist="city travellers"/>
 </block>
</form>
</vxml>
```

VoiceXML

- Variation von Systemprompts bei Wiederholung

```
<field name="travellers" type="number">
 <prompt count="1">
  How many are travelling to <value expr="city"/>?
 </prompt>
 <prompt count="2">
  Please tell me the number of people travelling.
 </prompt>
 <prompt count="3">
  To book a flight, you must tell me the number
  of people travelling to <value expr="city"/>.
 </prompt>
 <nomatch>
  <prompt>Please say just a number.</prompt>
  <reprompt/>
 </nomatch>
</field>
```


VoiceXML

- Werteüberprüfung

```
<field name="travellers" type="number">
 <prompt>
  How many are travelling to <value expr="city"/>?
 </prompt>

 <filled>
  <var name="num_travellers" expr="travellers + 0"/>
  <if cond="num_travellers > 12">
    <prompt>
     Sorry, we only handle groups of up to 12 people.
    </prompt>
    <clear namelist="travellers"/>
  </if>
 </filled>

</field>
```

VoiceXML

- Auswertung von Konfidenzwerten

```
<field name="city">
 <prompt>Which city?</prompt>
  ...
 <filled>
  <if cond="city$.confidence < 0. 3">
   <prompt>Sorry, I didn't get that</prompt>
   <clear namelist="city"/>
  <elseif cond="city$.confidence < 0.7"/>
   <assign name="utterance" expr="city$.utterance"/>
   <goto nextitem="confirmcity"/>
  </if>
 </filled>
</field>
<subdialog name="confirmcity" src="#ynconfirm" cond="false">
 <param name="user_input" expr="utterance"/>
 <filled>
  <if cond="confirmcity.result=='false'">
   <clear namelist="city"/>
  </if>
 </filled>
</subdialog>
```


VoiceXML

- Einbinden externer Grammatiken

```
<form name="trader">

 <field name="company">
  <prompt>
   Which company do you want to trade?
  </prompt>
  <grammar src="trade.xml#company"
     type="application/grammar+xml"/>
 </field>

 <field name="action">
  <prompt>
    do you want to buy or sell shares in
    <value expr="company"/>?
  </prompt>
  <grammar src="trade.xml#action"
    type="application/grammar+xml"/>
 </field>
</form>
```

VoiceXML

- CF-Grammatiken, Regelimport

```
<grammar xml:lang="en">
<import uri="http://please.com/politeness.xml"
  name="polite"/>
<rule name="command" scope="public">
 <ruleref import="polite#startPolite"/>
 <ruleref uri="#action"/>
 <ruleref uri="#object"/>
 <ruleref import="polite#endPolite"/>
</rule>
<rule name="action" scope="public">
 <choice>
  <item tag="buy"> buy </item>
  <item tag="sell"> sell </item>
 </choice>
</rule>
<rule name="company" scope="public">
 <choice>
  <item tag="ericsson"> ericsson </item>
  <item tag="nokia"> nokia </item>
 </choice>
</rule>
```


VoiceXML

- Übergang zu einfacheren Dialogkonstrukten im Problemfall

```
<form name="trader">
 <grammar src="trade.xml#command"
  type= "application/grammar+xml"/>
 <initial name="start">
  <prompt>What trade do you want to make?</prompt>
  <nomatch count="1">
   <prompt>
    Please say something like 'buy ericsson';
   </prompt>
   <reprompt/>
  </nomatch>
  <nomatch count="2">
   Sorry, I didn't understand your request.
   Let's try something simpler.
   <assign name="start" expr="true"/>
  </nomatch>
 </initial>
 <field name="company"> ... </field>
 <field name="action"> ... </field>
</form>
```

Offene Dialoge

- Analysefähigkeit statt Prädiktion
  - robuste Verarbeitung
- Erkennung des Dialogzustands
  - Aktivierung spezifischer Sprachmodelle
- Dialoggedächtnis
  - Detektion von Inkonsistenzen → Klärungsdialoge
- unerwartete Übernahme der Initiative durch den Nutzer (barge in)
  - Hauptgrund: Langatmigkeit von Systemprompts
  - technische Reaktion
    - Echounterdrückung einschalten
    - Sprachsynthese ausblenden
    - Dialogzustand mit maximaler Erwartung einnehmen

Erkennung fließender Sprache

- Fließende Sprache
- Sprachmodelle
- Fehlermaße
- Dialogmodelle
- Training und Erkennung
- Behandlung unbekannter Wörter
- Sprecherunabhängigkeit

Training

- Initialisierung der Phonmodelle
- Vortrainieren der Phonmodelle
  - Sprachsignal in Lautabschnitte segmentiert
  - Zuordnung von Phonsymbolen zu den Lautsegmenten
  - auch Bootstrapping: segmentieren mit schlechten Modellen, dann neu trainieren
- Training der Phonmodelle
  - unsegmentiertes Sprachsignal
  - Zuordnung einer Wortkette zum Sprachsignal
  - Verketten der Phonmodelle zu Wortmodellen
  - Verketten der Wortmodelle zu einem Gesamtmodell für das Trainingsmaterial
  - Baum-Welch-Training
  - beste Zuordnung der Wortgrenzen wird automatisch gefunden

Erkennung

- Wortgrenzen sind unbekannt: jedes Wort kann überall beginnen
- Viterbi-Annahme: Übergang im Markov-Modell erfolgt immer zu einem Knoten mit einem höheren Index
  - durch rekursive Verkopplung der Wortmodelle nicht mehr gültig
  - gesonderte Behandlung der Wortübergänge
  - $\rightarrow$ time-synchroneous Viterbi

Erkennung

- Suchraumeinschränkungen
  - beam search
    - Schwellwert für die Abweichung von der (momentan) wahrscheinlichsten Lösung
    - 80 - 90% des Suchraums kann unberücksichtigt bleiben, ohne Qualitätsverlust
  - level building
    - zuerst Ermitteln der besten Worthypothesen pro Zeitframe
    - Erweitern der Hypothesen auf 2, 3, ... Wortformen entsprechend den Restriktionen der Grammatik
    - nicht zeitsynchron

Erkennung

- Suchprobleme
  - VITERBI-Algorithmus ermittelt den optimalen Pfad, nicht die optimale Wortsequenz (nur Approximation)
- Alternative: stack decoding
  - wortweise links-rechts-Erweiterung von Pfaden (auf dem Stack)
  - dadurch eindeutige Vergangenheit jeder temporären Lösung garantiert
  - Entfaltung des Suchraums (Graph) als Baum
  - Probleme
    - finden einer geeigneten Pfadbewertung, die die Hypothesenmenge auf dem Stack wirksam genug einschränkt
    - Normierung der Bewertung auf die Pfadlänge (Pfade sind unterschiedlich lang)

Erkennung

- Ausbalancieren von akustischen und Sprachmodellen
- akustische Modelle sind unterbewertet
  - Übergang im Wort:

    $$p(x[1:t-1]) \cdot p_t(z_t|z_{t-1}) \cdot p_e(x|z_t)$$

  - Übergang zwischen zwei Wörtern:

    $$p(x[1:t-1]) \cdot p_t(w_i|w_{i-1})$$

  - Faktor m zur Anpassung der unterschiedlichen Modellwahrscheinlichkeiten:

    $$p(x[1:t-1]) \cdot p_t(w_i|w_{i-1}])^m$$

    Strafmaß für das Verlassen eines Wortmodells

Erkennung fließender Sprache

- Fließende Sprache
- Sprachmodelle
- Fehlermaße
- Dialogmodelle
- Training und Erkennung
- Behandlung unbekannter Wörter
- Sprecherunabhängigkeit

Behandlung unbekannter Wörter

- jedes endliche Wörterbuch ist unvollständig
- Viterbi-Suche zwingt aber zur Interpretation jedes Signalabschnitts durch ein Modell
  - sinnlose und unvorhersehbare Zuordnungen bei der Verwendung unbekannter Wortformen
- HMM's für unbekannte Wörter
  - Ziel: hohe *a posteriori* Wahrscheinlichkeit $p(w|x)$ für unbekannte Wortformen $w \notin L$
  - niedriges $p(w|x)$ für alle $w \in L$

Behandlung unbekannter Wörter

- Kopplung von Phonmodellen (Asadi, Schwartz und Makhoul 1991)
  1. beliebige Sequenz aus 5 Phonen
  2. beliebige Sequenz aus mindestens 3 Phonen
  3. beliebige Sequenz aus mindestens 5 Phonen
  4. beliebige Sequenz von links-kontextsensitiven Phonen

  2. ist beste Variante
- experimentelle Resultate
  - Erkennung 74%
  - Fehlalarm 3.4%
  - bei Grammatik mit Perplexität $Q = 100$
- auch: explizite Modellierung der phonotaktischen Regularitäten
- auch: Verkopplung der Modelle nach Triphonstatistik

Erkennung fließender Sprache

- Fließende Sprache
- Sprachmodelle
- Fehlermaße
- Dialogmodelle
- Training und Erkennung
- Behandlung unbekannter Wörter
- Sprecherunabhängigkeit

Sprechervarianz

- individuelle Besonderheiten der Stimmlippen
  - Grundfrequenz
- anatomische Parameter des Vokaltrakts
  - Spektrale Lautcharakteristika (Formanten, Antiformanten)
- Sprechtempo
- Dialekt
- intraindividuelle Varianz: sprachliche Fitness
- Fehlerrate erhöht sich um Faktor 3 . . . 5

Sprecherunabhängigkeit

- Sprechervarianz
- Sprecherunabhängige akustische Modelle
- Sprecherselektion
- Sprecheradaption

Sprecherunabhängige akustische Modelle

- mutiple Sprecher vs. Sprecherunabhängigkeit
  - multiple Sprecher: Testsprecher = Trainingssprecher
  - sprecherunabhängig: Testsprecher $\neq$ Trainingssprecher
- Training robuster Modelle
  - Training eines Modells mit den Sprachdaten verschiedener Sprecher
  - flache Emissionsverteilungen
  - Daten von mehreren Sprechern $\rightarrow$ viele Daten

Sprecherunabhängige akustische Modelle

- Verwendung multipler Kodebücher
  - Unterteilung der Merkmalsvektoren in n Abschnitte
    - LPC-Koeffizienten
    - Δ-LPC-Koeffizienten
    - Energie und Δ-Energie
  - separate Vektorquantisierung für die einzelnen Abschnitte
  - Annahme: die Emission der Vektorabschnitte ist unabhängig
    $\rightarrow$ Emissionswahrscheinlichkeit berechnet sich aus dem Produkt der Einzelwahrscheinlichkeiten
  - Vorteile
    - Reduktion des Quantisierungsfehlers
    - m^n Vektorkombinationen bei $m \cdot n$ Parametern
      $\rightarrow$ erheblicher Zuwachs an Dynamikumfang bei moderat erhöhter Parameterzahl
    - Multiplikation von n relativ flachen Emissionsverteilungen
      $\rightarrow$ Kontrastverschärfung

Sprecherselektion

- Einteilung der Sprecher in Cluster
  - Extremfall: jeweils nur ein Sprecher pro Cluster
- Clusterspezifische Modelle
  - Kodebücher der Vektorquantisierung
  - Emissionswahrscheinlichkeiten
  - Transitionswahrscheinlichkeiten
- Clustern von stochastischen Modellen
  - Iteration
    - paarweises Mischen der Modellparameter
    - Messen des Entropieverlustes
    - Zusammenfassen der beiden Modelle mit dem geringsten Entropieverlust
  - Vereinfachung: Ignorieren der Transitionswahrscheinlichkeiten
  - Paarweiser Elementeaustausch, solange dies zu einer Entropieerhöhung führt
- Anzahl der Cluster ist durch die verfügbaren Trainingsdaten begrenzt

Sprecherselektion

- Abhängigkeit der Sprecherclusterung vom Geschlecht der Sprecher (Lee 1989)

7	M	10	6	0	0	26	20	13
	F	2	1	7	20	0	0	0
6	M	10	6	0	26	20	13	
	F	2	1	27	0	0	0	
5	M	10	0	27	21	17		
	F	2	27	0	0	1		
4	M	10	0	23	42			
	F	2	27	0	1			
3	M	0	33	42				
	F	28	1	1				
2	M	0	75					
	F	30	0					

Sprecherselektion

- Sprecheridentifikation
  - mit vereinbarter Identifikationsäußerung (Testsatz mit zugehöriger Wortkette)
    - pro Sprechercluster: Bilden eines Modells für die Wortkette
    - Berechnen der Vorwärtswahrscheinlichkeit für jedes Sprechercluster
    - Maximumselektion
  - ohne vereinbarte Identifikationsäußerung
    - Erkennen der ersten Äußerung mit sprecherunabhängigen Modellen
    - Bestätigen einer korrekten Erkennung
    - manuelle Korrektur einer fehlerhaften Erkennung

- Lee 1989: kaum Verbesserung gegenüber sprecherunabhängigen Modellen

Sprecheradaption

- individuelle Modifikation eines sprecherunabhängigen Modells
  - offline: separate Trainingsdaten zur Adaption
    - überwachtes Training
      $\longrightarrow$ sprecherabhängige aber stark untertrainierte Modelle
    - nach 10 ... 30 Sätzen 3 ... 5 % höhere Erkennungsrate
  - offline: Berechnung einer Abbildungsvorschrift aus dem Langzeitspektrum eines Sprechers
    - Transformation des Merkmalsraums
  - online: Nachführung während der laufenden Erkennung
    - nichtüberwachtes Training

Grundlagen der Sprachsignalerkennung

- Spracherkennung als technisches und akademisches Problem
- Sprachsignal: Erzeugung, Wahrnehmung und Beschreibung
- Merkmalsextraktion
- Worterkennung
- Phon-basierte Worterkennung
- Erkennung fließender Sprache
- Systemarchitekturen

Systemarchitekturen

- Sprachverstehende Systeme
- Monolithische Systeme
- Modulare Systemarchitekturen
- Bimodale Spracherkennung

Sprachverstehende Systeme

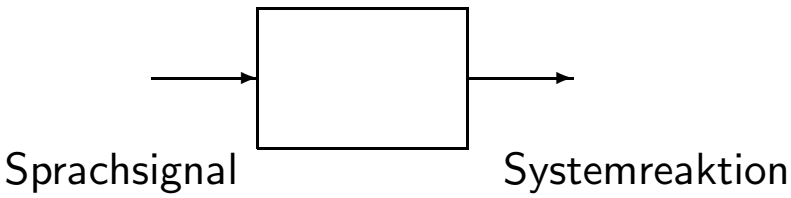
- Wortkette ist nur im Ausnahmefall Endresultat der Verarbeitung gesprochener Sprache
  - Diktiergerät
- meist inhaltliche Interpretation gewünscht:
  - Zugriff zu Informationssystemen
  - Zugriff zu technischen Gerätesteuerungen
  - Dolmetschen
- syntaktisch-semantische Analyse erforderlich
  - Analyse bei Erkennungsunsicherheit
    - Korrekturmechanismen
    - Verwaltung alternativer Hypothesen
    - unterspezifizierte Erkennungsresultate

Systemarchitekturen

- Sprachverstehende Systeme
- Monolithische Systeme
- Modulare Systemarchitekturen
- Bimodale Spracherkennung

Monolithische Systeme

- unmittelbare Abbildung Sprachsignal → Systemreaktion



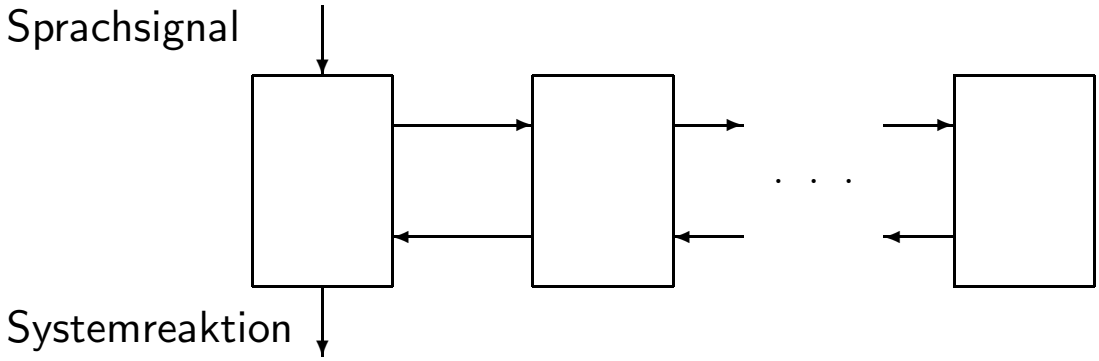
- Einzelworterkennung
  - eindeutige Zuordnung von Systemreaktionen zu Wortformen
  - direkte semantische Interpretation

  → nur für atomare Systemreaktionen (Kommandointerpreter)
- Strukturierte Beschreibungen der gewünschten Systemreaktion
  - Datenbankquery
  - Steuerprozeduren (z.B. Werkzeugmaschine, Roboter)

  → Lernen der Datenbankabfrage ???

Monolithische Systeme

- kaskadierte Architekturen



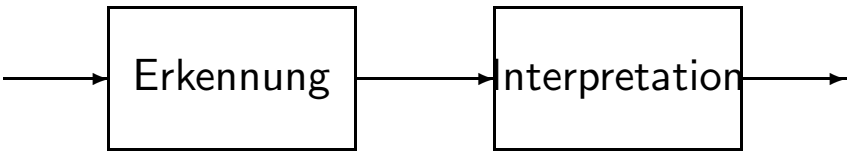
- Umsetzung erfolgt immer noch in einem Schritt

Systemarchitekturen

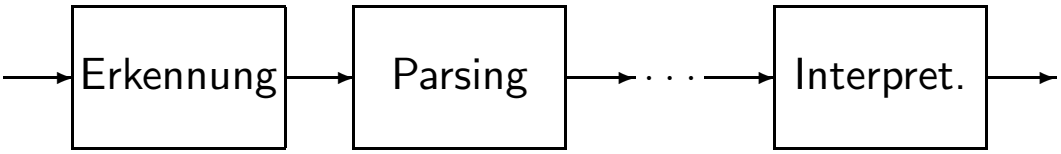
- Sprachverstehende Systeme
- Monolithische Systeme
- Modulare Systemarchitekturen
- Bimodale Spracherkennung

Modulare Systemarchitekturen

- sequentielle Architekturen mit *feed forward*-Schnittstellen
- zweistufig: Erkennung $\longrightarrow$ Interpretation



- mehrstufig:

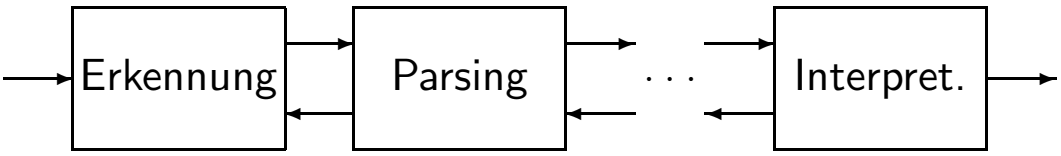


Modulare Systemarchitekturen

- Problem: Erkennungsunsicherheit
  - Übergabe der besten Wortkette
  - Übergabe der n besten Wortketten
    - heuristische Selektion
  - Wortgraphen
    - erweitertes Chart-Parsing

Modulare Systemarchitekturen

- Sprachverstehen ist ein erwartungsgesteuerter Prozeß
- sequentielle Architekturen mit bidirektionaler Interaktion



- Probleme:
  - Kombinatorik über mehrere Ebenen hinweg
  - Prosodie
  - Semantik
  - Domänenmodellierung
  - Dialogmodellierung
  - flexible Wahl der Erkennungseinheiten

Modulare Systemarchitekturen

- vernetzte Architekturen
  - Blackboard-Architektur
    - Hearsay-II (Reddy et al. 1973)
    - Engpaß: zentraler Scheduler

  → dezentrale Architekturen

Systemarchitekturen

- Sprachverstehende Systeme
- Monolithische Systeme
- Modulare Systemarchitekturen
- Bimodale Spracherkennung

Bimodale Spracherkennung