Sprachkommunikation

1 Motivation und Zielsetzung

1.1 Was ist Sprache

Sprachlaut: Segment, in das sich eine sprachliche Äußerung auditiv zerlegen lässt.
Phonem: Kleinste bedeutungsunterscheidene, aber nicht selbst bedeutungstragende Enheit
Morphem: Kleinste selbst **bedeutungstragende Einheit einer Sprache**

Syntax: Beziehung der Zeichen untereinander **(Regeln bspw. Pluralbildung)**

Semantik: Lehre von der **Bedeutung** der sprachlichen Zeichen
Pragmatik: Beziehung zwischen den Zeichen und ihrem Benutzer **Weltwissen hinter den sprachlichen Äußerungen**

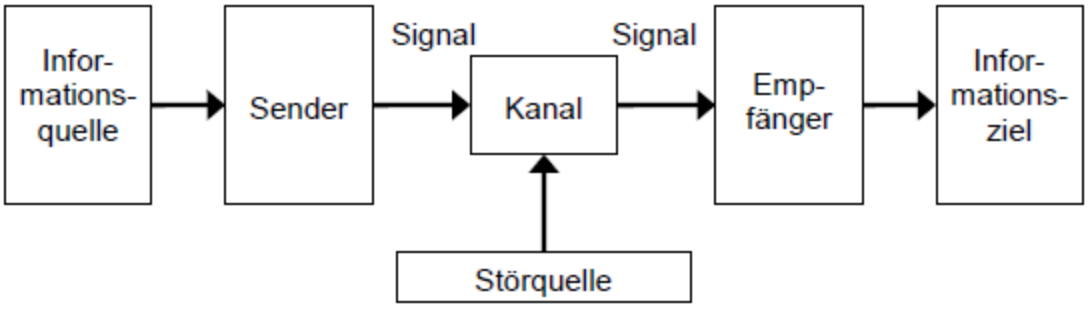
Prosodie
- Quantität (Lautdauer)
- Intensität oder Akzentuierung (Betonung Akzent)
- Intonation (melodische Aspekte, Haltung und Emotionen)

Dauer, Amplitude, Grundfrequenz

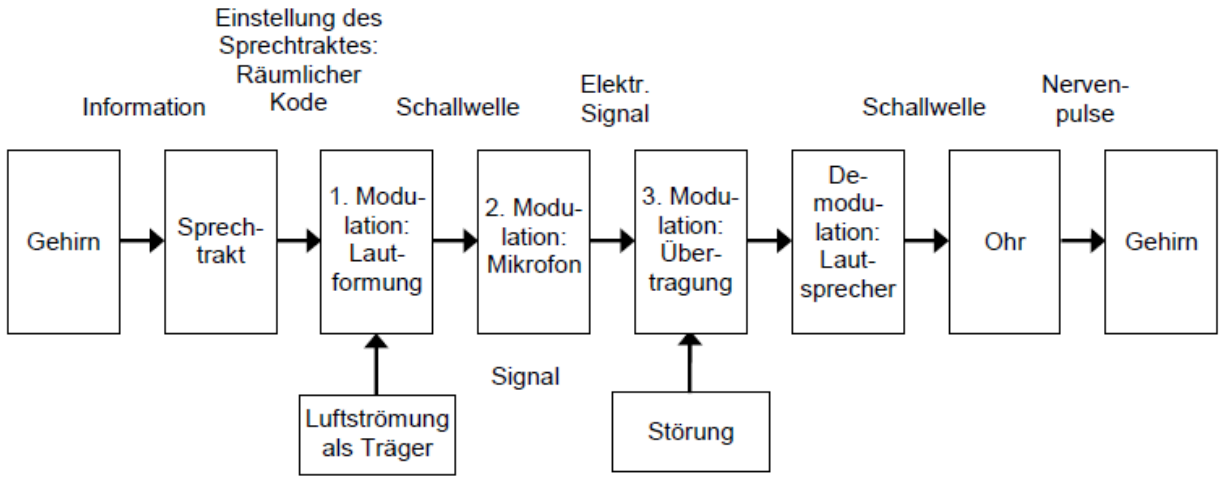
1.2 Anwendung

Vorteile:
+ Das Kommunikationsmedium des Menschen
+ intuitiv und natürlich
+ keine speziellen Kenntnisse
+ Hands-Busy-Eyes-Busy geeignet
+ Sehbehinderte
+ von jedem Ort anwendbar

1.4 Kommunikationsmodelle



Kommunikationsmodell nach Shannon und Weaver (1949).



Erweitertes Kommunikationsmodell, vgl. Heute (1990).

Störungen
- Ströung des Übertraungskanals
- Störung der Sprachproduktion (Schlaganfall)
- Störung der Sprachrezeption (Schwerhörigkeit)
- Kein gemeinsames Zeichensystem

2 Sprachsignaldarstellung und -eigenschaften

Mikrophon nimmt analoges Signal auf -> Zeit und Amplitude sind kontinuierlich. Muss in ein Zeit und Werte diskretes Signal überführt werden.

2.1 Darstellung kontinuierlicher Signale im Zeitbereich

Sprache ist *quasi-periodisch* und *nicht-periodisch*
Gesehen auf 20ms Konstant

In der Kommunikationstechnik haben wir es häufiger mit Systemen zu tun, die sich durch Differentialgleichung 1. Ordnung mit konstanten Koeffizienten berschreiben lassen. Solche Systeme verhalten sich linear und zeitinvariant, d.h. es gelten der **Überlagerungssatz** und der **Verschiebungszsatz**. *-> lineares, zeitinvariantes System*.

2.3 Darstellung kontinuierlicher Signale im Frequenzbereich, Spektrum

Fourier-Reihe:
Als Fourierreihe bezeichnet man die Reihenentwicklung einer **periodischen, abschnittsweise stetigen Funktion** in eine Funktionenreihe aus Sinus- und Kosinusfunktionen.

Fourier-Transformation
Die Fourier-Transformation (genauer die kontinuierliche Fourier-Transformation; Aussprache: fuʁie) ist eine Methode der Fourier-Analysis, die es erlaubt, **kontinuierliche, aperiodische Signale in ein kontinuierliches Spektrum zu zerlegen.**
-> kontinuierlichen Spektrum
Enthält Betragsanteil und Phasenateil -> komplexes Spektrum, wird aber meist getrennt dargestellt.

- Ein periodisches Signal besitzt ein diskretes Spektrum.
- Ein diskretes Signal besitzt ein sich periodisch wiederholendes Spektrum.

Symmetrie-Satz (Kann hin und rucktransformiert werden)
Verschiebungssatz
Faltungs- und Multiplikationssatz (Faltung im Zeitbereich ist gleich Multiplikation im Frequenzbreich)
Satz von Parseval (Energie im Zeitbereich ist gleich der Energie im Frequenzbreich)

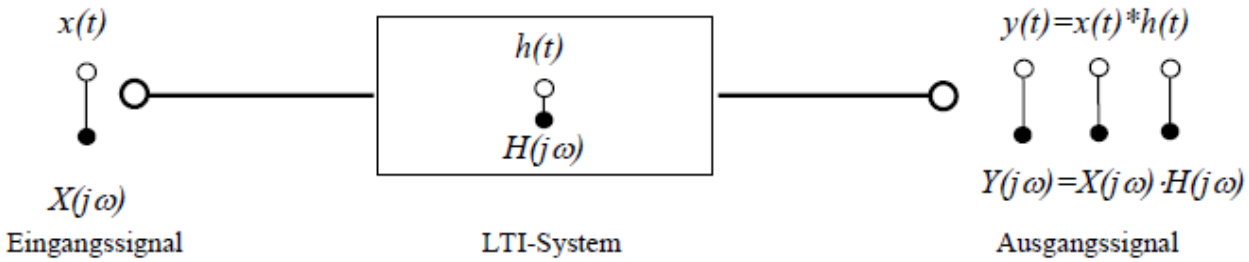
2.4 Impulsantwort und Übertragungsfunktion

Wenn ein LTI System mit einem Dirac Impuls angeregt wird, ist das entstehende Ausgangssignal $h(t)$ characteristisch für das System. $h(t)$ *Impulsantwort*
Faltungsintegral:

$$y(t) = \int_{-\infty}^{\infty} x(\tau) \cdot h(t-\tau) d\tau$$

$$= x(t) * h(t)$$

Faltung ist kommutativ!



2.5 Statistische Beschreibung von Sprachsignalen

Kumulative Verteilungsfunktion (bestimmen ob x unterhalb einer bestmmten schranke bleibt)
Verteilungsdichtefunktion (X in einem bestimmten intervall (integrieren des intervalls, besonders gut für kontinuierliche zufallsvariablen)

Auto-Korrelationsfunktion beschreibt die Ähnlichkeit zwischen den Signal x(t) und dem tau verschobenen signal x(t+tau)

Kreuz-Kovarianz = 0 dann sind die signale unkorreliert

2.6 Energiedichte- und Leistungsdichtespektrum

$|X(j\omega)|^2$ *Energiedichtesprektrum* -> Rücktransformiert ist das die Autokorrelationsfunktion
Nur für endliche Energie

Bei unendlicher Energie aber endlicher Leistung -> grenzwertbildung ->
Leistungsdichtespektrum

2.7 Darstellung diskreter Signale im Zeit- und Frequenzbereich

Zeit- und Wertdiskret
+ Störsicherheit
+ Universalität
+ Einfache Verarbeitung
- Quantisierungsfehler
- höherer Bandbreitenbedarf (kann durch intelligente Kodierung ausgeglichen werden)

Durch Abtastung wird ein kontinuierliches Signal Zeitdiskret. Muss mit mindestens der doppelten Frequenz abgetastet werden, wie die höchste im Signal vorhandene Frequenz.
Wenn dies nicht geschieht entstehen im Spektrum durch die wiederholung alias effekte, da sich krams überlagert.
Aliasfilterung

2.8 Langzeit- und Kurzzeit-Signaleigenschaften

Sprache zwischen 0-4kHz
Grundfrequenz bei Männer 125Hz bei Frauen 250Hz.
Formanten characteristisch für den Sprecher.

2.9 Spektrogramm

gleitende Fourier Transformation

kurzzeitsprektren dargestellt über die Zeit

Gezeigt wird Frequenz, Zeit und Leistungsdichte

2.10 Amplitudenverteilung

sprache häufig null. und lange pausen 60 % schweigen pro person Telefon

3 Grundlagen der menschlichen Spracherzeugung

Sprache wird über rückkopplung gesteuert. (Akkustisch und Sprech und Atmungsmuskulatur)

3.1 Anatomie des menschlichen Sprechapparates

Lunge liefert energie /luftstrom
Simmbänder sind ein schwingungsfähiges System
Vokaltrakt besteht aus Mund Rachen und Nasenraum bilden einen Resonanz Körper

3.2 Anregung

1 Periodische Anregung -> Glottis geht ryhtmisch auf und zu -> generiert dreiecksignale
2a Aperiodisch glottis bleibt geöffnet luftstrom wid aber turbulent
2b Aperiodisch glottis offen aber andere stelle im vokaltrakt ist geschlossen. (Plosives)

3.3 Lautformung

Artikulation Organe -> Lippen, Unterkiefer, Zungengifel, Gaumen
-> Röhrenförmiger Raum mit veränderbaren Querschnitt

3.4 Sprachlaute

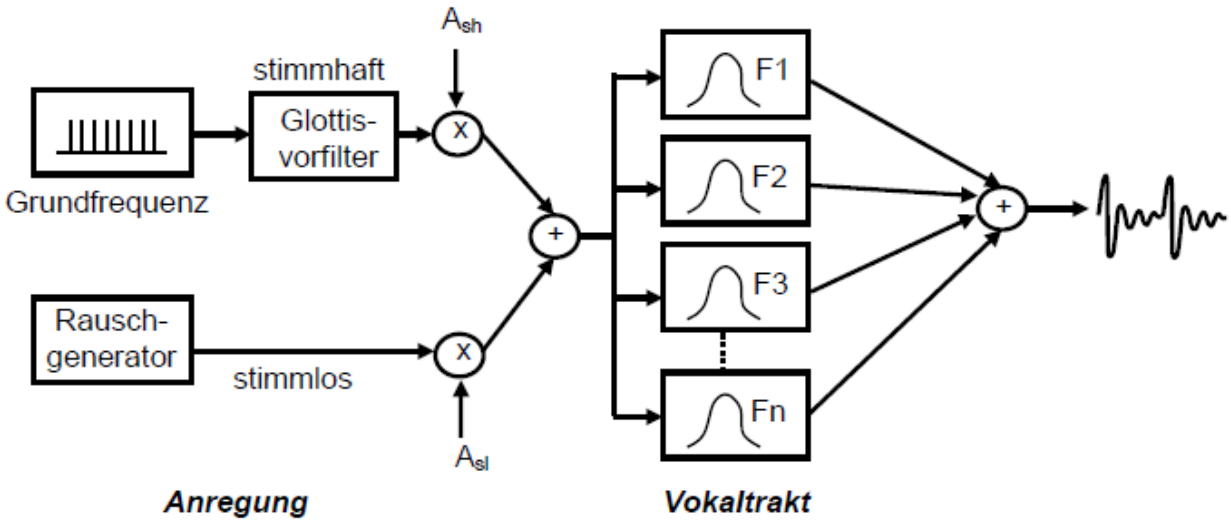
Vokale und Umlaute -> Periodische Anregung (Stimmhaft)
Konsonanten -> Verschlußlaute (b,p,d,t), Engelaute (z,s,w,f), Zitterlaute (rrr), Öffnungshlaute (h), Nasenlaute (m,n)

höhere Formaten sind relativ konstant, und können zur Sprecherekennung verwendet werden.

3.5 Modelle der Spracherzeugung

Vokaltrakt als akustische Röhre mit veränderlichem, aber stückweise konstanten Querschnitt.
-> Webstersche Differentialgleichung
Am Mund Druck = 0
An der Glottis Schnelle = 0

Quelle Filter Modell



Parallelschaltung einzelner Teilfilter für jeden formanten ein filter
-> Beinhaltet übrigens keine Sprunganregung
-> Anregung findet nimmer am unteren Endes Vokaltraktes statt

Alternativ Filter in reihe, dann entspricht dem Röhrenmodell

Leitungsmodell komplexer, stellt daher auch mehr da ;)

4 Sprachanalyse

4.1 Spektralanalyse

Durchstimmbarer Bandpass
-> Interessant wenn Filter konstanter relativer Bandbreite verwendet werden
Da Ähnlich wie frequenzanalyse wie im Ohr funktioniert.
Man kann das Sperrverhalten verbessern, indem man die Rechtecktfunktion ersetzt durch eine allgemeine Gewichtung. Hamming Hann Fenster

DFT ist eine spezielle Filterbank mit M äquidistanten Kanälen, einer Unterabtastung um den Faktor M, sowie mit speziellen Filtern, die als Impulsantwort eine Rechteckfuntkion aufweisen.

M^2 Komplexe addition und Multiplikation
FFT-> (Eine Gerade und ungerade k-werte berechnen)
- Nutzt periodizität der Koeffizienten aus.
- M muss grade Zahl sein
- Am besten Zweierpotenz
- Radix-2/Decimation in time algorithmus

4.2 Cepstrum

G ausgangssignal des Vokaltraktfilters
S Anregungsfunktion
H übertragungsfunktion

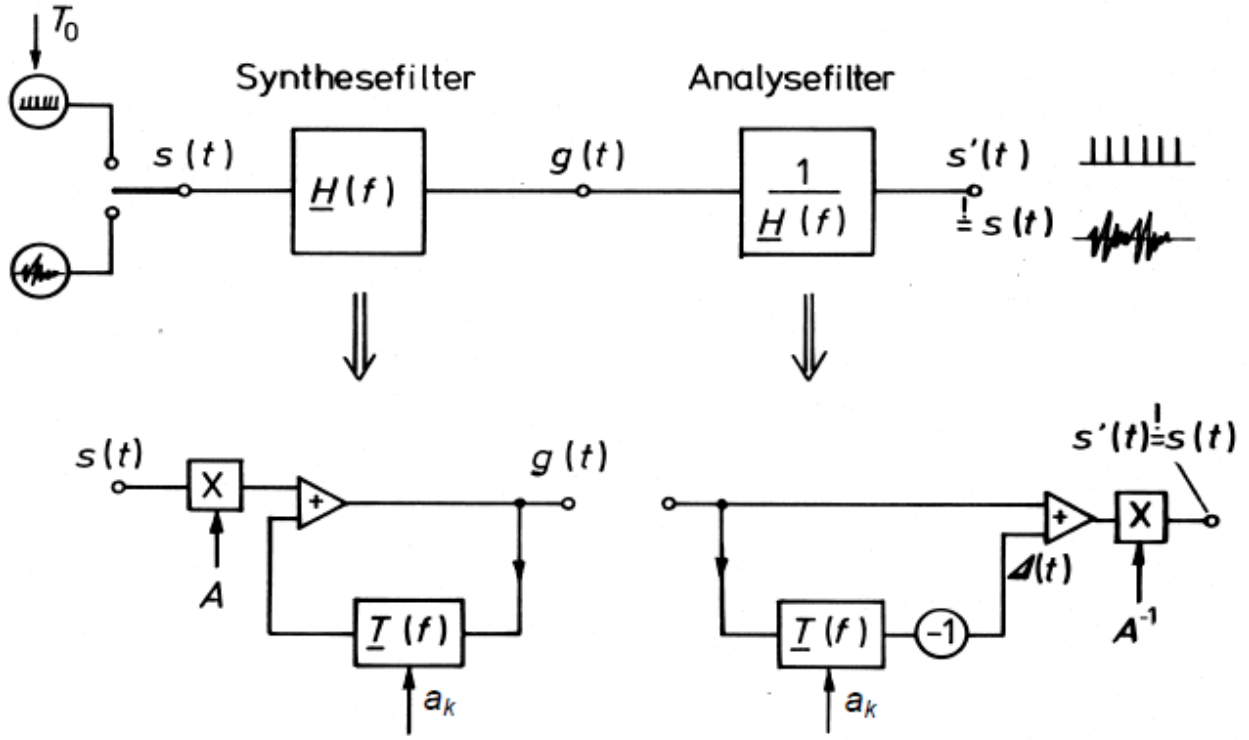
Durch Rück-Transformation aus dem Frequenz- in einen (weiteren, da logarithmischen) Zeit-Bereich erhält man das sog. *Cepstrum* von *g(t)*:

$$\begin{aligned} C(x) &= F\{\ln|G(j\omega)|\} \\ &= F\{\ln|S(j\omega)| + \ln|H(j\omega)|\} \\ &= C_1(x) + C_2(x) \end{aligned} \qquad (4.20)$$

Durch den Logarithmus als nichtlineare Operation wird nämlich aus dem Produkt zwischen Eingangsspektrum und Übertragungsfunktion des Vokaltraktes eine reine Addition; sofern sie sich nicht überdecken, können die Anteile also leicht (mittels eines liftering) getrennt werden. Das heißt, dass durch Filterung im Quefrency-Bereich Anregungssignal und Vokaltraktfilter getrennt werden können. Diese Eigenschaft kann man z.B. zur Bestimmung der Grundfrequenz, die ja nur im Anregungssignal vorkommt, zur Stimmhaft-Stimmlos-Entscheidung, sowie zur Formantbestimmung ausnutzen.

4.3 Lineare Prädikation

Die Idee ist folgende: Schickt man das aus dem Vokaltrakt hervorgegangene Signal durch ein zum Vokaltrakt inverses Filter, so lässt sich am Ausgang dieses inversen Filters das Anregungssignal messen (das im Menschen so nicht messbar ist).

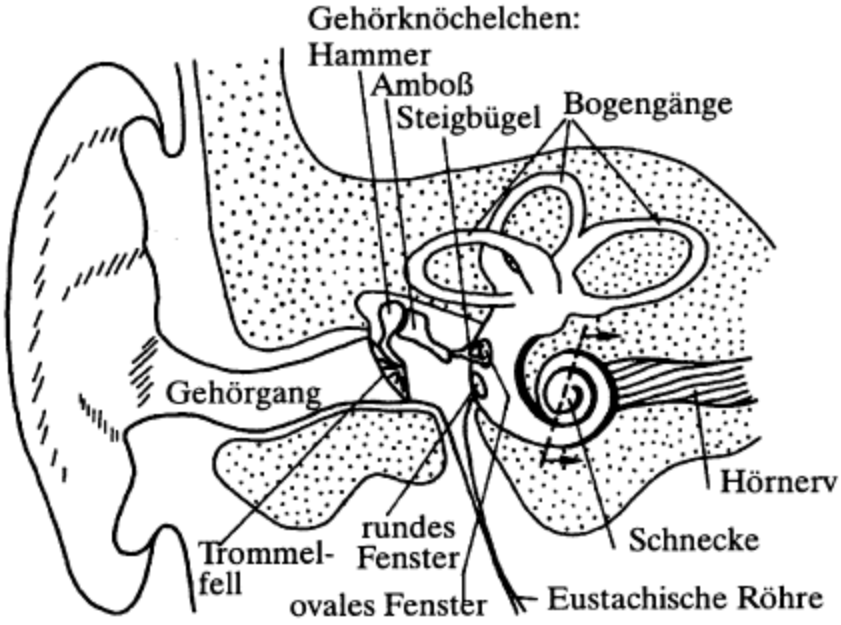


Die Analyseaufgabe besteht nun darin, die Koeffizienten ak und den Amplitudenfaktor A so zu bestimmen, dass das Ausgangssignal möglichst gut dem Anregungssignal s(t) entspricht. Ist dies der Fall, so hat man neben dem Anregungssignal gleich auch die Übertragungsfunktion des Vokaltraktes bestimmt.

Das Filter sagt also die Differenz zwischen Sprachsignal und Anregungssignal (mal Faktor A) als Linearkombination vergangener Signalwerte g(t) voraus; das Verfahren wird daher als lineare Prädiktion oder als LPC-Analyse (Linear Predictive Coding) bezeichnet.

Zur Bestimmung von A kann z.B. der Effektivwert (quadratischer Mittelwert) von g(t) verwendet werden. Die Koeffizienten ak werden so bestimmt, dass die Differenz zwischen s(t) und s'(t) im Zeitbereich bzw. die Differenz zwischen S(jbω) und S'(jω) im Frequenzbereich minimal wird.

5 Grundlagen der auditiven Wahrnehmung



5.1 Außenohr

Richtwirkung; es ermöglicht das räumliche Hören.
Die Ohrmuschel stellt zusammen mit dem Gehörgang ein Resonatorsystem dar, d.h. es verstärkt einzelne Frequenzen und schwächt andere ab.
Das Hören mit beiden Ohren (binaurales Hören) ist Voraussetzung für das räumliche Hören; nur so kann die Position und Entfernung einer Schallquelle adäquat bestimmt werden.

5.2 Mittelohr (HAS)

Die Hauptfunktion des Mittelohres besteht darin, die Impedanz (den Wellenwiderstand) des Außenohres, welches mit Luft gefüllt ist, an die Impedanz des mit Lymphflüssigkeit gefüllten

Innenohres (Schnecke) anzupassen. Die Anpassung geschieht zum einen über die Hebelwirkung der Gehörknöchelchen, zum anderen über das Verhältnis der Flächen von Trommelfell und ovalem Fenster.

-> minimale SChutzfunktion -> kann bei extremen schalldrücken ausweichen

5.3 Innenohr und Nervensystem

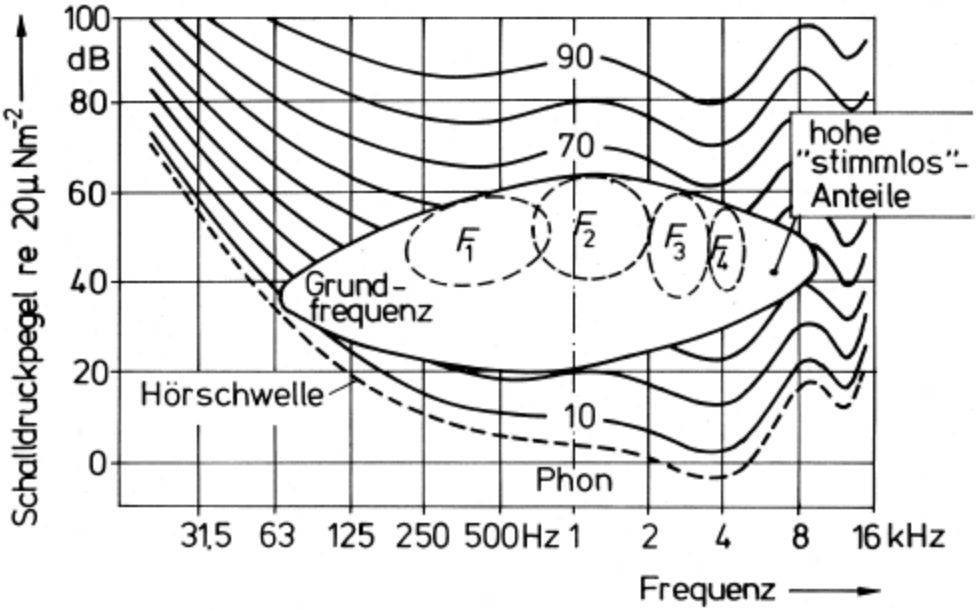
Der Kanal der abgewickelten Schnecke ist durch ein Membransystem in zwei Haupt- und Nebenkanäle aufgeteilt. Diese Kanäle sind mit Lymphflüssigkeit gefüllt. Die wichtigste Membran ist die sog. Basilarmembran. Auf ihr befinden sich die Sinneszellen, die die Hörwahrnehmung hervorrufen. Die Sinneszellen werden durch die Bewegung der Membranen angeregt.

Die Frequenzen des Anregungssignals (Druck auf das ovale Fenster) auf unterschiedliche Orte der Basilarmembran abgebildet, wo sie zur Erregung der betreffenden Nervenzellen führen. Auf der Basilarmembran findet also eine Frequenz-Orts-Transformation statt.

Im Corti'schen Organ findet also eine Analog-Digital-Wandlung statt, bei der die analog in der Schallwelle vorliegende Information über die Basilarmembran-Bewegung (die ihrerseits die Frequenz kodiert) in Spikefolgen der Nervenfasern umkodiert wird.

5.4 Frequenzauflösung und Tonhöhenwahrnehmung

Die Empfindlichkeit ist allerdings nicht bei allen Frequenzen gleich groß. Zur Beschreibung der Empfindlichkeit kann man den **Schalldruckpegel**, der bei einer bestimmten Lautstärke gehört wird, **über der Frequenz** auftragen. Man kommt dann zu Hörbereichs-Empfindlichkeits-Darstellungen oder sog. **Hörflächen**.



Lautstärkepegel: Pegel des als gleich laut empfundenen 1 kHz-Tones (Einheit phon).

Die wahrgenommene Höhe eines Tones hängt mit seiner Frequenz zusammen. Man bezeichnet allgemein die Tonhöhe als die Position eines Tones auf eine Skala. Dabei bestehen aber unterschiedliche Skalierungsmöglichkeiten:

1. **Schallereignisskala** der Tonhöhe (harmonische Tonhöhenskala)
Dabei entsprechen gleiche Intervalle auf der Skala Verdoppelungen der Frequenz (Oktavschritte, das heißt gleiche musikalische Intervalle). Es handelt sich also um eine logarithmische Frequenzmaß-Skala

2. **Hörereignisskala** der Tonhöhe (melodische Tonhöhenskala)
Hierbei wird die Frequenzauflösung der Basilarmembran, d.h. eine natürliche Tonhöhenskala abgebildet. Die Zuordnung ergibt sich in etwa wie in der folgenden Abbildung gezeigt. Hier ist insbesondere die Mel-Skala von Interesse;

Verhältnistonhöhe (gemessen in mel)
Tonheit (critical band rate, gemessen in Bark).

das menschliche Gehör ungefähr 600 Tonhöhen unterscheiden kann.

5.5 Lautheitswahrnehmung

Befragt man Versuchspersonen, wie stark sich zwei Töne unterschiedlichen Schalldruckpegels in ihrer Lautheit unterscheiden (bspw. halb oder doppelt so laut), so kommt man auf die Lautheit N mit der Einheit sone.

Offenbar scheint das Gehör bei der Lautheitsbildung über bestimmte Bereiche des Frequenzkontinuums zu integrieren. Man bezeichnet diese Bereiche als Frequenzgruppen (critical bands);

Reiht man die Frequenzgruppen willkürlich und lückenlos aneinander, so erhält man für den Bereich hörbarer Frequenzen 24 Frequenzgruppen. Jede dieser Gruppen ist 1 Bark = 100 mel breit.

7 Sprach Technologische Systeme

7.1 Spracherkennung

Es findet also eine Umsetzung von der Signal- auf die Symbolebene statt. Die dazu verwendete Methode hängt von einer Vielzahl von Randbedingungen ab.

- Sprache
- Sprecher
- Zieleinheiten
- Anzahl der Zieleinheiten
- Komplexität
- Umgebung

7.1.1 Problemstellungen

Problem, dass im Sprachsignal die einzelnen Einheiten (Phone, aber auch Wörter) nicht in isolierter Form vorliegen.

Aussprache einzelner Laute durch die benachbarten Laute stark beeinflusst; man bezeichnet diesen Effekt als Koartikulation.

Laute oder Silben werden beim schnellen Sprechen ausgelassen (reduziert)

zwei Sprachsignale desselben Satzes, wenn er von zwei unterschiedlichen Sprechern vorgelesen wird, zum Teil deutlich voneinander unterscheiden.

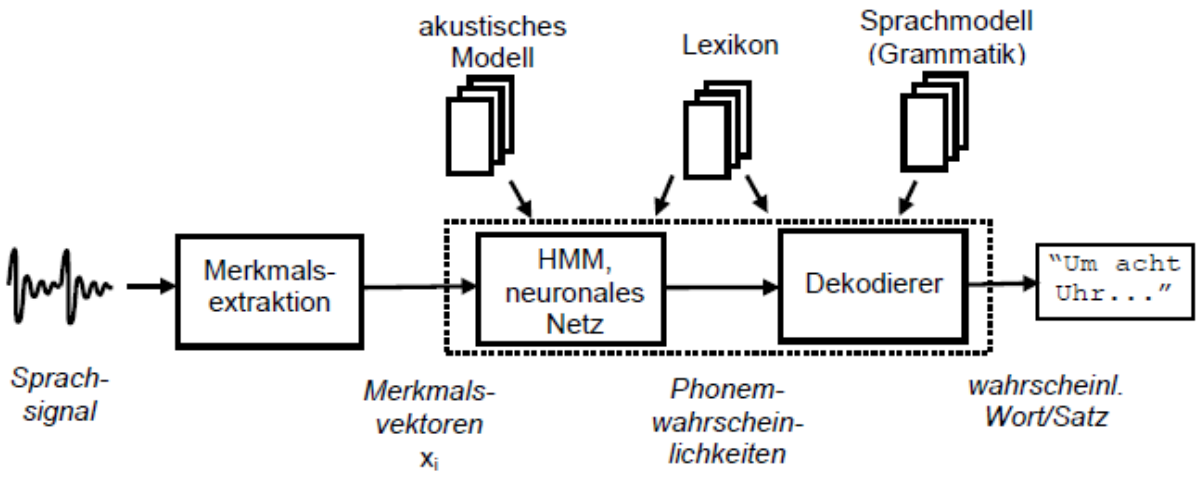
Ziel muss es also sein, möglichst viel Wissen über die Sprache im Erkennungsalgorithmus zu berücksichtigen. Hierbei ist an folgendes Wissen gedacht

Wissen über die Spracherzeugung
akustisches Modell
Vokabular
Grammatik

7.1.2 Aufbau eines Spracherkenners

Um der Variabilität des Sprachsignals Rechnung zu tragen wird die maschinelle Spracherkennung (automatic speech recognition, ASR) im Allgemeinen als statistischer Prozess durchgeführt. Dabei wird eine beobachtete Folge von Lauten bezüglich ihrer Ähnlichkeit mit mehreren vortrainierten Lautfolgen verglichen, um anschließend die ähnlichste (und damit die wahrscheinlichste) Lautfolge auszuwählen. Der Vergleich wird aber nicht nur auf Laut-, sondern auf einer (transformierten) Signalebene durchgeführt.

Zum effizienten Vergleich sind hilfreich:
• Eine **adäquate Repräsentation** der Sprachsignale (sog. **Merkmals-Repräsentation**)
• Eine Zuordnung von Merkmalen zu Lautmustern (sog. **akustisches Modell**)
• Eine Auflistung darüber, welche Lautmuster aufeinander folgen dürfen (**Lexikon**)
• Informationen darüber, wie häufig die entsprechenden Lautmuster in welcher Reihenfolge auftreten (sog. **Sprachmodell**), entweder als statistische Wahrscheinlichkeiten oder als feste Regeln
• **Effiziente Algorithmen zur Auswahl der wahrscheinlichsten Lautmuster**



Schematischer Aufbau eines Spracherkenners.

Zur Berechnung der Phonemwahrscheinlichkeiten werden üblicherweise entweder Hidden-Markov-Modelle (HMMs) oder sog. neuronale Netze eingesetzt; zur Dekodierung benutzt man meist HMMs.

7.1.3 Merkmalsextraktion

Die Merkmale sind die eigentlichen Informationsträger

die zur Unterscheidung der Sprachlaute (und nicht etwa der Sprecher oder der Sprechumgebung) relevanten Informationen aus dem Sprachsignal extrahieren und in eine geeignete Repräsentation überführen.

Hierzu haben sich Verfahren bewährt, die explizit die Lautinformation im Vokaltrakt extrahieren, z.B. das in Kapitel 4 vorgestellte Cepstrum oder die lineare Prädiktion.

Mel-Skaliertes Cepstrum

Das Cepstrum führt eine explizite Trennung von Anregungssignal und Lautformung durch und ist deshalb gut zur effizienten Erfassung der Laut-relevanten Informationen geeignet.

Zur Berechnung eines Mel-skalierten Cepstrums (sog. mel frequency cepstral coefficients, MFCCs) wird das Sprachsignal zunächst mittels einer Fourier-Transformation in den Spektralbereich transformiert. Dieses Spektrum wird nun in 20 bis 24 Bänder eingeteilt, wobei die Bänder selbst mit dreieckförmigen Filtern gewichtet werden,

Auf diese in Bändern zusammengefassten Spektralwerte wird nun die nichtlineare Operation (Logarithmus) angewendet und zurück in den Quefrency-Bereich transformiert. Hierbei ergeben sich (typischerweise 13) cepstrale Koeffizienten, die als Merkmalsvektor verwendet werden

können. Allerdings enthält der Vektor kaum Informationen über die Änderung der Koeffizienten (und damit des Vokaltraktes) über der Zeit; diese können nachträglich zugefügt werden, indem für jeden Koeffizienten die Differenz (Δ) und die zweite Ableitung (ΔΔ) zu den Koeffizienten der vorhergehenden Rahmen (Fenstern) berechnet wird. Die Dimension des Vektors steigt damit auf 26 oder 39 Einträge.

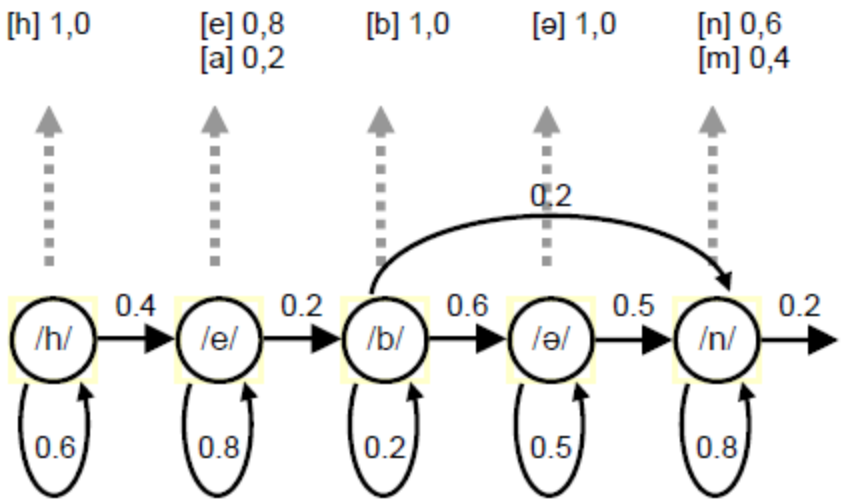
Perceptual Linear Predictive (PLP) Coding:
LPC auf Bark Skala

Relative Spektral Coding (RASTA)
Die RASTA-Analyse (ebenfalls von Hermansky entwickelt) nutzt aus, dass die Wechselgeschwindigkeit der Störkomponenten oftmals außerhalb der typischen Bewegungen des Vokaltraktes und damit der Stationaritätsdauer von Sprachsignalen (typischerweise 20 ms) liegen. Es werden also Komponenten unterdrückt, die sich schneller oder langsamer als typische Sprache ändern. Sich langsam ändernde Anteile (z.B. Hintergrundgeräusche) werden durch eine gesonderte Bandpass-Filterbank herausgefiltert.

7.1.4 Hidden-Markov-Modelle und neuronale Netze

Hidden-Markov-Modelle sind ein Standardwerkzeug zur Beschreibung und Modellierung von Zufallsprozessen.
Durch Vergleich zwischen generierten und beobachteten Merkmalen wird derjenige „Weg“ durch das Modell ermittelt, durch den die **beobachtete Merkmalsfolge am wahrscheinlichsten erzeugt worden sein könnte**. Die zugehörige Symbolfolge ist dann das Erkennungsergebnis.

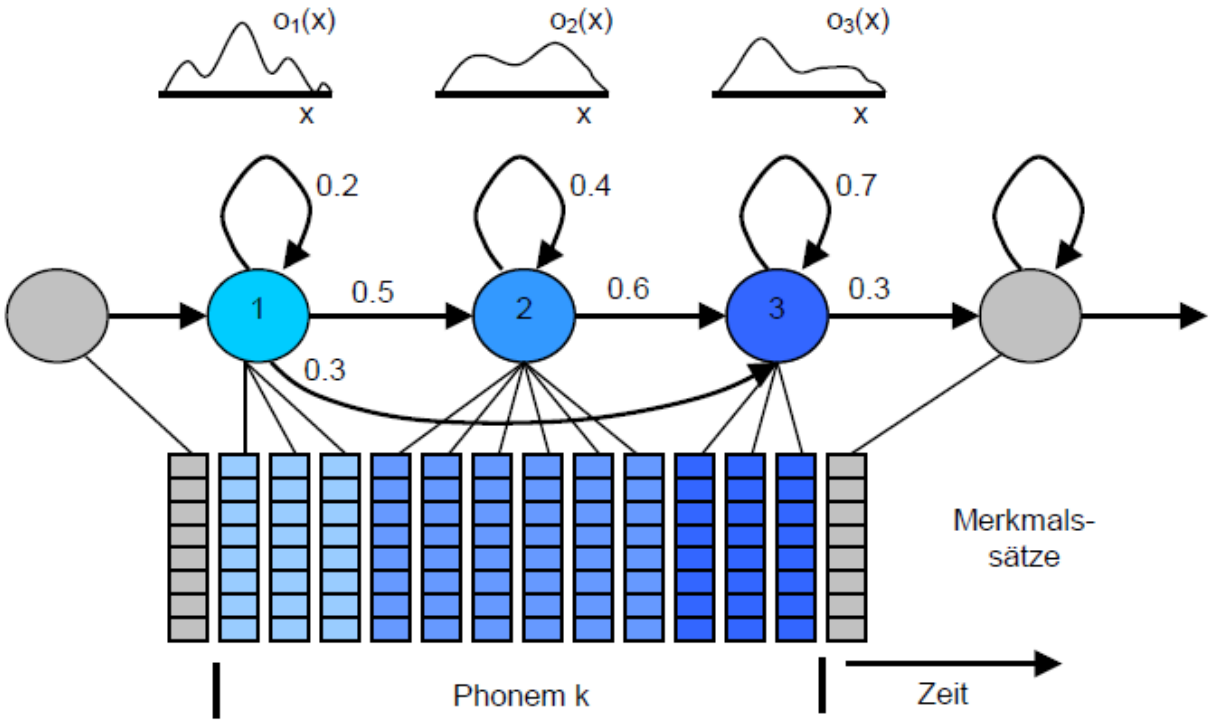
Angegeben sind jeweils die **Übergangswahrscheinlichkeiten** sowie die in jedem Zustand „emittierten“ Symbole. Es ist ersichtlich, dass jedem Zustand nicht ein Symbol ein-eindeutig zugeordnet ist, sondern dass ein Zustand mit unterschiedlichen Wahrscheinlichkeiten unterschiedliche Symbole emittieren kann. Man bezeichnet diese Wahrscheinlichkeiten als **Emissionswahrscheinlichkeiten**.



Beispiel eines einfachen Hidden-Markov-Modells für Sprache.

Durch die nicht ein-eindeutige Zuordnung von Ausgabesymbolen zu Zuständen lässt sich aus einer beobachteten Folge von Ausgabesymbolen nicht direkt auf die dahinter liegende Folge von Zuständen schließen. Man bezeichnet diese Markov-Modelle deshalb als „hidden“.

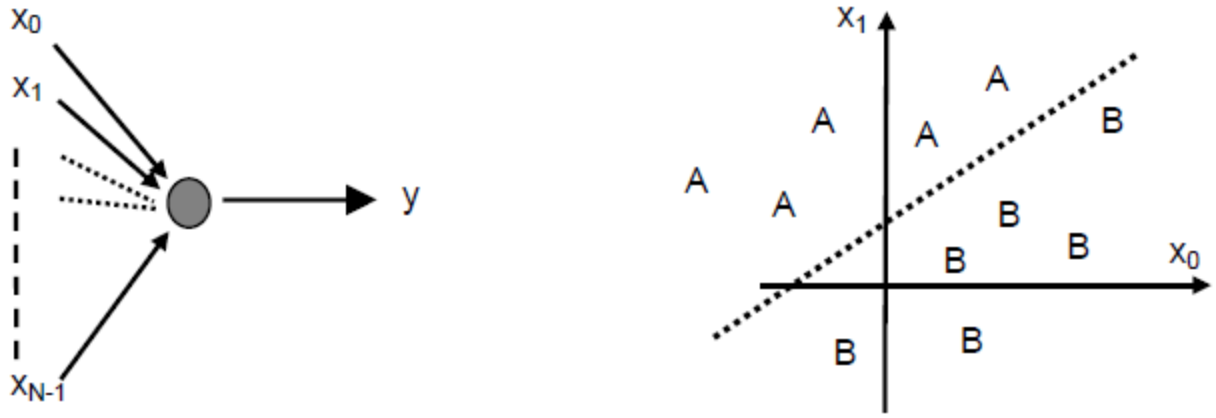
Meist wird ein einzelnes Phonem als eine Kette von 3 Zuständen dargestellt, einen für den **Übergang vom vorangehenden Phonem**, einen für den **mittleren Zustand**, und einen für den **Übergang zum nächsten Phonem**. Dadurch wird die **Koartikulation** im Modell berücksichtigt.



Phonem-HMM.

HMM muss trainiert werden, wahrscheinlichkeiten an die zu erkennende sprache angepasst werden.

NEURONALE NETZE
PERZETRON
Ein Perzeptron ist zunächst ein einfacher Klassifikator. Es verfügt über einen Eingangsvektor x (z.B. den beobachteten Merkmalsvektor), einen Gewichtsvektor w, und einen (ein-dimensionalen) Ausgangswert y.d.h. das Ausgangssignal kann nur zwei Werte (+1 oder -1) annehmen.



Single-Layer-Perzeptron.

Mit einem solchen Perzeptron lassen sich also zwei Klassen unterscheiden, anhand eines Vektors von Eingangsgrößen. Die Klassifikation lässt sich bei einem 2-dimensionalen Eingangsvektor als Linie interpretieren, die die Fläche in zwei Klassen teilt; bei höherer Dimensionalität des Eingangsraumes entspricht dies einer Fläche (drei Dimensionen) bzw. einer Hyper-Fläche.

7.1.5 Sprachmodelle

Zum einen kann versucht werden, eine explizite Grammatik zu definieren, die den Wortschatz des Erkenners (d.h. die mögliche Aufeinanderfolge von Wörtern) möglichst gut repräsentiert, Eine solche Grammatik wird oft als sog. kontextfreie Grammatik formuliert.

Hierzu wird ein großes Korpus ausgezählt und die Wahrscheinlichkeit, dass zwei oder mehr (allgemein n) Wörter aufeinander folgen, berechnet. Man bezeichnet eine solche Grammatik als n-gram (bigram, trigram).

7.1.6 Erkennungsleistungen

Zum Vergleich verwendet man meist die Wortfehlerrate (Word Error Rate, WER) oder die Rate der richtig erkannten Wörter (Word Accuracy, WA). Diese können mit Hilfe von etikettierten Daten bestimmt werden, d.h. mit Sprachdaten, zu denen richtige, von einem menschlichen Hörer angefertigte Transkriptionen vorliegen.
Nach dem Alignment werden alle korrekt erkannten Wörter (cw), vertauschten Wörter (sw), gelöschten Wörter (dw) und eingefügten Wörter (iw) gezählt.

Neben den Maßen für die Erkennungsrate auf Wortebene können auf gleiche Weise Maße auf Satzebene (Sentence Accuracy, SA, und Sentence Error Rate, SER) berechnet werden.

7.2 Sprachsynthese

Die Aufgabe der Sprachsynthese besteht darin, aus auf Symbolebene vorliegenden Text ein Signal zu generieren, welches als Sprache wahrgenommen wird.

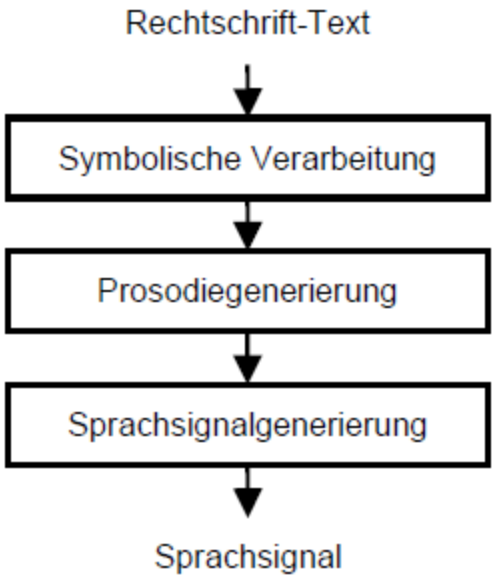
Die erste neuzeitlich entwickelte Sprechmaschine, die nicht nur einzelne Laute, sondern zusammenhängende Sprecheinheiten erzeugen konnte, war der von Dudley 1939 entwickelte Voder (VOice DEmonstratoR). Mit Hilfe des sog. Vocoders (VOice CODER) gelang es erstmals, gesprochene Sprache in eine parametrische Darstellung zu überführen und daraus wiederum verständliche Sprache zu generieren (Dudley, 1939).

Eine vollständige Generierung auf Basis von Text ist nicht unbedingt immer notwendig, vor allem dann nicht, wenn die zu synthetisierenden Äußerungen vorher bekannt sind.

Zur einfachen Generierung von Sprache sehr begrenzten Wortschatzes reicht es oft aus, eine oder mehrere Sprachsignale aus einem Vorrat von Signalen auszuwählen und hintereinander abzuspielen (sog. canned speech).

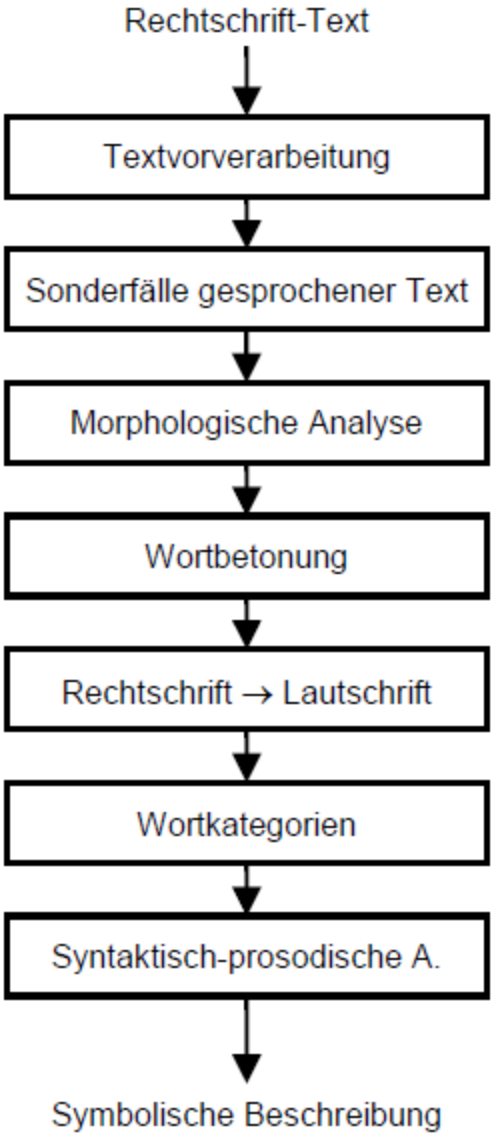
Man kann Sprachausgabesysteme nach ihren Leistungsmerkmalen in folgende Klassen einteilen (vgl. Blauert und Schaffert, 1985):
• Ansageautomaten (Verkettungssynthese)
• Aussageautomaten (Concept-to-Speech-Systeme)
• Vorleseautomaten (Text-to-Speech)

7.2.1 Struktur eines Vorleseautomaten



Bei einem Concept-to-Speech-System (CTS) wird als Eingabe nicht ausformulierter Text, sondern eine abstrakte Darstellung des zu synthetisierenden Sachverhaltes verwendet.

7.2.2 Symbolische Verarbeitung



Textvorverarbeitung:
in eine einheitliche Form gebracht werden. Hierzu müssen Sonderfälle im geschriebenen Text, bspw. Ziffern, Abkürzungen oder Nummerierungen, in eine sprachliche Form gebracht werden.

Behandlung von Sonderfällen im gesprochenen Text:
Hierbei sind z.B. die Aussprache von Eigennamen oder fremdsprachlicher Wörter zu nennen.

Morphologische Analyse:
Um eine zuverlässige phonetische Transkription durchzuführen können einzelne Worte in Wortstämme, Präfixe, Suffixe, Flexionsendungen und Fugenelemente aufgeteilt werden.

Bestimmung der Wortbetonung:
Betonungsstufen der Silben erkannt werden
Allerdings kann sich die Wortbetonung durch die Satzbetonung noch ändern.

Rechtschrift-nach-Lautschrift-Umsetzung:
Die Umsetzung von Rechtschrift nach Lautschrift geschieht meist an mehreren Stellen der Synthese. Bspw. können erkannte Morphe direkt mit Lautschrift-Informationen versehen werden.

Bestimmung von Wortkategorien:

Bestimmung der syntaktischen und prosodischen Struktur:

7.2.3 Prosodiegenerierung

Neben der textlichen Information stellt die Prosodie wichtige Informationen zur Interpretation einer Aussage bereit.

Regelbasierte Verfahren
Zur Generierung der Satzmelodie kann z.B. das Modell nach Fujisaki (1983) verwendet werden, welches die Satzmelodie als Folge von rechteck- oder impulsförmigen Anregungsfunktionen beschreibt, welche dann durch Filter 2. Ordnung in Konturen für die Grundfrequenz umgewandelt werden.

Datenbasierte und statische Verfahren
Hierbei werden z.B. Silbendauern aus großen Korpora annotierter Sprache parametrisch beschrieben und einem statistischen Modell zugeführt. Ein solches Modell kann z.B. ein neuronales Netz oder ein Klassifikationsbaum

7.2.4 Sprachsignalgenerierung

Parameterische Synthese
Die Verwendung von Modellen der menschlichen Spracherzeugung bei der Synthese hat den Nachteil, dass nicht alle Aspekte genügend genau modelliert werden können. Deshalb ist die damit erzielbare Qualität im Allgemeinen recht begrenzt.

Die Idee ist hierbei, die menschliche Spracherzeugung parametrisch zu beschreiben und aus den Parametern mittels eines Modells Sprechschalle zu erzeugen. Die dabei verwendeten Ideen orientieren sich stark an den in Kapitel 3 beschriebenen Modellen zur Spracherzeugung (insbes. dem Quelle-Filter-Modell)

Eine andere Variante der parametrischen Synthese ist der LPC-Synthetisator. Dabei werden aus natürlicher Sprache zunächst LPC-Parameter bestimmt, mit deren Hilfe dann neu synthetisiert werden kann. Hierzu wird meist eine rein stimmhafte oder stimmlose Anregung gewählt, und mit Hilfe des Prädiktions-Synthesefilters (vgl. Abschnitt 3.5) das Sprachsignal synthetisiert. Benötigt

werden also die LPC-Parameter, die Amplitude des Anregungssignals und eine Stimmhaft-Stimmlos-Entscheidung.

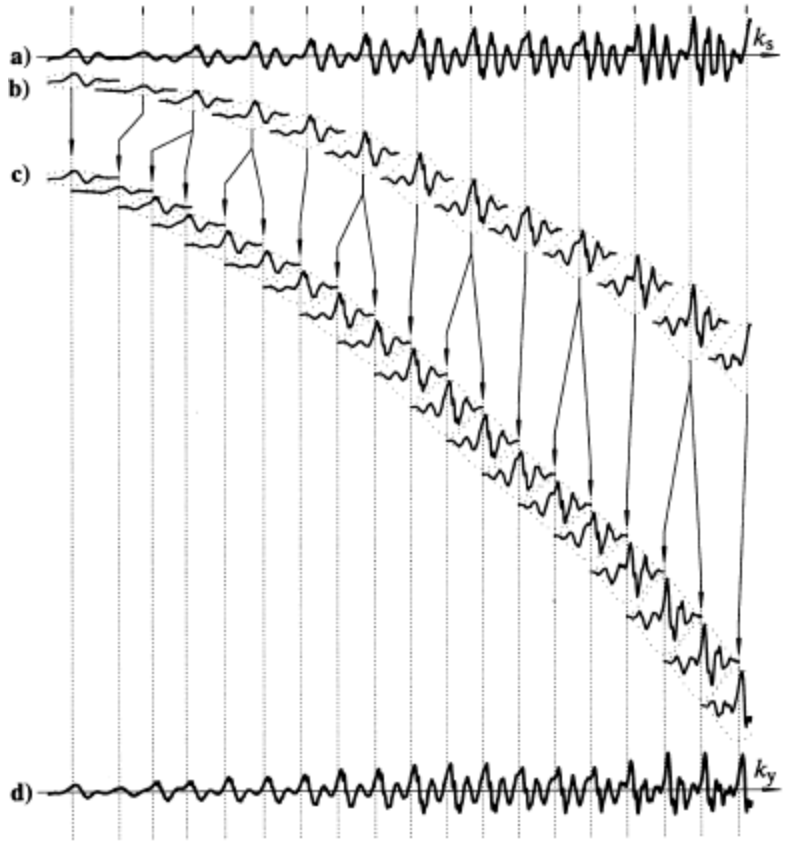
Artikulatorische Synthese

Die genauen Bewegungen der Artikulationsorgane nachzubilden.
Allerdings ist die Komplexität der Modellierung wie auch der Berechnung sehr hoch.

Verkettungssynthese

Daher versucht man, durch Verkettung einzelner Bausteine natürlich erzeugter Sprache ein neues Sprachsignal zusammenzusetzen. Im Gegensatz zum eingangs erwähnten canned speech wird hier allerdings eine umfangreichere Signalmanipulation durchgeführt.

Einheiten, die sich zur Verkettung eignen, können Phone, Diphone, Halbsilben, Silben, ganze Wörter oder sogar Wortketten sein.



Zum Verketten hat sich das sog. PSOLA-Verfahren (pitch-synchronous overlap-and-add) bewährt. Bei PSOLA werden im natürlichen Sprachmaterial (Synthesebausteine) zunächst Marker für die Grundperiode gesetzt; dies können z.B. die Zeitpunkte des Glottisverschlusses oder andere Extremwerte im Zeitsignal sein. Man bezeichnet diese Marker als Periodenmarken. Um jede Periodenmarke herum wird nun ein Abschnitt des Sprachsignals herausgefenstert, meist mit einem sanft ein- und ausblendenden Hann-Fenster. Dadurch wird das Eingangssignal (Synthesebausteine) zunächst in eine diskrete Folge von Elementarbausteinen zerlegt, die durch die Grundperioden des Eingangssignals vorgegeben werden. Das Ausgangssignal wird gebildet,

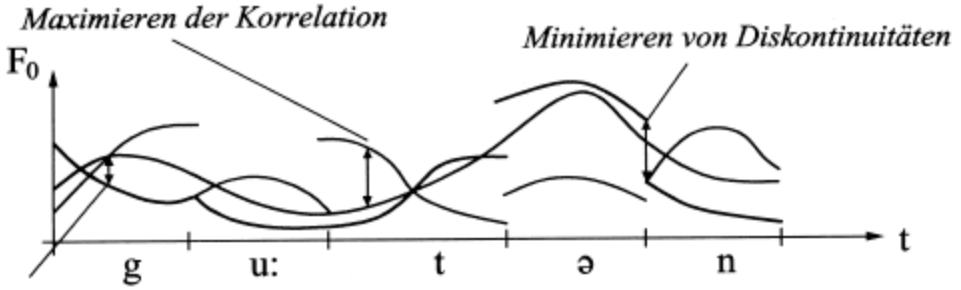
indem die Signalwerte von typischerweise zwei beteiligten Elementarbausteinen – verschoben auf die Grundperiode des Ausgangssignals – addiert werden. Dadurch entsteht eine Verschiebung auf der Zeitachse, ohne dass die Formantstruktur des Eingangssignals nennenswert beeinträchtigt würde.

Mit dem PSOLA-Verfahren lassen sich also Signalabschnitte synchron zur Grundperiode zusammensetzen. Die dadurch erzeugte Sprache hat gegenüber einfacher Verkettung den Vorteil, dass sich keine direkten Grundfrequenzsprünge im Zeitsignal ergeben. Zudem ist keine Transformation des Zeitsignals notwendig.

Units-Selection-Synthese

Bei der Unit-Selection-Synthese versucht man, Zeitsignalabschnitte möglichst ohne oder nur mit geringer Signalmanipulation zu verketten. Die Signalabschnitte sollten dabei so lang wie möglich sein, um die Anzahl der Verkettungsstellen zu minimieren. Dadurch steigt wiederum die Größe des Vokabulars, was heutzutage zwar kein prinzipielles Problem mehr darstellt, aber den Aufwand zur Erstellung des Vokabulars (u.U. von mehreren Sprechern) erheblich erhöht.

Um die wahrnehmbaren Sprünge an den Verkettungsstellen so gering wie möglich zu halten, werden Signalabschnitte ausgewählt, die bezüglich ihrer prosodischen Struktur möglichst gut dem zu synthetisierenden Sprachabschnitt entsprechen. Dazu werden einzelne Sprachbausteine oft in mehreren Varianten im Syntheseinventar abgelegt, Varianten, die sich nur durch ihre Prosodie unterscheiden.



Die Passgenauigkeit wird über eine Kostenfunktion definiert

- Einheitenkosten: wie gut die im Inventar vorhandenen Bausteine zu den zu synthetisierenden passen. Diese Kosten müssen online zur Laufzeit bestimmt werden.
- Verkettungskosten: passgenauigkeit Bausteine des Inventars zueinander. Verkettungskosten können offline vor bestimmt werden.

HMM-basierte Synthese

Das führt zu der (verwirklichten) Idee, dass sich Sprachsynthese als Erkennungsaufgabe mit Hilfe eines HMMs lösen lässt. Von einem natürlichen Sprecher werden zunächst Äußerungen aufgenommen und in Bausteine zerlegt. Wie bei der Unit-Selection-Synthese liegen die Bausteine zumeist mehrfach im Inventar vor. Diese Bausteine entsprechen dann den Zuständen eines HMMs, wobei jeder Zustand des HMMs mit einer parametrischen Darstellung des Zeitsignals (z.B. LPC-Parameter) verknüpft ist.

Die HMMs werden zunächst mit einer Datenbasis des Sprechers mit vielen (phonetisch ausbalancierten) Sätzen trainiert, d.h. es werden Übergangswahrscheinlichkeiten und Emissionswahrscheinlichkeiten bestimmt. Zur Synthese wird nun der optimale Pfad durch das HMM gesucht, um die gewünschte Äußerung zu synthetisieren. Dieser Pfad ergibt die Parameter, aus denen schließlich das Sprachsignal synthetisiert wird.

7.3 Natürlichsprachliche Dialogsysteme

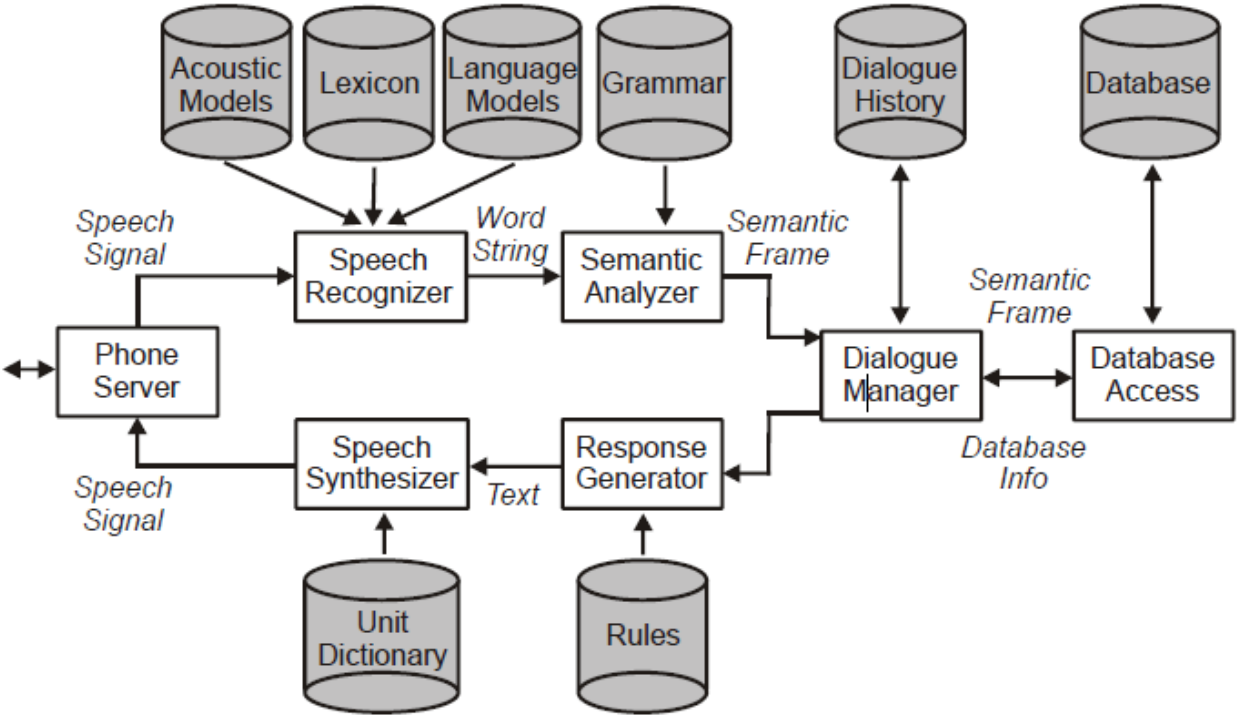
Die Aufgabe des Sprachdialogsystems besteht darin, die Interaktion zwischen Benutzer und Anwendungssystem aufrecht zu erhalten und dabei die nötigen Informationen auszutauschen. Diese Aufgabe umfasst z.B.

• die **Verifikation** der Kohärenz der **Benutzereingaben**,
• die **Verhandlung** von kommunikativen und aufgabenbezogenen **Zielen**,
• die **Lösung** von kommunikativen **Problemen**,
• die **Auflösung** von **Auslassungen und Referenzen**,
• die **Vorhersage** der wahrscheinlich nächsten **Benutzeräußerung**, und schließlich
• die **Generierung** einer adäquaten **natürlichsprachlichen Äußerung** für den Benutzer.

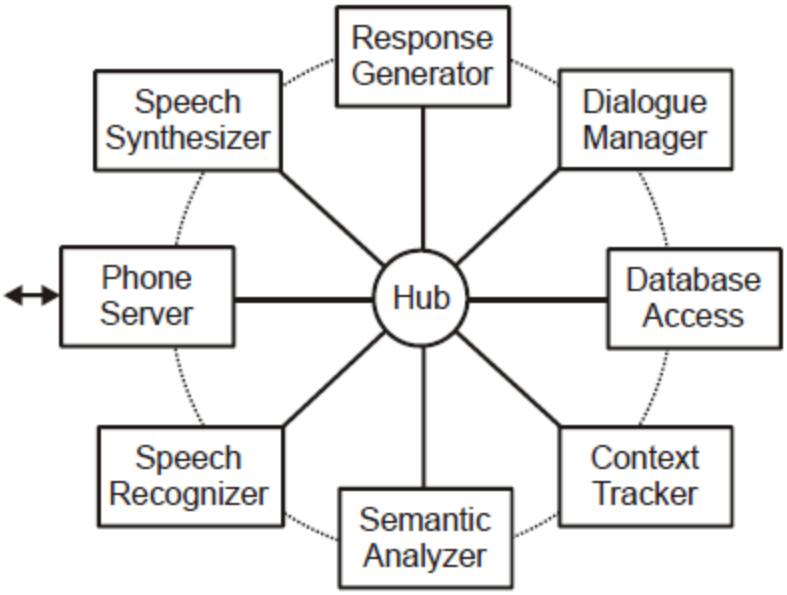
Unterschieden wird in:
- Kommandosysteme
- Menüorientierte Systeme
- Sprachdialogsysteme
- Multimodale Dialogsysteme

Struktur eines Sprachdialogsystems

Sequenztielle Struktur



Hub Struktur



Auch in der Hub-Struktur wird die Information im Prinzip sequentiell verarbeitet. Darüber hinaus sind aber auch Strukturen denkbar, in denen einzelne Module autonom operieren und jeweils die Initiative zum Lenken des Dialogs übernehmen können.

Sprachverstehen

Die Aufgabe des sprachverstehenden Moduls ist es, aus der Worthypothese des Spracherkenners die semantischen Informationen zu extrahieren, die für den Dialogverlauf wichtig sind, und diese in geeigneter Form dem Dialogmanager zur Verfügung zu stellen.

Aufgrund der Natur spontaner Sprache ist eine komplette grammatikalische Analyse meist nicht möglich. Deshalb ist es notwendig, dass der Parser auch mit Satzteilen bzw. unvollendeten Äußerungen, Einwürfen etc. zurechtkommt. In einfachen Fällen kann auch eine Schlüsselworterkennung (keyword spotting) ausreichend sein.

Dialogmanagement

Der Dialogmanager muss einen glatten Verlauf des Dialogs sicherstellen, bei dem alle wichtigen Informationen zur Lösung der Aufgabe ausgetauscht werden und der letztendlich zur „richtigen“ Lösung der Aufgabe führt. Dazu muss Wissen über die Aufgabe sowie ein allgemeines „Weltwissen“ vorhanden sein. Der Gesprächsverlauf muss mitverfolgt werden.

Weitere wichtige Aufgaben des Dialogmanagers bestehen u.a. darin,
• die **Initiative** im Dialogverlauf zu verteilen,
• **Feedback** über erkannte und verstandene Dinge zu geben,
• dem Benutzer **Hilfestellungen** zu geben,
• **Fehler** und Missverständnisse zu **korrigieren**,
• Komplexe Dialogphänomene wie **Auslassungen (Ellipsen) und Referenzen aufzuklären**, sowie
• die **Informationsausgabe** zum Benutzer zu steuern.

Die geforderten Funktionen können auf unterschiedliche Arten implementiert werden.
Dialog-Grammatiken:
Der Dialogverlauf wird hier z.B. als Kette von Zuständen vorgegeben. Jeder Zustand steht für eine Äußerung oder Aktion des Systems. Übergänge zwischen den Zuständen sind von den Äußerungen des Benutzers – und was davon erkannt und verstanden wurde – abhängig.
Plan-basierte Ansätze:
Hierbei wird explizit versucht, einzelne aufgabenbezogene Dialogziele zu modellieren.
Kollaborative Ansätze:
Im Gegensatz zur Modellierung einzelner aufgabenbezogener Dialogziele versuchen kollaborative Ansätze, die Motivation, die hinter einem Dialog steckt, und die Dialogmechanismen, die der Mensch zu Erreichung seiner Ziele benutzt, zu modellieren.

Um den Dialogverlauf verfolgen zu können benutzt ein Dialogmanager eine Reihe von Speichern und Modellen:

• **Dialog History:** Eine Abfolge aller im Verlauf des Dialogs gemachten Vorschläge des Benutzers und des Systems.
• **Task Record:** Eine Repräsentation der **Aufgabe**, die mit Hilfe des Systems erledigt werden kann (z.B. als slots), **und die dazu im Gesprächsverlauf gesammelten Informationen**.
• **World Knowledge Model:** Eine Repräsentation von **Hintergrundinformationen**, die für die Aufgabe wichtig sind (Kalender, etc**.)**.
• **Domain Model:** Eine Beschreibung der **Aufgaben-Domäne,** d.h. des Zugverkehrs, der Fahrpreise, etc.
• **Conversational Model:** Ein allgemeines **Modell der kommunikativen Fähigkeiten**.
• **User Model:** Eine Repräsentation der **Präferenzen, Ziele, Annahmen und Intentionen des Benutzers.**

Eine Hauptaufgabe des conversation model ist die Steuerung der Initiative zwischen Benutzer und System.
• System-Initiative, (Telefonservice) d.h. die Initiative bleibt beim System und die Aufgabe des Benutzers ist es, Fragen zu beantworten,
• Benutzer-Initiative, (Siri) d.h. das System reagiert hauptsächlich auf die Fragen/Äußerungen des Benutzers, und
• gemischte Initiative, d.h. beiden Gesprächspartnern ist es erlaubt, Fragen zu stellen oder Vorschläge zu machen.

Das System Rückmeldungen darüber gibt, was es verstanden hat:
- Explizit: Verstandene wird vom System wiederholt und vom Benutzer bestätigt
- Implizit: Verstande wird im nächsten Satz wiederholt aber muss nicht bestätigt werden

Sprachausgabe

Die Formulierung der Antwort sollte die Entscheidung darüber umfassen, welche und wieviele Informationen zu welchem Zeitpunkt und in welcher Form an den Benutzer ausgegeben werden. Zur Formulierung der Aussage können einfache Grammatiken oder Schablonen verwendet werden, in die die jeweiligen Informationen eingetragen werden. Bei der Wortwahl sollte auf eine konsistente und verständliche Formulierung geachtet werden. Es ist davon auszugehen, dass der Benutzer Wörter, die das System in seinen Äußerungen verwendet, selbst wiederum in seiner Antwort verwendet; deshalb sollte das Vokabular der Systemäußerungen auch vom System erkannt und verstanden werden können.

8. Multimodale Dialogsysteme

Auch wir Menschen kommunizieren nicht allein über den akustischen Kanal, sondern wir zeigen gleichzeitig, was wir tun und wollen, benutzen Mimik, Gestik, Bewegungen, Berührungen, etc. um miteinander zu kommunizieren. Eine Interaktion mit Maschinen unter Einbeziehung verschiedener Kanäle zur Informationsübermittlung könnte also vorteilhaft sein. Hinzu kommen technologische und prinzipielle Einschränkungen rein sprachbasierter Systeme.

Dabei verstehen wir unter dem Begriff „Medium“ ein Kommunikationsmittel (Material oder Gerät), welches einen bestimmten physikalischen Kanal benutzt, und unter dem Begriff „Modalität“ die Verwendung dieses Mediums zur Kommunikation, z.B. in Form von Intonation, Blick, Geste, Mimik, etc. Modalitäten sprechen verschiedene Sinne an, z.B. bei der visuellen, auditiven, oder den haptischen Wahrnehmung. Unterschiedliche Modalitäten eignen sich unterschiedlich gut für verschiedene Zwecke der Informationsübermittlung.

8.1 Eigenschaften von Modalitäten

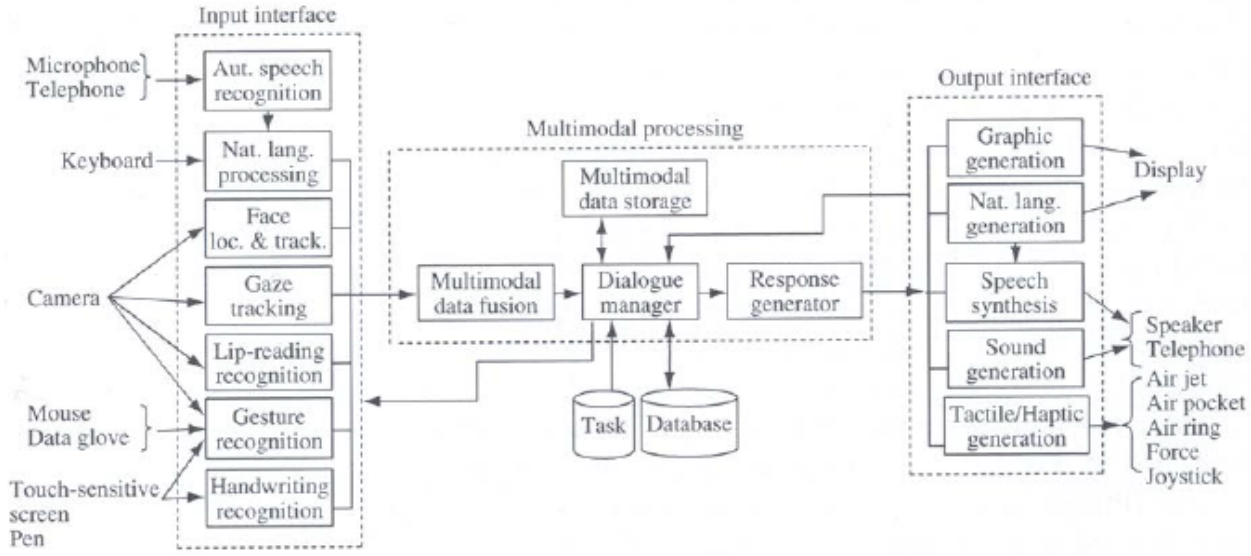
- Linguistisch vs. nicht-linguistisch:
- Analog vs. nicht-analog:
- Arbiträr vs. nicht-arbiträr:
- Statisch vs. dynamisch:
- Klasse von Medien: Grafisch (visuell wahrnehmbar), akustisch (auditiv wahrnehmbar) oder haptisch.

No.	*Modality*	*Modality Property*
1	Linguistic input/output	Linguistic input/output modalities have interpretational scope, which makes them eminently suited for conveying abstract information. They are therefore unsuited for conveying high-specificity information including detailed information on spatial manipulation and location.
2	Linguistic input/output	Linguistic input/output modalities, being unsuited for specifying detailed information on spatial manipulation, lack an adequate vocabulary for describing the manipulations.
3	Arbitrary input/output	Arbitrary input/output modalities impose a learning overhead which increases with the number of arbitrary items to be learned.
4	Acoustic input/output	Acoustic input/output modalities are omnidirectional.
5	Acoustic input/output	Acoustic input/output modalities do not require limb (including haptic) or visual activity.
6	Acoustic output	Acoustic output modalities can be used to achieve saliency in lowacoustic environments. They degrade in proportion to competing noise levels.
7	Static graphics/ haptics input/ output	Static graphic/haptic input/output modalities allow the simultaneous representation of large amounts of information for free visual/tactile inspection and subsequent interaction.
8	Dynamic input/output	Dynamic input/output modalities, being temporal (serial and transient), do not offer the cognitive advantages (wrt. attention and memory) of freedom of perceptual inspection.
9	Dynamic acoustic output	Dynamic acoustic output modalities can be made interactively static (but only small-piece-by-small-piece).
10	Speech input/output	Speech input/output modalities, being temporal (serial and transient) and non-spatial, should be presented sequentially rather than in parallel.
11	Speech input/output	Speech input/output modalities in native or known languages have very high saliency.
12	Speech output	Speech output modalities may complement graphic displays for ease of visual inspection.
13	Synthetic speech output	Synthetic speech output modalities, being less intelligible than

No.	*Modality*	*Modality Property*
		natural speech output, increase cognitive processing load.
14	Non-spontaneous speech input	Non-spontaneous speech input modalities (isolated words, connected words) are unnatural and add cognitive processing load.
15	Discourse input/output	Discourse input/output modalities have strong rhetorical potential.
16	Discourse input/output	Discourse input/output modalities are situation-dependent.
…	…	…

8.2 Allgemeine Architektur eines multimodalen Dialogsystems

In einem multimodalen Dialogsystem sind im Allgemeinen das Dialogmodel, das Aufgabenmodel, die interne Präsentation der Daten und die zur Verfügung stehenden Modalitäten voneinander unabhängig. Daraus folgt, dass ein und dieselbe Information auf unterschiedlichen Wegen – mit unterschiedlichen Modalitäten oder Kombinationen derselben – in das System gegeben werden kann, bzw. von diesem zur Verfügung gestellt werden kann.



Dabei ist es wichtig, die einzelnen Informationskanäle nicht getrennt zu betrachten, sondern in ihrer zeitlichen und inhaltlichen Kombination.
Die erhaltenen Informationen müssen also sinnvoll zusammengeführt werden; man nennt diesen Prozess Fusion (engl. fusion).
Die Entscheidung darüber, welche Informationen über welche Modalität ausgegeben wird, trifft in oben dargestelltem Schema die Response-Generation-Komponente: Sie führt die Aufteilung (engl. fission) der Informationen – also das Gegenstück zur Fusion – durch.
Vorteilhaft, wenn durch das Zusammenspiel verschiedener Modalitäten Vieldeutigkeiten aufgelöst werden.

8.3 Multimodale Eingabe-Schnittstelle

Um weitere Informationen über einen menschlichen Sprecher zu erhalten ist normalerweise die Bestimmung der Gesichtsposition notwendig. Dies zum einen, um Gesten abzulesen, zum anderen, um die Lippen zu lokalisieren, mit deren Hilfe dann z.B. die Spracherkennung verbessert werden kann, oder auch Emotionen erkannt werden können.
Zur Gesichtserkennung werden unterschiedliche Verfahren verwendet.
- Regelbasierter Ansatz:
- Invariante Merkmale
- Mustervergleich:
- Farbe:

Blickbewegungs-Detektoren arbeiten z.B. nach den folgenden Prinzipien:
- Cornea-Reflex-Methode
- Elektro-Okulorgramme (EOG)
- Kontaktlinsen:

Das sog. Lippen-Lesen hilft Menschen bei der Erkennung schwer unterscheidbarer Laute und kann auch die maschinelle Spracherkennung (vor allem in gestörten Umgebungen und bei mehreren parallelen Sprechern) deutlich verbessern. (Visemen, den kleinsten bedeutungsunterscheidenden visuellen Korrespondenten der Phoneme)

Audio-Visual Automatic Speech Recognition, AVASR

Erkennung von Gesten:
Zur Eingabe eignen sich **intrusive** (z.B. Datenhandschuhe) oder **nicht-intrusive Geräte** (z.B. mittels einer Kamera).

Typen von Gesten:
- **Symbolische:** Diese benutzen **Symbole**, um Bedeutung zu übermitteln, bspw. eine Handbewegung, um Zustimmung oder Ablehnung auszudrücken.
- **Deiktische**: **Zeige-Gesten**, um Objekte oder Positionen zu referenzieren.
- **Ikonische:** Gesten, mit deren Hilfe Objekte, **Positionen oder Aktionen visuell** beschrieben werden.
- **Metaphorische:** Gesten, mit denen **abstrakte Ideen** beschrieben werden.
- **Schlagen oder rythmische** Gesten.

8.4 Multimodale Verarbeitung

- Welche Informationen gehören zusammen?
- Auf welcher Ebene lassen sich die Informationen am besten zusammenfassen?
- Wie soll im Falle von widersprüchlichen Informationen reagiert werden?

Informationen, die zeitlich eng zusammen eintreffen, gehören meist zusammen. Allerdings muss dabei die unterschiedliche Verarbeitungszeit der verschiedenen Eingabemodalitäten beachtet werden. Zeitstempel gestatten hierbei eine exakte Zuordnung.

Wie bei der audio-visuellen Spracherkennung können die Informationen auf niedriger (Signal-) oder auf einer höheren (semantischen) Ebene fusioniert werden. Die Fusion auf Signalebene kommt insbesondere dann in Frage, wenn die betrachteten Modalitäten synchron eingehen, bspw. bei der AVASR.

Auf Basis der Konfidenzen können widersprüchliche Informationen im Sinne einer besten Gesamt-Erkennungsrate aufgelöst werden.

8.5 Multimodale Ausgabe-Schnitstellen

Neben der rein sprachlichen Ausgabe werden in jüngerer Zeit vermehrt animierte Agenten (Embodied Conversational Agent, ECA) verwendet, um sprachliche Informationen – verbunden mit Mimik und Gestik – auszugeben. Gegenüber der rein akustischen Ausgabe können solche Avatare verschiedene Vorteile haben, sofern sie gut ausgeführt werden: Sie stellen eine Bezugsperson für den Benutzer dar und können seine Aufmerksamkeit lenken (bspw. auf verschiedene Bereiche des Bildschirmes, in denen Informationen angezeigt werden); sie können den Systemzustand ausdrücken (bspw. durch Mimik Unverständnis anzeigen, oder zeigen, dass auf eine Eingabe gewartet wird); sie können Emotionen besser transportieren; und sie können bei entsprechend genauer Modellierung der Artikulation auch die Sprachverständlichkeit steigern.

Zusätzlich zu Grafiken und animierten Agenten verwenden multimodale Dialogsysteme häufig Icons oder – als akustisches Gegenstück – Auditory Icons (manchmal auch Eracons genannt) – um dem Nutzer schnell und auf einprägsame Weise Informationen zu übermitteln. Icons erfordern aufgrund ihres Charakters meist kaum kognitive Resourcen zur Verarbeitung, und sie sind meist sehr intuitiv und in verschiedenen Sprach- und Kulturgemeinden gleich.

Flipped Classroom: Auditory Perception

1. Wie funktioniert das Richtungshören?
2. Wieso benötigt man ein Mittelohr zwischen Außen- und Innenohr?
3. Erläutern Sie die Funktionsweise des Innenohres, von der Auslenkung des ovalen Fensters bis zum Nervenimpuls!
4. Was versteht man unter der Frequenz-Orts-Transformation im Innenohr?
5. Was versteht man unter dem Lautstärkepegel, und wie bestimmt man ihn?
6. Was versteht man unter der Lautheit, und wie bestimmt man sie?
7. Wie kann man die wahrgenommene Höhe eines Tones skalieren?
8. Erläutern Sie den Effekt der Maskierung!
9. Erläutern Sie den Begriff der Frequenzgruppen!

Flipped Classroom: Multimodal Dialog Systems

1. Benennen Sie Vor- und Nachteile multimodaler gegenüber rein sprachbasierter Mensch-Maschine Interaktion.
2. Was ist der Unterschied zwischen einem multimedialen und einem multimodalen Dialogsystem?
3. Nach welchen Regeln können Sie für einen Anwendungsfall geeignete Ein- und Ausgabemodalitäten auswählen?
4. Erläutern Sie den Aufbau und Funktionsweise eines multimodalen Dialogsystems.
5. Wie können Sie ein Gesicht maschinell erkennen?
6. Wie können Sie Blickrichtung erkennen und Verfolgen?
7. Welche Vorteile birgt die audiovisuelle gegenüber der rein akustischen Spracherkennung?
8. Erläutern Sie den Begriff Fusion und Fission.
9. Welche Arten von Gesten kennen Sie, und wie können Sie diese maschinell erkennen?
10. Was ist ein Embodied Conversational Agent (ECA), und wie funktioniert er?